

RESEARCH

Open Access



Origin of the nuclear proteome on the basis of pre-existing nuclear localization signals in prokaryotic proteins

Olga M. Lisitsyna¹, Margarita A. Kurnaeva², Eugene A. Arifulin¹, Maria Y. Shubina^{1,2}, Yana R. Musinova^{1,3,4}, Andrey A. Mironov^{2,5,6,7} and Eugene V. Sheval^{1,8,9*}

Abstract

Background: The origin of the selective nuclear protein import machinery, which consists of nuclear pore complexes and adaptor molecules interacting with the nuclear localization signals (NLSs) of cargo molecules, is one of the most important events in the evolution of eukaryotic cells. How proteins were selected for import into the forming nucleus remains an open question.

Results: Here, we demonstrate that functional NLSs may be integrated in the nucleotide-binding domains of both eukaryotic and prokaryotic proteins and may coevolve with these domains.

Conclusion: The presence of sequences similar to NLSs in the DNA-binding domains of prokaryotic proteins might have created an advantage for nuclear accumulation of these proteins during evolution of the nuclear-cytoplasmic barrier, influencing which proteins accumulated and became compartmentalized inside the forming nucleus (i.e., the content of the nuclear proteome).

Reviewers: This article was reviewed by Sergey Melnikov and Igor Rogozin.

Open peer review: Reviewed by Sergey Melnikov and Igor Rogozin. For the full reviews, please go to the Reviewers' comments section.

Keywords: Nucleus, Nuclear proteome, Nuclear localization signal, Evolution

Background

Acquisition of a cell nucleus enabled the spatial segregation of transcription and translation and likely permitted the evolution of more sophisticated mechanisms of gene expression regulation [1]. Because proteins are translated in the cytoplasm, the emergence of a reliable and efficient nuclear import mechanism was the essential event leading to the origin of the eukaryotic cell. Nucleocytoplasmic

transport across the nuclear envelope occurs predominantly through nuclear pore complexes (NPCs). Although small proteins can freely diffuse through NPCs, globular molecules larger than ~ 40 kDa are selectively transported via an energy-dependent mechanism that requires additional transport factors, called karyopherins, which recognize nuclear localization signals (NLSs) in cargo proteins [2]. Past studies have revealed some important events in the evolution of the nuclear envelope and possible ancestors of the key elements of the import machinery: NPCs and karyopherins [3–7]. However, it remains unclear how the proteins were selected for import into the forming nucleus, i.e., how the nuclear proteome evolved.

* Correspondence: sheval_e@belozersky.msu.ru

¹Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, 119991 Moscow, Russia

⁸Department of Cell Biology and Histology, Faculty of Biology, Lomonosov Moscow State University, 119991 Moscow, Russia

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Methods

Human proteins containing NLSs were collected from NLSdb (<https://www.rostlab.org/services/nlsdb1/browse.php/>) and the UniProt database. Annotations of protein domain structure were obtained from the UniProt/Swiss-Prot database. Regions between the nearest annotated domains were analyzed as out-of-domain regions. Orthologs of human proteins with NLSs were found in the *Branchiostoma floridae*, *Danio rerio*, *Xenopus laevis*, *Pelodiscus sinensis*, and *Gallus gallus* proteomes using OrthoDB release 10 (<https://www.orthodb.org/>) (Supplementary Table S1).

Multiple alignment of orthologous sequences was performed with Clustal Omega. The conservation degree of multiple alignments was evaluated as the information content (IC) [8], which was calculated as follows:

$$IC = \frac{\sum_j I_j}{L} \quad (1)$$

$$I_j = \sum_b I(b, j) \quad (2)$$

$$I(b, j) = F(b, j) \times \log_2 \frac{F(b, j)}{p_b} \quad (3)$$

$$F(b, j) = \frac{N(b, j) + p_b}{r + 1} \quad (4)$$

where I_j is the IC of the j^{th} alignment column, L is the length of multiple alignment, $I(b, j)$ is the IC of amino acid residues type "b" in the j^{th} alignment column, $F(b, j)$ is the frequency of amino acid residues type "b" in the j^{th} alignment column, $N(b, j)$ is the number of amino acid residue type 'b' in the j^{th} alignment column, (p_b) (pseudo count) is the base frequency of amino acid residue type "b", and r is the number of rows in the alignment [9].

