ISSN 1993-6842 (on-line); ISSN 0233-7657 (print)
Biopolymers and Cell. 2017. Vol. 33. N 2. P 147–154
doi: http://dx.doi.org/10.7124/bc.00094C

UDC 577.112

Comparative analysis of nuclear localization signal (NLS) prediction methods

O. M. Lisitsyna[1], V. B. Seplyarskiy[2], E. V. Sheval[1,3]

[1] A. N. Belozersky Institute of Physico-Chemical Biology, M. V. Lomonosov Moscow State University
Leninskie gory, house 1, building 40, Moscow, Russian Federation, 119992

[2] A. A. Kharkevich Institute for Information Transmission Problems,
19/1, Bolshoy Karetny per. Moscow, Russian Federation, 127051

[3] LIA 1066 French-Russian Joint Cancer Research Laboratory
Villejuif, France–Moscow, Russian Federation
lisitsynaom@gmail.com

Aim. Comparative analysis of six state-of-the-art nuclear localization signal (NLS) prediction methods (PSORT II, NucPred, cNLSMapper, NLStradamus, NucImport and seqNLS). **Methods.** Each program was tested for correct predictions using a dataset of 155 experimentally determined NLSs and for false-positives using a dataset of 155 transmembrane proteins, which putatively lack NLS. **Results.** The most suitable NLS predictors wer fond to be NucPred, NLStradamus and seqNLS; these programs provide the maximum rate of correct to wrong predictions among the tested programs. However, the best results obtained by these programs were only ~45 % of the correct predictions. **Conclusion.** The identification of novel NLSs by predictors still requires experimental verification.

K e y w o r d s: nuclear localization signal, prediction

Introduction

The nuclear envelope separates the nucleus from the cytoplasm and provides bi-directional traffic via nuclear pore complexes [1, 2]. Small proteins (up to ~40 kDa) can freely permeate the nuclear envelope [3, 4], whereas the traffic of the larger proteins is an active process that depends on the binding of short stretches of amino acids referred to as nuclear localization signals (NLSs) with special adaptor proteins, karyopherins [5].

The best-characterized NLSs are the classical NLSs (cNLSs) [6], which are recognized by the carrier protein karyopherin-α (importin-α) [7]. cNLSs include two types of signals: monopartite NLSs having a single cluster of basic amino acid residues and bipartite NLSs having two clusters of basic amino acids separated by a 10–12 amino acid linker [6]. In addition to the

cNLS, several alternative types of NLSs have been characterized, including the PY-NLSs with consensus sequence [basic/hydrophobic]-Xn-[R/H/K]-X2–5-PY [8], the acidic M9 domain of hnRNP A1 [9], the sequence KIPIK in yeast transcription repressor Matα2 [10], the complex signals of U snRNPs [11], PTHrP domain [12], IBB domain [13], and many others. Predominantly, these non-classical NLSs (ncNLSs) are translocated into the nucleus via interaction with karyopherin-β [14].

The identification of novel NLSs is still a quite complicated and time-consuming task for experimental biology. Developing methods of computational biology predicting possible variants of NLSs can significantly contribute to progress in this field. Some predictor programs with different algorithms are available to identify the putative NLS (Table). Recently, it has been demonstrated that the information about the protein localization, predicted with the bioinformatic approaches using data from protein databases, such as Protein Atlas, UniProt, LocDB and Gene Ontology, does not fully concur with the nuclear proteome data [15]. Moreover, the NLS prediction cannot completely guarantee the accurate identification of novel NLSs [16], which indicates that the precision of prediction may be a major factor limiting the effectiveness and rapidity of the experimental NLS research. Here, we analyzed six state-of-the-art NLS prediction programs to detect the restrictions of NLS prediction methods and find the most effective method.

Table. **Prediction programs used for NLS identification**

Predictor	Web address
PSORT II	http://psort.hgc.jp/form2.html
NucPred	https://www.sbc.su.se/~maccallr/nucpred/
cNLSMapper	http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi
NLStradamus	http://www.moseslab.csb.utoronto.ca/NLStradamus/
NucImport	http://bioinf.scmb.uq.edu.au:8080/NucImport/
seqNLS	http://mleg.cse.sc.edu/seqNLS/

