

A peer-reviewed version of this preprint was published in PeerJ on 3 February 2016.

[View the peer-reviewed version](https://doi.org/10.7717/peerj-cs.45) (peerj.com/articles/cs-45), which is the preferred citable publication unless you specifically need to cite this preprint.

Habich M, Djuranovic S, Szczesny P. 2016. PATACSDB—the database of polyA translational attenuators in coding sequences. PeerJ Computer Science 2:e45 <https://doi.org/10.7717/peerj-cs.45>

PATACSDB - The database of polyA translational attenuators in coding sequences

Malgorzata Habich, Sergej Djuranovic, Pawel Szczesny

Recent addition to the repertoire of gene expression regulatory mechanisms are polyadenylate (polyA) tracks encoding for poly-lysine runs in protein sequences. Such tracks stall translation apparatus and induce frameshifting independently of the effects of charged nascent poly-lysine sequence on the ribosome exit channel. As such they substantially influence the stability of mRNA and amount of protein produced from a given transcript. Single base changes in these regions are enough to exert a measurable response on both protein and mRNA abundance, and makes each of these sequences potentially interesting case studies for effects of synonymous mutation, gene dosage balance and natural frameshifting. Here we present the PATACSDB, a resource that contain comprehensive list of polyA tracks from over 250 eukaryotic genomes. Our data is based on Ensembl genomic database of coding sequences and filtered with algorithm of 12A-1 which selects sequences of polyA tracks with a minimal length of 12 A's allowing for one mismatched base. The PATACSDB database is accesible at: <http://sysbio.ibb.waw.pl/patacsdb>. Source code is available for download from GitHub repository at <http://github.com/habich/PATACSDB>, including the scripts to recreate the database from the scratch on user's own computer.

PATACSDB - The Database of PolyA Translational Attenuators in Coding Sequences

4

5 Malgorzata Habich¹, Sergej Djuranovic², Pawel Szczesny^{1,3#}

6

7 1 Institute of Biochemistry and Biophysics Polish Academy of Sciences, Department of Bioinformatics,
8 Pawinskiego 5a, 02-106 Warsaw, Poland

9 2 Washington University School of Medicine, Department of Cell Biology and Physiology, St. Louis,
10 MO63110, USA

11 3 Faculty of Biology, Institute of Experimental Plant Biology and Biotechnology, University of Warsaw,
12 ul. Miecznikowa 1, 02-096 Warsaw, Poland

13

14

15 # Correspondence should be addressed to:

16 Pawel Szczesny

17 Department of Bioinformatics

18 Institute of Biochemistry and Biophysics Polish Academy of Sciences

19 ul. Pawinskiego 5A

20 02-106 Warsaw

21 Poland

22 Email: szczesny@ibb.waw.pl

23

24 Abstract

25 Recent addition to the repertoire of gene expression regulatory mechanisms are polyadenylate
26 (polyA) tracks encoding for poly-lysine runs in protein sequences. Such tracks stall translation
27 apparatus and induce frameshifting independently of the effects of charged nascent poly-lysine
28 sequence on the ribosome exit channel. As such they substantially influence the stability of
29 mRNA and amount of protein produced from a given transcript. Single base changes in these
30 regions are enough to exert a measurable response on both protein and mRNA abundance, and
31 makes each of these sequences potentially interesting case studies for effects of synonymous
32 mutation, gene dosage balance and natural frameshifting. Here we present the PATACSDb, a
33 resource that contain comprehensive list of polyA tracks from over 250 eukaryotic genomes. Our
34 data is based on Ensembl genomic database of coding sequences and filtered with algorithm of
35 12A-1 which selects sequences of polyA tracks with a minimal length of 12 A's allowing for one
36 mismatched base. The PATACSDb database is accesible at: <http://sysbio.ibb.waw.pl/patacsdb>.
37 Source code is available for download from GitHub repository at
38 <http://github.com/habich/PATACSDb>, including the scripts to recreate the database from the
39 scratch on user's own computer.

40 Background

41 The classical view of the genetic information flow inside living cells, that is transcription from
42 DNA to RNA and finally translation of mRNA into protein, is of course a bit simplistic. Over

decades of research we keep accumulating evidences of several control points at different levels of these processes. The first studies were focused on transcriptional regulation, but more recently regulation of gene expression at the level of translation gathered researchers' attention. Translational regulation generally controls the amount of protein synthesised from a given mRNA through several mechanisms, targeting recruitment of ribosomes to the transcript, elongation speed, termination and as a proxy to all these processes mRNA stability. Ribosome stalling, that is pausing of ribosome during translational cycle, is recognized by components of several mRNA surveillance pathways. As a result of impeded rate of ribosome along the mRNA, the transcript is endonucleolytically cleaved and nascent albeit incomplete protein product is degraded by proteasome (Shoemaker & Green, 2012). Over the years we have got to know that certain sequence features can trigger ribosome stalling. These are damaged bases (Cruz-Vera et al., 2004), stable stem-loop structures (Doma & Parker, 2006), rare codons (Letzring, Dean & Grayhack, 2010), mRNAs lacking stop codons (so called non-stop mRNAs) (Dimitrova et al., 2009), runs of codons that encode consecutive basic aminoacids (Kuroha et al., 2010; Brandman et al., 2012), or finally, runs of adenines encoding poly-lysine tracks (Koutmou et al., 2015; Arthur et al., 2015).