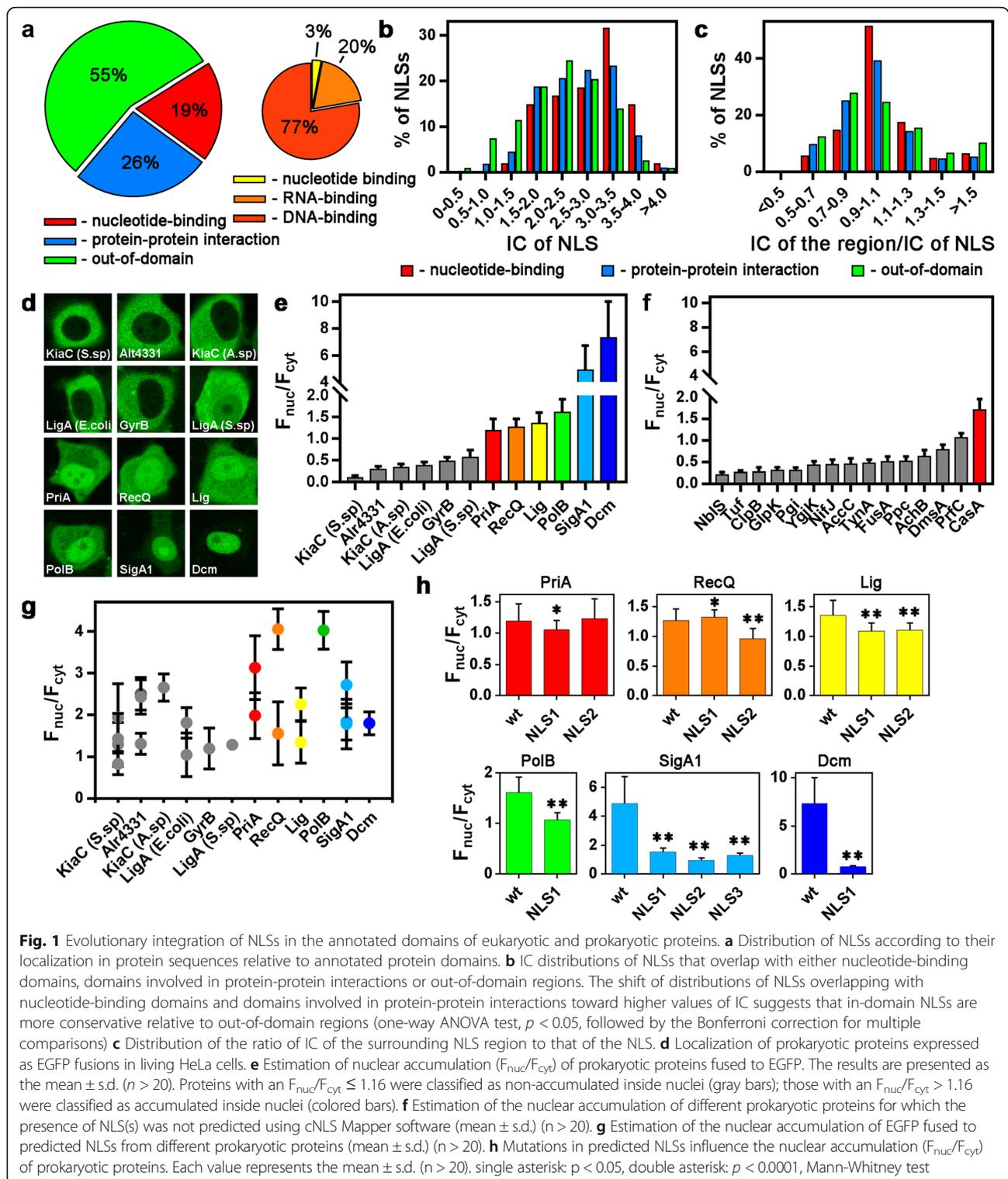
The *Thermococcus sibiricus* lineage was kindly provided by E.A. Bonch-Osmolovskaya. Genomic DNA of *Synechococcus* sp. and *Anabaena* sp. was provided by O.A. Koksharova and that of *Vibrio harveyi* by Y.V. Bertsova and A.V. Bogachev. Genes encoding target prokaryotic proteins were amplified by PCR from corresponding genomic DNA and inserted into the pEGFP-C1 vector (Clontech). Mutated genes of prokaryotic proteins were obtained by PCR site-directed mutagenesis. Double-stranded oligonucleotides encoding predicted NLSs of prokaryotic proteins were inserted into the pEGFP-C1 vector (Clontech). DNA fragments encoding M9M and Bimax2 peptides were inserted into the pTagRFP-C vector (Evrogen).

