

Sequence analysis

NucPred—Predicting nuclear localization of proteins

Markus Brameier^{1,*}, Andrea Krings² and Robert M. MacCallum^{3,*}

¹Bioinformatics Research Center (BiRC), University of Aarhus, 8000 Aarhus C, Denmark, ²Stockholm Bioinformatics Center (SBC), Stockholm University, 106 91 Stockholm, Sweden and ³Division of Cell & Molecular Biology, Imperial College London, South Kensington Campus, London, UK

Received on November 30, 2006; revised on February 9, 2007; accepted on February 19, 2007

Advance Access publication March 1, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: NucPred analyzes patterns in eukaryotic protein sequences and predicts if a protein spends at least some time in the nucleus or no time at all. Subcellular location of proteins represents functional information, which is important for understanding protein interactions, for the diagnosis of human diseases and for drug discovery. NucPred is a novel web tool based on regular expression matching and multiple program classifiers induced by genetic programming. A likelihood score is derived from the programs for each input sequence and each residue position. Different forms of visualization are provided to assist the detection of nuclear localization signals (NLSs). The NucPred server also provides access to additional sources of biological information (real and predicted) for a better validation and interpretation of results.

Availability: The web interface to the NucPred tool is provided at <http://www.sbc.su.se/~maccallr/nucpred>. In addition, the Perl code is made freely available under the GNU Public Licence (GPL) for simple incorporation into other tools and web servers.

Contact: brameier@birc.au.dk, r.maccallum@imperial.ac.uk

INTRODUCTION

Subcellular location provides at least some information about protein function and contributes to understanding protein interactions and signaling pathways in the cell. It is important for the identification of drug targets and may serve as an indicator of several diseases, like cancer and Alzheimer. Experimental determination of subcellular locations is often expensive and time-consuming. Instead, computational methods can make fast and accurate predictions. In recent years, several bioinformatics tools have been developed to identify different kinds of subcellular compartment(s) (Emanuelsson *et al.*, 2000; Nair and Rost, 2005; Nakai and Horton, 1999; Reinhardt and Hubbard, 1998).

An important compartment in eukaryotic cells is the nucleus, where many essential biological processes take place. Because proteins are manufactured outside the nucleus, they must get back into the nucleus if they are needed there. This import is mediated by short binding sites on the protein sequence, called nuclear localization signals (NLSs). NLSs are not yet well understood, and so the set of currently known NLSs may be incomplete.

Only few tools for predicting nuclear localization of proteins make use of sequence motifs to match certain sequence patterns. Cokol *et al.* (2000) identify relevant substructures of known NLSs and use these for building more general motifs. Our approach to nuclear localization applies genetic programming (GP), a machine-learning technique that automatically develops computer programs in an artificial evolutionary process (Koza, 1992). Our evolved predictors incorporate multiple regular expressions which are matched against the input (amino acid) sequence. These sequence motifs are evolved together with the actual classification rules. One important advantage of NucPred is that it is not restricted to a predefined set of NLS patterns and, thus, has the potential to discover new, unknown NLSs. In fact, the program classifiers were trained without using knowledge about known NLSs, but only about the nuclear/non-nuclear location of protein sequences.

Other approaches train predictors with *global* information, like the amino acid composition (Nair and Rost, 2005; Reinhardt and Hubbard, 1998) or provide this in addition to the *local* sequence motifs (Nakai and Horton, 1999).

The method behind NucPred has earlier been reported to be competitive with existing methods for nuclear localization of proteins (Heddad *et al.*, 2004), including PredictNLS (Cokol *et al.*, 2000) and PSORT II (Nakai and Horton, 1999). Hwang *et al.* (2006) used all three computational approaches for identifying the nuclear proteome in human T leukemia cells, together with experimental methods. The NucPred predictor has also been incorporated into other bioinformatics web services, such as the POGs/PlantRBP database (<http://plantrbp.uoregon.edu>).

Table 1 compares NucPred with state-of-the-art approaches to nuclear localization, including more recent methods like LOCtree (Nair and Rost, 2005) and BaCelLo (Pierleoni *et al.*, 2006). The test dataset used here contains all human protein sequences—not predicted to be transmembrane by TMHMM—for which there is at least one TAS evidenced Gene Ontology (GO) annotation for cellular component in UniProt GOA. Positive test cases have GO term ‘nucleus’, ‘chromosome’, or one of their child terms assigned. The performance of NucPred, PredictNLS and PSORT II are roughly equivalent, but at different sensitivities. Conjunction of predictions from different methods, e.g. NucPred and PredictNLS, gives higher specificity, while disjunction increases sensitivity.

Second generation predictors, like LOCtree and BaCelLo, achieve a higher sensitivity by a hierarchical architecture of

*To whom correspondence should be addressed.