Materials and Methods

Datasets

We used 155 experimentally determined NLSs from 128 human proteins from Uniprot database (http://www.uniprot.org/). We only used the proteins with manually annotated descriptions. To provide high protein diversity, we excluded the closely related proteins (identity between amino acid sequences is more than 65 %) in our dataset (available on request from the corresponding author). According to the published data, known cNLS could be described by the following amino acid patterns: K(R/K)X(R/K) [17], K(K/R)X(K/R) [18], KR(R/X)K [19], KRRR [20], (P/R) XXKR(^DE)(K/R), KRX(W/F/Y)XXAF, (R/P)XXKR(K/R)(^DE), KR(K/R)R or K(K/R)RK [21] for a monopartite cNLS, and (K/R)(K/R)$X_{10\text{-}12}$(K/R)$_{3/5}$ [22], KR$X_{10\text{-}12}$KRRK [19], KR$X_{10\text{-}12}$K(K/R)(K/R) or KR$X_{10\text{-}12}$K(K/R)X(K/R) [21] for a bipartite cNLS. Comparison of NLSs from a created dataset of experimental NLSs with these patterns demonstrates that the majority of them (120 of 155) may be classified as cNLSs.

In total, 155 random transmembrane proteins from the Protein Data Bank of Transmembrane proteins (http://pdbtm.enzim.hu/)

were selected for the control dataset. Two extra datasets of transmembrane proteins (alpha type and beta type), each of the same size, were created to validate the results obtained for the first transmembrane protein dataset.

Prediction performance evaluation

To measure prediction performance, we used the following criteria:

(1) True positive rate = $N_{true\ postitive}/N_{expNLS}$

(2) False positive rate = $N_{false\ positive}/N_{TMP}$

where $N_{true\ positive}$ is the number of correct predictions in protein dataset with experimental NLS (N_{expNLS}), thus

$$N_{expNLS} = N_{true\ postitive} + N_{false\ negative},$$

To be able to calculate a false positive rate, we considered no more than one NLS per transmembrane protein and ignored any NLS outside the experimentally predicted ones in our positive cohort of proteins. We determined the correct prediction as a result that overlapped with experimental NLS by more than three amino acid residues.

$N_{false\ positive}$ is the number of transmembrane proteins with predicted NLS, N_{TMP} is the total number of transmembrane proteins in dataset, thus

$$N_{TMP} = (N_{true\ negative} + N_{false\ positive})$$

The Matthews' Correlation Coefficient (MCC) [23] was also defined to measure the correlation between prediction and observation:

$$MCC = \frac{N_{true\ positive} \times N_{true\ negative} - N_{false\ positive} \times N_{false\ negative}}{\sqrt{N_{expNLS} \times (N_{true\ positive} + N_{false\ positive}) \times N_{TMP} \times (N_{true\ negative} + N_{false\ negative})}}$$

Statistical analysis

The statistical analysis was performed by R statistical computing.

Results and Discussion

An approach

We compared the prediction performance of the following six programs: PSORT II [24], NucPred [25], cNLSMapper [20], NLStradamus [26], NucImport [27] and seqNLS [28] (Table). The number of correct predictions and the rate of false negative results were evaluated using the dataset of proteins with experimental NLSs. However, the amount of false positive predictions and true negative values were calculated based on a transmembrane protein dataset (155 proteins) suggesting that transmembrane proteins do not contain any NLSs. For equalization of true positive and false positive results, we considered the prediction of multiple NLSs within one transmembrane protein as one predicted NLS. Validation of the datasets of transmembrane proteins with two extra datasets of alpha and beta types of transmembrane proteins demonstrated the similar results for all predictors (data not shown); thus, the first dataset of 155 random transmembrane proteins could be applied as a negative control.