HeLa cells were grown in Dulbecco's modified Eagle's medium supplemented with L-glutamine (Paneco), 10% fetal calf serum (HyClone) and an antibiotic/antimycotic solution (Gibco). Cell transfection was performed using Lipofectamine 2000 reagent (Thermo Fisher Scientific) according to the manufacturer's instructions. Images of at least 20 living HeLa cells expressing EGFP-fused proteins were acquired in two different experiments using a Nikon C2 confocal laser scanning microscope. The ratio of nucleoplasmic EGFP concentration to cytoplasmic EGFP concentration (F_{nuc}/F_{cyt}) was measured as described elsewhere [10]. Statistical analysis and graph preparations were performed using Prism 6 (GraphPad software).

Results

To detect possible mechanisms of NLS origin, we analyzed data for NLSs localization relative to protein domains in modern organisms. We collected a dataset consisting of 592 annotated NLSs from 496 human proteins, among which 234 NLSs were identified experimentally and the other 358 NLS sequences were annotated in silico (Supplementary Table S1). Forty-five percent of all NLSs overlapped with some annotated domains (19% with nucleotide-binding domains and 26% with domains involved in protein-protein interactions); the other 55% of NLSs exhibited out-of-domain localization (Fig. 1a). The majority (77%) of the nucleotide-binding domains matching with NLSs are annotated as DNA-binding domains (Fig. 1a). Our data are in agreement with published data about colocalization of NLSs with DNA- and RNA-binding domains [11, 12].

We hypothesized the existence of an evolutionary link between NLSs and domains. To test this hypothesis, the conservation of NLSs and surrounding regions was analyzed by comparing the human protein sequences with their orthologs from five different species of phylum Chordata (*Branchiostoma floridae*, *Danio rerio*, *Xenopus laevis*, *Pelodiscus sinensis* and *Gallus gallus*). The degree of conservation of NLSs and surrounding regions (domains or out-of-domain regions) was evaluated as the IC of the obtained multiple alignment when the most conserved position in the alignment had a higher IC value. Comparison of the calculated IC distribution in three groups of NLSs demonstrated that NLSs overlapping annotated domains (both nucleotide-binding domains and domains involved in protein-protein interactions) are more conserved than NLSs located outside annotated domains (one-way ANOVA test, $p < 0.05$, followed by the Bonferroni correction for multiple comparisons) (Fig. 1b). To compare conservation between protein regions and NLSs overlapping these regions, the ratio of IC of each region to the IC of the corresponding NLS was calculated (Fig. 1c). Approximately half of all NLSs



overlapping nucleotide-binding domains had the same conservation degree as the corresponding domains (for 51% of NLSs, the ratio was within the 0.9–1.1 interval). These NLSs are integrated into nucleotide-binding

domains, and their evolution might depend on the evolution of the domains. Those NLSs overlapping domains involved in protein-protein interactions exhibited lower similarity with the surrounding protein regions (for 39%

of NLSs, the ratio was within the 0.9–1.1 interval); NLSs located outside domains did not demonstrate substantial similarity with the surrounding regions (the ratio was within the 0.9–1.1 interval only for 25% of NLSs).