Table 1. Comparison of different nuclear localization methods

Method	Specificity	Sensitivity
NucPred (0.8 threshold)	0.615	0.307
NucPred (0.5 threshold)	0.480	0.626
PredictNLS	0.625	0.230
PSORT II	0.466	0.697
NucPred(0.8) AND PredictNLS	0.726	0.166
NucPred(0.8) OR PredictNLS	0.571	0.432
LOCtree (<i>ab initio</i>)	0.587	0.633
BaCelLo	0.668	0.614

different support vector machines (SVM) classifiers which mimics biological pathways. Each such classifier discerns another class of subcellular compartment. Both methods make use of evolutionary information from sequence profiles, besides sequence composition. LOCtree further improved coverage by training on additional proteins whose subcellular location was predicted by sequence homology or keyword similarity.

METHOD AND APPLICATIONS

The NucPred web server provides three major functionalities for predicting nuclear localization. A likelihood score is calculated either (1) for up to 1000 input sequences (in batch mode), (2) for each sequence in a multiple alignment or (3) over whole proteomes.

The NucPred core is an ensemble (or jury) of 100 sequence-based predictors. Each makes a Boolean (yes/no) decision to whether a protein (sequence) has a nuclear role or not. To improve accuracy, the individual predictions are combined by a majority voting scheme. The consensus score is basically the fraction of predictors which vote ‘nuclear’. For a discrete prediction, the user has to decide on a score threshold (preferably 0.8). Sequences which score greater than or equal to this threshold are predicted to spend time in the nucleus.

For single sequences, NucPred calculates a per-residue scoring. Each amino acid location is colored according to its influence on a ‘nuclear’ classification (see example in Fig. 1). The closer the color of a subsequence lies at the red (blue) end of the color spectrum, the more positive (negative) is its effect. Unlike black-box predictors, such as neural networks (Reinhardt and Hubbard, 1998) or support vector machines (SVMs, Nair and Rost, 2005), the classifying programs in NucPred may be interpreted in terms of the regular expressions contained. For each such sequence motif, a fixed—positive or negative—degree of influence has been pre-calculated over all training sequences (selected from UniProt). The per-residue coloring reflects the sum of the influences of all regular expressions that match at the particular position in the sequence, highlighting potential regions for experimental verification or manipulation.

NucPred offers the possibility to project the per-residue coloring onto the 3D protein structure found in the protein data bank (PDB <http://www.rcsb.org/pdb>). The colored structure (modified PDB file) may be viewed using Rasmol (<http://www.openrasmol.org>) or another PDB viewer. Such an analysis may reveal the spatial arrangement of putative or known NLSs.

Another NucPred option allows the upload of multiple sequences from the same protein family to be aligned by the



Fig. 1. >Protein sequence of human gene AUTS2 (UniProt ID Q8WXX7) colored with per-residue NucPred scores (monochrome version). The overall sequence score is 1. First 500 out of 1259 residues shown.

ClustalW algorithm. Each sequence is colored (individually) according to the per-residue NucPred scores. This enables the comparative analysis of homologs from the same or different species, and may be useful for the study of splice variants.

NucPred also provides pre-calculated sequence scores over full proteomes of different eukaryotes—including human, mouse, rat, fish, fly, worm and yeast—and for all 98 716 eukaryotic proteins in Uniprot (release 8.5 <http://www.ebi.uniprot.org>). To increase confidence, the NucPred predictions may be compared to and combined with the discrete predictions from PredictNLS and PSORT II. TMHMM (Krogh *et al.*, 2001) predictions of transmembrane proteins are provided to potentially rule out a nuclear location.

In addition, Uniprot annotations for subcellular location and GO annotations for cellular compartment are given if available. This information allows, for instance, the discovery of proteins which are associated with human diseases and are confidently predicted to have a nuclear function, but are not currently known to be nuclear. The protein from Fig. 1 is believed to increase the susceptibility for Autism and is predicted to be nuclear by all three methods. A further interesting example of 94 human proteins meeting these criteria is oncogene ETC2 (UniProt ID Q9H8V3).

Conflict of Interest: none declared.

REFERENCES

Cokol,M. *et al.* (2000) Finding nuclear localization signals. *EMBO Rep.*, **1**, 411–415.
Emanuelsson,O. *et al.* (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
Heddad,A. *et al.* (2004) Evolving regular expression-based sequence classifiers for protein nuclear localization. In *Proc. of EvoBIO 2004 Conf., Lect. Notes Comput. Sci.*, vol. 3005, pp. 31–40.
Hwang,S.I. *et al.* (2006) Systematic characterization of nuclear proteome during apoptosis: A quantitative proteomic study by differential extraction and stable isotope labeling. *Mol. Cell Proteomics*, **5**, 1131–1145.
Koza,J.R. (1992) *Genetic Programming: On the Programming of Computer Programs by Natural Selection*. MIT Press, Cambridge, MA.
Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.*, **348**, 85–100.
Nakai,K. and Horton,P. (1999) PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.
Pierleoni,A. *et al.* (2006) BaCelLo: A balanced subcellular localization predictor. *Bioinformatics*, **22**, 408–416.
Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, **26**, 2230–2236.