Search for optimal program operation modes

Algorithms of seqNLS, cNLSMapper and NLStradamus have a cut-off score option for

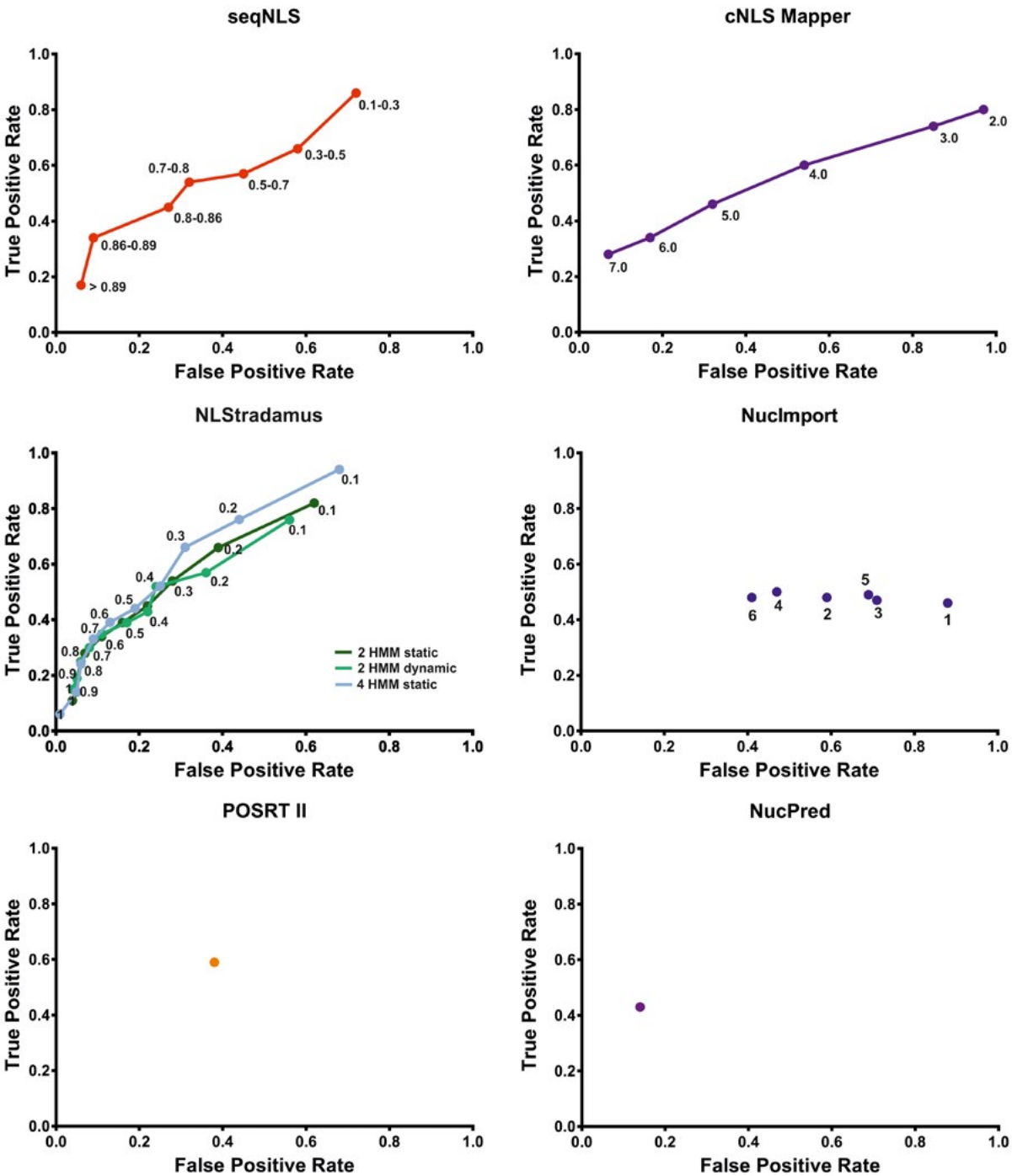

Fig. 1. Evaluation of the prediction performance of different NLS predictors (True Positive Rate *versus* False Positive Rate). Different cut-off scores are labeled for seqNLS, cNLS Mapper and NLStradamus as well as six types of training models for NucImport.

their prediction results. Based on this function, we obtained the ROC-curve to evaluate the True Positive Rate and False Positive Rate at different prediction cut-off scores (Fig. 1). The NLStradamus has not only a cut-off score option but also the following three different prediction algorithms: simple two-state static or dynamic Hidden Markov Models (HMM) algorithms and a four-state static HMM algorithm. The ROC-curves were evaluated for each of these algorithms. For other predictors (NucPred, PSORT II and NucImport), only one value of the true positive to false positive results ratio was obtained (Fig. 1). The output of NucPred provides the colored query sequence from blue (small probability of nuclear localization) to red (high probability of nuclear localization). In the case of prediction with strict conditions (colored from orange and red), only 18 % of experimental NLSs were correctly predicted (data not shown). For this reason, the prediction performance criteria of NucPred were evaluated with less strict conditions (colored from green to red) with an increase in the numbers of correct predictions (43 %). NucImport has six training models as well as the parameter "name of species" (mouse or yeast) that can be used for predictions. We tested NucImport at each of the six models, but only with the "mouse" parameter as the "name of species" because it was more related to our dataset of human proteins.