NLSs are short and structurally simple sequences. For example, a monopartite ‘classical’ NLS has a degenerate consensus sequence of K(K/R)X(K/R) [13]. As nucleotide-binding domains are enriched in positively charged amino acids, the occasional appearance of such NLSs in such domains seems probable. As similar nucleotide-binding domains may be found in prokaryotic proteins, it seems plausible that these domains already contain sequences that potentially function as NLSs. If this supposition is correct, then such prokaryotic proteins would accumulate inside nuclei after expression in eukaryotic cells. We cloned 12 large (> 45 kDa) prokaryotic proteins with nucleotide-binding domains. The NLSs of all of these proteins were predicted using cNLS Mapper [14], and at least one overlapped with a nucleotide-binding domain (Supplementary Table S2). To produce a control group of proteins, we cloned and analyzed 15 large (> 45 kDa) proteins without predicted NLSs (Supplementary Table S2). The proteins were fused to enhanced green fluorescent protein (EGFP), and their localization was investigated in living HeLa cells. Approximately half of all the proteins accumulated inside nuclei, though to different degrees (Fig. 1d). To quantify the efficiency of nuclear accumulation in the nucleus, the ratio of nucleoplasmic to cytoplasmic ($F_{\text{nuc}}/F_{\text{cyt}}$) fluorescence was measured for all the proteins, as described elsewhere [10]. Proteins with $F_{\text{nuc}}/F_{\text{cyt}}$ values higher than that of EGFP, i.e., > 1.16, were classified as accumulating inside nuclei (Fig. 1e, Supplementary Table S2). No correlation between the efficiency of nuclear accumulation and the molecular weight of prokaryotic proteins was detected (Pearson correlation coefficient = 0.13), indicating that the transfer of proteins into the nucleus was not due to diffusion but rather due to an active process. Among control proteins, only one accumulated inside nuclei ($F_{\text{nuc}}/F_{\text{cyt}} = 1.71 \pm 0.25$), and this protein was the only one among the 15 control proteins with DNA-binding activity (Fig. 1f; Supplementary Table S2). These data are in agreement with published results indicating that NLSs are present not only in the proteins of eukaryotes but also in the proteins of prokaryotes [15–22] and bacteriophages [23].

It is possible that nuclear import of large (> 40 kDa) proteins depends on the presence of NLSs, which were predicted in all investigated prokaryotic proteins using cNLS Mapper [13] (Supplementary Table S2). To confirm that these protein regions are indeed functionally active NLSs, we constructed plasmids coding the predicted NLSs fused to EGFP. All of the predicted NLSs were able to accumulate EGFP inside nuclei (Fig. 1g, Supplementary Table S3); however, the $F_{\text{nuc}}/F_{\text{cyt}}$ values

of the predicted NLSs did not correlate with the $F_{\text{nuc}}/F_{\text{cyt}}$ values of the corresponding full-length proteins (Supplementary Fig. S1) (Pearson’s correlation coefficient between the $F_{\text{nuc}}/F_{\text{cyt}}$ of the strongest among all predicted NLSs and the $F_{\text{nuc}}/F_{\text{cyt}}$ of the proteins = 0.14). Therefore, these results can be considered only an indication of potential NLS activity. We also employed site-directed mutagenesis to directly detect the presence of NLSs.

Substitutions of all positively charged amino acids in each predicted NLS with alanine decreased the nuclear accumulation ($F_{\text{nuc}}/F_{\text{cyt}}$) of all proteins that had been classified as accumulated inside nuclei ($F_{\text{nuc}}/F_{\text{cyt}} > 1.16$) (Fig. 1h, Supplementary Table S4).

Nuclear import of proteins containing a classical NLS depends on interaction of the NLS sequence with karyopherin- α and karyopherin- β ; nonclassical NLSs directly interact with karyopherin- β for nuclear import. We next applied the inhibitors Bimax2 [24] and M9M [25], which bind highly specifically to karyopherin- α and karyopherin- β , respectively. Nuclear accumulation of PriA, Lig, PolB and SigA1 was decreased by coexpression of both Bimax2 (Supplementary Fig. S2) and M9M (Supplementary Fig. S3), whereas nuclear accumulation of Dcm was reduced only by Bimax2. Accordingly, these proteins accumulate inside nuclei via the ‘classical’ karyopherin- α/β -dependent pathway. Additionally, nuclear accumulation of RecQ was decreased by coexpression of M9M but not Bimax2, indicating the presence of a nonclassical NLS in this protein.

