BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 17 2014, pages 2501–2502
doi:10.1093/bioinformatics/btu310

Sequence analysis

Advance Access publication May 13, 2014

PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition

Alex K. Lancaster[1,2,3], Andrew Nutter-Upham[1], Susan Lindquist[1,4,5,*] and Oliver D. King[6,*]

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, [2]Department of Pathology, Beth Israel Deaconess Medical Center, [3]Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115, USA, [4]Department of Biology, [5]Howard Hughes Medical Institute, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139 and [6]Department of Cell and Developmental Biology, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, MA 01655, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Prions are self-templating protein aggregates that stably perpetuate distinct biological states and are of keen interest to researchers in both evolutionary and biomedical science. The best understood prions are from yeast and have a prion-forming domain with strongly biased amino acid composition, most notably enriched for Q or N. PLAAC is a web application that scans protein sequences for domains with **p**rion-**l**ike **a**mino **a**cid **c**omposition. Users can upload sequence files, or paste sequences directly into a textbox. PLAAC ranks the input sequences by several summary scores and allows scores along sequences to be visualized. Text output files can be downloaded for further analyses, and visualizations saved in PDF and PNG formats.

Availability and implementation: http://plaac.wi.mit.edu/. The Ruby-based web framework and the command-line software (implemented in Java, with visualization routines in R) are available at http://github.com/whitehead/plaac under the MIT license. All software can be run under OS X, Windows and Unix.

Contact: oliver.king@umassmed.edu or lindquist_admin@wi.mit.edu

Received on January 10, 2014; revised on April 1, 2014; accepted on April 28, 2014

1 INTRODUCTION

Prions are proteins that can switch from non-aggregated states to self-templating highly ordered aggregates. This property allows them to confer stable changes in biological states that are of great interest in molecular and evolutionary biology (Newby and Lindquist, 2013). For example, they create neurodegenerative diseases, perpetuate activity states in neural synapses and provide access to a broad realm of phenotypic diversification in microbes. The ability to identify potential prion-like proteins from sequence data would speed the search for new prions across a wide variety of taxa. We previously developed (Alberti *et al*., 2009) a hidden Markov model (HMM) to identify candidate prions and parse these candidates into prion-like domains (PrLDs) and non-PrLDs, on the basis of amino acid (AA) composition. Briefly, the HMM has two hidden states, for *PrLD* and *background*, and the output symbols are the 20 AAs. The output probabilities for the PrLD state were constructed based on the AA frequencies in the PrLDs of four prions of *Saccharomyces cerevisiae* that were known at the time. This algorithm and extensions have since been used in several studies to identify prion-like sequences in yeast (Holmes *et al*., 2013) and also in humans (Kim *et al*., 2013; King *et al*., 2012), in which several proteins with PrLDs are associated with ALS and related neurodegenerative disorders. Here we describe a web-based front end to the prion-prediction algorithm, PLAAC, and give an overview of implementation and extensions; further details are provided on the PLAAC Web site.

*To whom correspondence should be addressed.

2 FEATURES AND METHODS

PLAAC supports the scanning of single protein sequences for potential PrLDs, as well as the scanning of whole proteomes. The user can specify a minimum length for prion domains (set by a textbox, by default $L_{\text{core}} = 60$), and can optionally use organism-specific background AA frequencies in the HMM instead of the default *S.cerevisiae* background frequencies. These frequencies can be computed from the uploaded sequences, or selected from precomputed organism-specific frequencies (set by a dropdown list). A parameter α (set via a slider) allows continuous interpolation between organism-specific background frequencies ($\alpha = 0$) and *S.cerevisiae* background frequencies ($\alpha = 1$). We have used $\alpha = 0.5$ when scanning other species, reflecting our uncertainty in the degree to which the corresponding PrLD AA frequencies are skewed toward *S.cerevisiae* background frequencies (as opposed to being species-independent).

Resulting output including per-protein summary tables and per-residue tables for selected proteins can be downloaded as text files. Visualizations can also be downloaded as PNG or PDF files (Fig. 1). The command-line program allows additional control over plots [which tracks to display, and whether to show sliding averages of per-residue scores (Alberti *et al*., 2009) or sliding averages of these sliding averages (Kim *et al*., 2013)].

Single sequence: To search for PrLDs in a single sequence, the user pastes into a textbox or uploads the protein sequence, either in FASTA format or as bare sequence, and may modify the L_{core} and α parameters, if desired. After submission, scores (including `COREscore`, `LLR` and `PAPA` scores described below) for the sequence are displayed along with a graphical visualization of the location, if any, of predicted PrLDs.

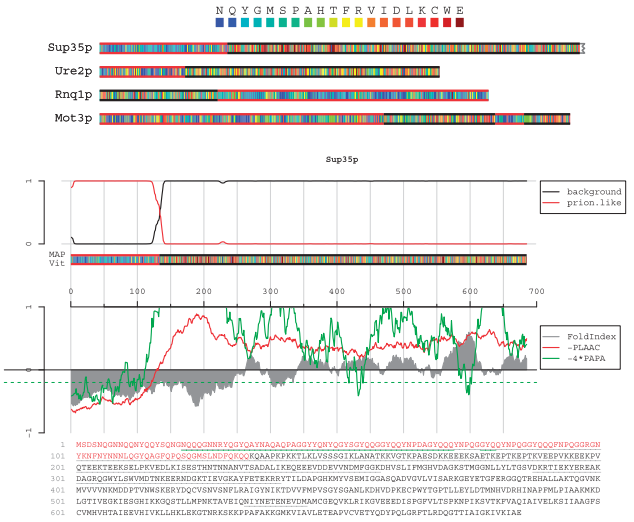

Fig. 1. Visualization outputs from PLAAC. Top: four known yeast prion proteins with each AA color-coded by its enrichment log-likelihood ratio in PrLDs (styled after the Sequence Enrichment Visualization Tool; http://jura.wi.mit.edu/cgi-bin/bio/draw_enrichment.pl), with HMM parse indicated by outer bars. Bottom: detailed visualization of the Sup35 protein, including several prion-prediction scores discussed in the main text