Comparison of the predictor programs

Figure 1 shows the prediction results for the six considered computational approaches. ROC-curve comparison revealed that a lower cut-off score provided the maximum false positive results as well as the correct predictions of experimental NLS. At the points with lower cut-off scores, the number of correct predictions was approximately equal to the number of false predictions. However, the higher cut-off scores allow for a more than 4-fold correct prediction to the false positive ratio in the best cases for NLStradamus. Among six evaluated programs NucPred, NLStradamus (at cut-off scores of 0.5–1) and seqNLS service (at cut-off scores of 0.8–0.86) showed the best prediction achievements. Additionally, the evaluation of the prediction performance for each NLStradamus HMMs did not show significant differences between them at the cut-off score from 0.5 to 1 (Fig. 1). PSORT II can be compared with the NLStradamus at cut-off score of 0.2 (Fig. 1). At the all range of cut-off scores cNLSMapper provided less true positive and more false positive predictions than NLStradamus and seqNLS. Only at the strongest cut-off score (7.0) prediction achievements of cNLSMapper were similar to NLStradamus (Fig. 1). In the case of NucImport, the rate of correct predictions was the same for all six models, but the minimum of the false positive results was calculated for model 6 (Fig. 1). Nevertheless, the best NucImport model 6 provided an equal ratio of correct and incorrect predictions, which was the worse prediction achievement among the estimated programs.

To evaluate the correlation between prediction and observation, the Matthews' Correlation Coefficient (MCC) [23] was calculated for each predictor at its best settings (cut-off score, prediction model). A coefficient of +1 represented a perfect prediction, 0 indicated a result no better than the random result and –1 indicated total disagreement between prediction and observation. The highest MCC (~0.3) was

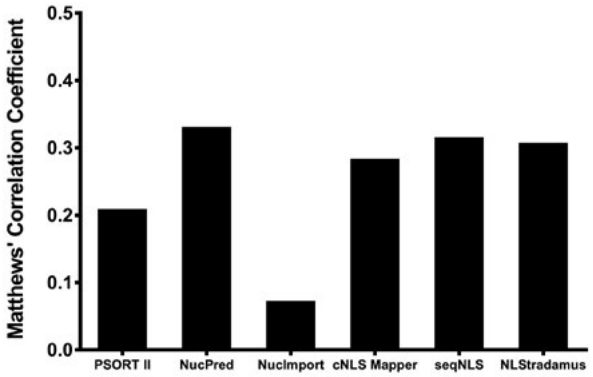

Fig. 2. Calculated Matthews' Correlation Coefficient. The best values are presented for NucPred, cNLSMapper, seqNLS and NLStradamus.

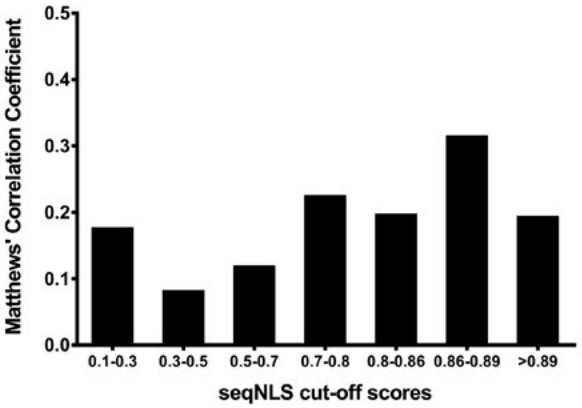

Fig. 3. Calculated Matthews' Correlation Coefficient for seqNLS at different cut-off scores.

obtained for NucPred, seqNLS (cut-off score 0.8–0.86) and NLStradamus (cut-off score 0.5), when the best values of cNLSMapper and PSORT II were also close (0.28 and 0.2 correspondingly). According to MCC, the best prediction model of NucImport demonstrated random prediction (Fig. 2). Variation in the cut-off score of the predictors also influenced MCC; the decrease of the cut-off score led to random results (MCC is near 0) (Fig. 3).

Conclusion

In this study, we estimated the prediction performance of six NLS predictors using the following two types of datasets: human proteins with experimentally identified NLS and transmembrane proteins. The best True Positive Rate and False Positive Rate and the highest MCC were obtained for NucPred, NLStradamus (at cut-off scores of 0.5–1) and seqNLS service (at cut-off scores of 0.8–0.86). The prediction achievements of cNLS Mapper and PSORT II were a little bit worse. Our data are in agreement with Lin & Hu [28] who demonstrated that the seqNLS was a better predictor than cNLSMapper. However, our results indicated that NLStradamus showed the same or even better results than the seqNLS on our dataset of human proteins. It should be stressed that even at the highest True Positive Rate and minimum False Positive Rate, the best programs (NucPred, NLStradamus, seqNLS) correctly identified only ~45 % of the experimental NLSs. Therefore, the identification of novel NLS by predictors still requires experimental verification.