Discussion

Overall, our data indicate that regions enriched with the positively charged amino acids of nucleotide-binding domains can indeed serve as genuine NLSs. These NLSs are integrated into domains, and their evolution might depend on the evolution of the corresponding domains. Such NLSs, even if they are present in prokaryotic proteins, can interact with karyopherins. Karyopherins have many functions in the cell and, in particular, can act as chaperones [26, 27]. The protein domains interacting with karyopherins might have evolved before the origin of the nuclear envelope, with these domains containing sequences that potentially play a role in NLSs. The presence of sequences similar to NLSs in DNA-binding domains of prokaryotic proteins might create an advantage for nuclear accumulation of these proteins during evolution of the nuclear-cytoplasmic barrier, influencing which proteins accumulated and became compartmentalized inside the forming nucleus (the content of the nuclear proteome). Proteins that did not harbor such integrated NLSs might have acquired them *de novo* after nuclear envelope formation, and such NLSs can be considered separate units of genome evolution. Interestingly, sequences that are similar to NLS can also be predicted

and experimentally defined as being present in some cytoplasmic proteins of modern organisms. This indicates that during evolution, some proteins, albeit possibly resident inside nuclei due to the presence of an integrated NLS, were excluded from the nucleus via different mechanisms, as discussed elsewhere [28].

Reviewers' comments.

Reviewer's report 1.

Sergey Melnikov.

Reviewer comments:

I reviewed this manuscript in detail when it was submitted to *Molecular Biology and Evolution*. I recommended the authors to make numerous changes, and they addressed every single of my comments. I therefore have no reason to criticize this work any further. This study is important to the field as it shows that the nuclear localization signals in modern eukaryotic proteins could simply emerge from DNA-/RNA-binding domains of cellular proteins, because having a DNA- or RNA-binding domain is frequently sufficient for a protein to be recognized as a nucleus resident. This is an important finding and I encourage you to publish this work as is.

In this concise and thought-provoking manuscript, Olga Lisitsyna et al. investigate a central evolutionary enigma: the origin of the cell nucleus. The authors convincingly show that, in most instances, all that a protein needs to enter the cell nucleus is a DNA-binding domain. For instance, in their experiments with prokaryotic proteins, they show that – even in the absence of predicted NLS sequences – some DNA-binding prokaryotic proteins are actively transported into the cell nucleus (Fig. 1). This experiment, along with their analysis of NLS overlaps with DNA-binding domains in protein structures, suggests that NLSs have initially evolved from (and within) DNA-binding domains of chromatin-binding proteins – the conclusion that makes the perfect sense from the point of evolutionary contingency. Furthermore, in their supplementary data, the authors have collected a wonderful review of the experimentally identified and predicted nuclear localization signals. This information alone will be very useful for other scientists working in the field of the origin of eukaryotes and origin of the nucleus.

My only suggestion to the authors is to divide their data set of NLSs into two groups – experimentally-defined vs in silico predicted: when they describe their statistics on the % of NLSs overlap with RNA/DNA-binding domains, it seems useful to me to provide it first for the experimentally-defined NLSs (as a more reliable data), and then complement these numbers with additional data for in silico-identified NLSs.

Author's response:

We thank the reviewer for the critical evaluation of our work and the positive feedback. Of course, we agree that results based only on analysis of experimentally

defined NLSs should be more robust and reliable than those based on analysis of consolidated datasets (both experimentally defined and in silico-predicted NLSs). Unfortunately, the number of experimentally defined NLSs is not as large as necessary for the appropriate statistical analysis. Therefore, we used a dataset of NLSs, including both experimentally defined and in silico-predicted NLSs.

Reviewer's report 2.

Igor Rogozin.

Reviewer comments:

The authors demonstrated that NLS and NLS-like motifs may be integrated inside nucleotide binding domains of both eukaryotic and prokaryotic proteins and may co-evolve with these domains. They proposed that there are NLS-like motifs inside prokaryotic proteins that may be functionally important.

The authors need to choose the theoretical framework. If the authors would like to operate within the framework of evolutionary biology, they cannot use sentences like: "We propose that the pre-existence of NLSs inside prokaryotic proteins dictated, at least partially, the nuclear proteome composition.". Prokaryotes do not have nucleus thus they do not have NLS and those NLS-like sequences cannot "... dictated, at least partially, the nuclear proteome composition" (due to the absence of the nucleus). Those NLS-like sequences may have some functional roles, this is possible. Just an example, fragments of mobile elements (MEs) may be a part of promoter or protein coding regions. However I doubt that the "pre-existence" of MEs "dictated" regulatory pathways or functions of protein coding genes. According to Wojtek Makalowski it is something like scrap yard (Makalowski W. Genomic scrap yard: how genomes utilize all that junk. *Gene*. 2000, 259(1–2):61–7). I think that the authors need to use something like "prokaryotic sequences similar to NLSs or NLS like signals etc." (if they are willing to operate within the framework of evolutionary biology). If the authors would like to operate within frameworks of alternative hypotheses, it is better to notify readers about that. Otherwise a careful correction of logic and language is required.