Multiple sequences and whole proteomes: To scan multiple protein sequences (including whole proteomes), the user again pastes or uploads them in FASTA format. (Upload is recommended for more than a few sequences.) L_{core} and α parameters may be adjusted, and background frequencies may be computed directly from the provided sequences (not recommended if uploading or pasting just a few sequences). The user is presented with a summary table with a row for each uploaded protein ranked by `COREscore` from highest to lowest and then may select candidates in this summary list to generate plots for further visualization.

Output ranking and scores: Multiple sequences are ranked for prion-like properties by the `COREscore` metric (Alberti *et al.*, 2009), which is the maximum sum of per-residue log-likelihood ratios for any subsequence of length L_{core} that falls entirely within the PrLD state in the HMM Viterbi parse, provided a sequence of this length exists, and is undefined (NaN) otherwise. In addition, we compute a score `LLR` (for log-likelihood ratio) that is otherwise identical to `COREscore`, but without the requirement that the sequence falls entirely within the PrLD state of the HMM parse. Because `LLR` does not impose a hard cutoff, it can be useful when doing exploratory whole-proteome analyses, e.g. on the overall distribution of (near) PrLDs. However, examining whether the region with the highest `LLR` score falls entirely within the PrLD state in the HMM parse may be informative, e.g. when selecting domains to clone for studies of candidate PrLDs fused to reporter proteins.

Algorithm updates: Since the publication of (Alberti *et al.*, 2009), the AA frequencies for the PrLD state of the HMM have been updated based on 28 candidate PrLDs that showed switching behavior or strong amyloid formation experimentally (see source code for details). These updated AA frequencies were used in later publications (Holmes *et al.*, 2013; Kim *et al.*, 2013; King *et al.*, 2012).

A subsequent algorithm called PAPA (Toombs *et al.*, 2010, 2012) that uses AA scores derived from a random mutagenesis screen can downweigh many of the apparent false positives from Alberti *et al.* (2009), and can give sharper predictions for the results of point mutations (Kim *et al.*, 2013). It appears that a small number of hydrophobic residues can speed amyloid formation in regions otherwise highly enriched for polar uncharged residues such as Q and N (Toombs *et al.*, 2010). PLAAC and PAPA are complementary, as PLAAC identifies such regions, and PAPA has been validated only on such regions. (Single scores based on local averages of per-residue AA scores do not adequately capture the trade-off between hydrophobic and polar uncharged residues.) We reimplemented PAPA, and included this score in the output and visualizations, along with predictions of intrinsically unfolded protein regions from a reimplementation of FoldIndex (Prilusky *et al.*, 2005). It is also important to note that there are several known prions that are not strongly Q/N-rich (e.g. het-S, PrP, Mod5), but as systematic experimental screening for prion-like propagation is lacking for non-Q/N-rich proteins, it is difficult to estimate the false-negative rates of these algorithms.

PLAAC has been developed as a web application to allow users to scan single protein sequences as well as whole proteomes for the presence of PrLDs. We have also augmented the original algorithm with additional scores, making unified comparisons possible.

ACKNOWLEDGEMENTS

We thank R. Halfmann, S. Alberti, J. Shorter, A. Gitler, J.P. Taylor, the Lindquist Lab, S. McCallum of the Information Technology and F. Lewitter of the Bioinformatics and Research Computing (BaRC) groups at the Whitehead Institute for additional help and advice. R. Latek and K. Walker, former members of BaRC, developed the Sequence Enrichment Visualization Tool.

Funding: This work was supported by the G. Harold and Leila Y. Mathers Foundation [to S.L.]; Howard Hughes Medical Institute [to S.L.]; and the National Institutes of Health [GM025874 to S.L.].

Conflicts of Interest: none declared.

REFERENCES

Alberti,S. *et al.* (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell*, **137**, 146–158.

Holmes,D.L. *et al.* (2013) Heritable remodeling of yeast multicellularity by an environmentally responsive prion. *Cell*, **153**, 153–165.

Kim,H.J. *et al.* (2013) Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature*, **495**, 467–473.

King,O.D. *et al.* (2012) The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. *Brain Res.*, **1462**, 61–80.

Newby,G.A. and Lindquist,S. (2013) Blessings in disguise: biological benefits of prion-like mechanisms. *Trends Cell Biol.*, **23**, 251–259.

Prilusky,J. *et al.* (2005) FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.

Toombs,J.A. *et al.* (2010) Compositional determinants of prion formation in yeast. *Mol. Cell. Biol*, **30**, 319–332.

Toombs,J.A. *et al.* (2012) *De novo* design of synthetic prion domains. *Proc. Natl Acad. Sci. USA*, **109**, 6519–6524.