Acknowledgment

We are grateful to Prof. A. A. Mironov for valuable suggestion and critical comments.

Funding

The work was supported by the Russian Science Foundation (project 14-15-00199).

REFERENCES

1. *Dickmanns A, Kehlenbach RH, Fahrenkrog B.* Nuclear pore complexes and nucleocytoplasmic transport: from structure to function to disease. *Int Rev Cell Mol Biol.* 2015; **320**; 171–233.

2. *Sheval EV, Musinova YR.* Structural plasticity of the nuclear envelope and the endoplasmic reticulum. *Biopolym Cell.* 2014; **30**(5): 335–42.
3. *Keminer O, Peters R.* Permeability of single nuclear pores. *Biophys J.* 1999; **77**(1): 217–28.
4. *Feldherr CM, Akin D.* The location of the transport gate in the nuclear pore complex. *J Cell Sci.* 1997;**110** (Pt 24): 3065–70.
5. *Pemberton LF, Paschal BM.* Mechanisms of receptor-mediated nuclear import and nuclear export. *Traffic.* 2005; **6**(3): 187–98.
6. *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**(8): 5101–5.
7. *Miyamoto Y, Yamada K, Yoneda Y.* Importin α: a key molecule in nuclear transport and non-transport functions. *J Biochem.* 2016; **160**(2):69–75.
8. *Lee BJ, Cansizoglu AE, Süel KE, Louis TH, Zhang Z, Chook YM.* Rules for nuclear localization sequence recognition by karyopherin beta 2. *Cell.* 2006; **126**(3): 543–58.
9. *Siomi H, Dreyfuss G.* A nuclear localization domain in the hnRNP A1 protein. *J Cell Biol.* 1995; **129**(3): 551–60.
10. *Hall MN, Hereford L, Herskowitz I.* Targeting of E coli beta-galactosidase to the nucleus in yeast. *Cell.* 1984; **36**(4): 1057–65.
11. *Fischer U, Sumpter V, Sekine M, Satoh T, Luhrmann R.* Nucleo-cytoplasmic transport of U snRNPs: definition of a nuclear location signal in the Sm core domain that binds a transport receptor independently of the m3G cap. *EMBO J.* 1993; **12**(2): 573–83.
12. *García-Martín A, Ardura JA, Maycas M, Lozano D, López-Herradón A, Portal-Núñez S, García-Ocaña A, Esbrit P.* Functional roles of the nuclear localization signal of parathyroid hormone-related protein (PTHrP) in osteoblastic cells. *Mol Endocrinol.* 2014; **28**: 925–34.
13. *Lee SJ, Sekimoto T, Yamashita E, Nagoshi E, Nakagawa A, Imamoto N, Yoshimura M, Sakai H, Chong KT, Tsukihara T, Yoneda Y.* The structure of importin beta bound to SREBP 2: nuclear import of a transcription factor. *Science.* 2003; **302**(5650): 1571–5.
14. *Marfori M, Mynott A, Ellis JJ, Mehdi AM, Saunders NFW, Curmi PM, Forwood JK, Bodén M, Kobe B.* Molecular basis for specificity of nuclear import and prediction of nuclear localization. *BBA – Mol Cell Res.* 2011; **1813**(9): 1562–77.
15. *Wühr M, Güttler T, Peshkin L, McAlister GC, Sonnett M, Ishihara K, Groen AC, Presler M, Erickson BK, Mitchison TJ, Kirschner MW, Gygi SP.* The nuclear proteome of a vertebrate. *Curr Biol.* 2015; **25**(20): 2663–71.
16. *Nagarajan UM, Long AB, Harreman M, Corbett A, Boss JM.* A hierarchy of nuclear localization signals governs the import of the regulatory factor X complex subunits and MHC class II expression. *J Immunol.* 2004; **173**(24): 410–419.
17. *Chelsky D, Ralph R, Jonak G.* Sequence requirements for synthetic peptide-mediated translocation to the nucleus. *Mol Cell Biol.* 1989; **9**(6): 2487–92.
18. *Hodel MR, Corbett AH, Hodel AE.* Dissection of a nuclear localization signal. *J Biol Chem.* 2001; **276**(2): 1317–25.
19. *Fontes MRM, Teh T, Jan D, Brinkworth RI, Kobe B.* Structural basis for the specificity of bipartite nuclear localization sequence binding by importin-alpha. *J Biol Chem.* 2003; **278**(30): 27981–7.
20. *Kosugi S, Hasebe M, Entani T, Takayama S, Tomita M, Yanagawa H.* Design of peptide inhibitors for the importin α/β nuclear import pathway by activity-based profiling. *Chem Biol.* 2008; **15**(9): 940–9.
21. *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**(25): 10171–6.
22. *Robbins J, Dilworth SM, Laskey RA, Dingwall C.* Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: Identification of a class of bipartite nuclear targeting sequence. *Cell.* 1991; **64**(3): 615–23.
23. *Matthews BW.* Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Struct.* 1975; **405**(2): 442–51.
24. *Nakai K, Kanehisa M.* A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics.* 1992; **14**(4): 897–911.