This structure: ... However, it remains unclear how the proteins were selected for import into the forming nuclei, i.e., how the nuclear proteome evolved." The Methods section The Results section To address this question, we analysed data on NLSs and their localization relative to protein domains. .. does not look good to me. The question and attempts to answer are separated by the Methods section.

Author's response:

We thank the reviewer for taking the time to review our manuscript and for providing these comments.

We substantially modified the sentence “We propose that the pre-existence of NLSs inside prokaryotic proteins dictated, at least partially, the nuclear proteome composition”. Our logic was based on the data presented as well as on some published results (references [15–23]), which indicate that the NLSs in modern eukaryotic proteins might have evolved from the DNA-binding domains of prokaryotic proteins. As a result, some DNA-binding domains are sufficient for interaction with karyopherins, and as a consequence, a protein may have had features of a nuclear protein before the origin of the cell nucleus. Of course, these features would not be useful before the origin of the nuclear envelope. Interestingly, sequences that are similar to NLSs can also be found in some domains of cytoplasmic proteins of modern organisms (Kharitonov A.V., Shubina M.Y., Nosov G.A., Mamontova A.V., Arifulin E.A., Lisitsyna O.M., Nalobin D.S., Musinova Y.R., Sheval E.V. Switching of cardiac troponin I between nuclear and cytoplasmic localization during muscle differentiation. *Biochimica et Biophysica Acta – Molecular Cell Research*. 2020. 1867(2):118601). We described this as follows: “The presence of sequences similar to NLSs in DNA-binding domains of prokaryotic proteins might create an advantage for nuclear accumulation of these proteins during evolution of the nuclear-cytoplasmic barrier, influencing which proteins accumulated and became compartmentalized inside the forming nucleus (the content of the nuclear proteome). Proteins that did not harbor such integrated NLSs might have acquired them de novo after nuclear envelope formation, and such NLSs can be considered separate units of genome evolution. Interestingly, sequences that are similar to NLS can also be predicted and experimentally defined as being present in some cytoplasmic proteins of modern organisms. This indicates that during evolution, some proteins, albeit possibly resident inside nuclei due to the presence of an integrated NLS, were excluded from the nucleus via different mechanisms, as discussed elsewhere [28]”.

We modified the first sentence of the “Results” section as follows: “To detect possible mechanisms of NLS origin, we analyzed data for NLSs localization relative to protein domains in modern organisms.”

Finally, it should be noted that the manuscript was edited by American Journal Experts to improve phrasing and remove grammar and writing errors.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13062-020-00263-6>.

Additional file 1: Supplementary Table S1. NLSs identified experimentally or predicted in silico.

Additional file 2: Supplementary Table S2. Prokaryotic proteins with or without predicted NLSs.

Additional file 3: Supplementary Table S3. Predicted NLSs from prokaryotic proteins are able to target EGFP to the cell nucleus.

Additional file 4: Supplementary Table S4. Detection of NLSs inside prokaryotic proteins by site-directed mutagenesis.

Additional file 5: Supplementary Fig. S1. Comparison of nuclear accumulation (F_{nuc}/F_{cyl}) of all predicted NLSs and nuclear accumulation (F_{nuc}/F_{cyl}) of the full-length proteins fused with EGFP in living HeLa cells.

Supplementary Fig. S2. Decrease in nuclear accumulation of prokaryotic proteins by a peptide inhibitor of karyopherin- α (Bimax2). NLS from the T antigen of SV40 virus (NLS^{SV40}) fused with EGFP was used as a positive control. Expression of TagRFP-Bimax2 leads to a decrease in the nuclear accumulation of NLS^{SV40}. A decrease in nuclear accumulation was also detected for five prokaryotic proteins, namely, PriA, Lig, PolB, SigA1 and Dcm.

Supplementary Fig. S3. Decrease in the nuclear accumulation of prokaryotic proteins by a peptide inhibitor of karyopherin- β (M9M). The NLS from FUS protein (NLS^{FUS}) fused with EGFP was used as a positive control. The expression of TagRFP-M9M leads to a decrease in the nuclear accumulation of NLS^{FUS}. A decrease in nuclear accumulation was also detected for five prokaryotic proteins, namely, PriA, RecQ, Lig, PolB and SigA1.

Acknowledgments

We are grateful to E.A. Bonch-Osmolovskaya for providing *Thermococcus sibiricus* lineage, O.A. Koksharova for providing genomic DNA of *Anabaena* sp. and *Synechococcus* sp., Y.V. Bertsova and A.V. Bogachev for providing genomic DNA of *Vibrio harveyi*. We thank Y.S. Vassetzky for valuable discussion.

Authors' contributions

OML, AAM and EVS conceived the study and design the experiments; OML, MAK, EAA, MYS, YRM and EVS performed the experiments; OML, AAM and EVS performed the data analysis; OML, AAM and EVS prepared the manuscript; All authors edited and reviewed the manuscript; All authors read and approved the final manuscript.

Funding

This work was supported by the Russian Science Foundation (grant 18–14–00195).

Availability of data and materials

All supporting data are submitted in Supplementary Materials.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, 119991 Moscow, Russia. ²Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119991 Moscow, Russia. ³Koltzov Institute of Developmental Biology, Russian Academy of Sciences, 119334 Moscow, Russia. ⁴Skobelkin State Scientific Center of Laser Medicine FMBA, 121099 Moscow, Russia. ⁵Skolkovo Institute of Science and Technology, 121205, Moscow, Russia. ⁶Institute for Information Transmission Problems, Russian Academy of Sciences, 127051 Moscow, Russia. ⁷Faculty of Computer Science, National Research University Higher School of Economics, 101000 Moscow, Russia. ⁸Department of Cell Biology and Histology, Faculty of Biology, Lomonosov Moscow State University, 119991 Moscow, Russia. ⁹LIA 1066 LFR2O French-Russian Joint Cancer Research Laboratory, 94805 Villejuif, France.