25. *Brameier M, Krings A, MacCallum RM.* NucPred – predicting nuclear localization of proteins. *Bioinformatics 2007;* **23***(9): 1159–60.*
26. *Nguyen Ba AN, Pogoutse A, Provart N, Moses AM.* NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics.* 2009; **10**(1): 202.
27. *Mehdi AM, Sehgal MSB, Kobe B, Bailey TL, Bodén M.* A probabilistic model of nuclear import of proteins. *Bioinformatics.* 2011; **27**(9): 1239–46.
28. *Lin J, Hu J.* SeqNLS: nuclear localization signal prediction based on frequent pattern mining and linear motif scoring. *PLoS One.* 2013; **8**(10): e76864.

Порівняльний аналіз методів передбачення сигналів ядерної локалізації (NLS)

О. М. Лисиціна, В. Б. Сеплярський, Є. В. Шеваль

Мета. Ідентифікація сигналів ядерної локалізації (NLS) в амінокислотній послідовності білків за допомогою експериментальних методів залишається коштовним і тривалис процесом. Тому в останній час велику популярність отримали комп'ютерні методи прогнозування NLS. **Методи.** В даній статті ми провели порівняльний аналіз достовірності прогнозування NLS шести різних програм (PSORT II, NucPred, cNLSMapper, NLStradamus, NucImport та SeqNLS). Для кожного алгоритма було оцінена доля істинно позитивних прогнозів на вибірки з 155 експериментально визначених NLS з 128 білків людини, а також частку помилкових подій увибірці з 155 трансмембранних білків людини, які, як видно, позбавлені NLS. **Результати.** Найбільшу кількість вірнопрогнозованих NLS при найменшій частці хибнопозитивні результатів було отримано для трьох програм: NucPred, NLStradamus та seqNLS. **Висновки.** Однак навіть при набільшій ступені достовірності дані алгоритми прогнозують вірно не більше 45 % експериментально визначених NLS, тобто використання будь-яких алгоритмів прогнозування NLS вимагає експериментальної перевірки отриманих результатів.

Ключові слова: сигнал ядерної локалізації; передбачення.

Сравнительный анализ методов предсказания сигналов ядерной локализации (NLS)

О. М. Лисицына, В. Б. Сеплярский, Е. В. Шеваль

Цель. Идентификация сигналов ядерной локализации (NLS) в аминокислотной последовательности белка экспериментальными методами остается дорогостоящим и долгим процессом. Поэтому в последнее время большую популярность получили компьютерные методы предсказания NLS. **Методы.** В данной статье мы провели сравнительный анализ достоверности предсказания NLS шести различных программ (PSORT II, NucPred, cNLSMapper, NLStradamus, NucImport и SeqNLS). Для каждого алгоритма была оценена доля истинно положительных предсказаний на выборке из 155 экспериментально определенных NLS из 128 человеческих белков, а также доля ложноположительных предсказаний на выборке из 155 трансмембранных белков человека, которые, предположительно, лишены NLS. **Результаты.** Наибольшее количество правильно предсказанных NLS при наименьшей доле ложноположительных результатов было получено для трех программ: NucPred, NLStradamus и seqNLS. **Выводы.** Однако даже при наибольшей степени достоверности данные алгоритмы предсказывают правильно не более 45 % экспериментально определенных NLS, т.е. использование любых алгоритмов предсказания NLS требует экспериментальной проверки получаемых результатов.

Ключевые слова: сигнал ядерной локализации, предсказание.

Received 13.10.2016