Received: 21 January 2020 Accepted: 7 April 2020

Published online: 28 April 2020

References

- Martin W, Koonin EV. Introns and the origin of nucleus-cytosol compartmentalization. *Nature*. 2006;440:41–5.
- Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH. Classical nuclear localization signals: definition, function, and interaction with importin alpha. *J Biol Chem*. 2007;282:5101–5.
- Devos D, Dokudovskaya S, Alber F, Williams R, Chait BT, Sali A, Rout MP. Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol*. 2004;2:e380.
- Mans BJ, Anantharaman V, Aravind L, Koonin EV. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. *Cell Cycle*. 2004;3:1612–37.
- Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, et al. The molecular architecture of the nuclear pore complex. *Nature*. 2007;450:695–701.
- Brohawn SG, Leksa NC, Spear ED, Rajashankar KR, Schwartz TU. Structural evidence for common ancestry of the nuclear pore complex and vesicle coats. *Science*. 2008;322:1369–73.
- DeGrasse JA, DuBois KN, Devos D, Siegel TN, Sali A, Field MC, Rout MP, Chait BT. Evidence for a shared nuclear pore complex architecture that is conserved from the last common eukaryotic ancestor. *Mol Cell Proteomics*. 2009;8:2119–30.
- Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *J Mol Biol*. 1986;188:415–31.
- Gaur R. Amino acid frequency distribution among eukaryotic proteins. *IIOAB J*. 2014;5:6–11.
- Musinova YR, Lisitsyna OM, Golyshev SA, Tuzhikov AI, Polyakov VY, Sheval EV. Nucleolar localization/retention signal is responsible for transient accumulation of histone H2B in the nucleolus through electrostatic interactions. *Biochim Biophys Acta*. 2011;1813:27–38.
- LaCasse EC, Lefebvre YA. Nuclear localization signals overlap DNA- or RNA-binding domains in nucleic acid-binding proteins. *Nucleic Acids Res*. 1995; 23:1647–56.
- Cokol M, Nair R, Rost B. Finding nuclear localization signals. *EMBO Rep*. 2000;1:411–5.
- Chelsky D, Ralph R, Jonak G. Sequence requirements for synthetic peptide-mediated translocation to the nucleus. *Mol Cell Biol*. 1989;9:2487–92.
- Kosugi S, Hasebe M, Tomita M, Yanagawa H. Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc Natl Acad Sci U S A*. 2009;106:10171–6.
- Rossi L, Hohn B, Tinland B. The VirD2 protein of *Agrobacterium tumefaciens* carries nuclear localization signals important for transfer of T-DNA to plant. *Mol Gen Genet*. 1993;239:345–53.
- Nederlof PM, Wang HR, Baumeister W. Nuclear localization signals of human and *Thermoplasma* proteasomal alpha subunits are functional *in vitro*. *Proc Natl Acad Sci U S A*. 1995;92:12060–4.
- Perić M, Schedewig P, Bauche A, Kruppa A, Kruppa J. Ribosomal proteins of *Thermus thermophilus* fused to beta-galactosidase are imported into the nucleus of eukaryotic cells. *Eur J Cell Biol*. 2008;87:47–55.
- Lee JH, Jun SH, Baik SC, Kim DR, Park J-Y, Lee YS, Choi CH, Lee JC. 2012 prediction and screening of nuclear targeting proteins with nuclear localization signals in *Helicobacter pylori*. *J Microbiol Methods*. 2008;91:490–6.
- Moon DC, Gurusu M, Lee JH, Lee YS, Choi CW, Kim SI, Lee JC. Screening of nuclear targeting proteins in *Acinetobacter baumannii* based on nuclear localization signals. *Res Microbiol*. 2012;163:279–85.
- Kim J-M, Choe M-H, Asaithambi K, Song J-Y, Lee YS, Lee JC, Seo J-H, Kang H-L, Lee KH, Lee W-K, et al. *Helicobacter pylori* HP0425 targets the nucleus with DNase I-like activity. *Helicobacter*. 2016;21:218–25.
- Kwon YC, Kim S, Lee YS, Lee JC, Cho M-J, Lee W-K, Kang H-L, Song J-Y, Baik SC, Ro HS. Novel nuclear targeting coiled-coil protein of *Helicobacter pylori* showing Ca²⁺-independent, Mg²⁺-dependent DNase I activity. *J Microbiol*. 2016;54:387–95.
- Melnikov S, Kwok HS, Manakongtreecheep K, van den Elzen A, Thoreen CC, Söll D. Archaeal ribosomal proteins possess nuclear localization signal-type motifs: implications for the origin of the cell nucleus. *Mol Biol Evol*. 2019;37: 124–33.
- Redrejo-Rodríguez M, Muñoz-Espín D, Holguera I, Mencía M, Salas M. Functional eukaryotic nuclear localization signals are widespread in terminal proteins of bacteriophages. *Proc Natl Acad Sci U S A*. 2012;109:18482–7.
- Kosugi S, Hasebe M, Entani T, Takayama S, Tomita M, Yanagawa H. Design of peptide inhibitors for the importin alpha/beta nuclear import pathway by activity-based profiling. *Chem Biol*. 2008;15:940–9.
- Cansizoglu AE, Lee BJ, Zhang ZC, Fontoura BMA, Chook YM. Structure-based design of a pathway-specific nuclear import inhibitor. *Nat Struct Mol Biol*. 2007;14:452–4.
- Jäkel S, Mingot J-M, Schwarzmaier P, Hartmann E, Görlich D. Importins fulfill a dual function as nuclear import receptors and cytoplasmic chaperones for exposed basic domains. *EMBO J*. 2002;21:377–86.
- Apta-Smith MJ, Hernandez-Fernaud JR, Bowman AJ. Evidence for the nuclear import of histones H3.1 and H4 as monomers. *EMBO J*. 2018;37: e98714.
- Kharitonov AV, Shubina MY, Nosov GA, Mamontova AV, Arifulin EA, Lisitsyna OM, Nalobin DS, Musinova YR, Sheval EV. Switching of cardiac troponin I between nuclear and cytoplasmic localization during muscle differentiation. *Biochim Biophys Acta Mol Cell Res*. 2020;1867:118601.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