We have recently shown that polyA tracks trigger a response in a different manner than runs of basic aminoacids (Arthur et al., 2015). In addition to stalling, occasionally they lead to ribosome sliding on mRNA transcript which results in production of additional frameshifted product next to the known and well annotated gene protein product. As such polyA track sequences may support programed translational frameshifts in such mRNA transcripts giving rise to alternative protein products from those genes. This feature of polyA track genes resembles programmed frameshifting observed in viral genes with slippery sequences however without a need for

additional mRNA structures that induces ribosome stalling in known viral transcripts (Chen et al., 2014; Yan et al., 2015). The ultimate control over the production and stability of alternative transcripts from polyA track genes in Eukaryotes would be based on mRNA surveillance mechanisms, mainly non-sense mediated mRNA decay (NMD) or if the kinetic stall persists by no-go mRNA decay (NGD). PolyA tracks are highly conserved in genes among Eukaryotes and it is likely that they represent a universal translational attenuators or programed translational frameshift signals. Intrinsically this novel RNA motif plays an important role in balancing gene dosage and homeostasis of cellular environment. The level of attenuation, frameshifting and exact role of polyA tracks in organisms homeostasis is still to be elucidated.

PATACSDb server

Given that no genomic database reports polyA tracks in coding sequences, we have designed PATACSDb (PolyA Translational Attenuators in Coding Sequences DataBase), a resource devoted to collection of such features among eukaryotic organisms. In concordance with our experimental data from the controlled expression of reporter sequences or natural gene expression profiles we have designed a 12A-1 pattern, that is pattern of twelve adenines in coding region allowing for one mismatch. This pattern should result in reduction of expression by roughly 30%, a magnitude that can potentially have a measurable biological impact (Arthur et al., 2015). We have analyzed eukaryotic Ensembl genomes (Flicek et al., 2014) for the presence of this pattern in coding sequences, using only these entries for which coding sequence matched reported translated sequence. This was done not only on standard Ensembl genomes but its additional eukaryotic databases like Ensembl Protists and Ensembl Metazoa. As a result, we have identified 197964 genes in 254 genomes that carry 446206 polyA tracks.

The main table consists of protein common name, gene and transcripts Ensembl ids, location of the polyA track expressed as percentage (allows for quick identification of cases where polyA track is either at the end or at the beginning of the protein) and finally, the identified polyA track with a context of surrounding sequence. All columns are sortable. By default, the table is sorted by protein name, alphabetically. Sorting gene and transcript ids is also alphabetical. Location is sorted numerically. The rows with polyA sequences is sortable by polyA track length, so the user can quickly identify sequences with the longest track in particular organism. Obviously, due to used pattern, the shortest polyA tracks have length of 12 nucleotides. To facilitate quick interaction with tables, we have used Bootstrap-table library that allows for easy and intuitive sorting and searching through all fields in particular genome.

Project was created using Python 2.7. To parse biological data we used Biopython 1.65. To compare protein and cdna sequences we used local version of NCBI blast+ software v. 2.2.31.

To run the web service we used Flask v.0.10.1 . We used SQLite3 database engine and SQLAlchemy for database access. To query Ensembl database we used mysql client. We also used two other Python libraries: *xmltodict* and *requests*. The most difficult task was to ensure short page load times given the large dataset we worked on. To solve this problem we have created additional tables in database which contain metadata with the heaviest queries. This solution decreased time of loading more than 20 times.

We have designed two step architecture. In the first step we analyse data from Ensembl database and create our database with 12A-1 pattern. In the second step we use created database to provide information to web service. This architecture allows to separate obtaining data and running web service thus during analysis of new version of Ensembl data we still can provide data about old version, and change between versions can be done in seconds without user

noticing. In the future we will work on parallelization process of Ensembl data analysis to speedup first step. It is likely that polyA segments are not the only sequence determinants of translation efficiency in coding sequences and future studies will discover more of such motifs. Design of the PATACSDB engine allows for easy modification towards finding and cataloguing of novel sequence patterns.

References

- Arthur L., Pavlovic-Djuranovic S., Smith-Koutmou K., Green R., Szczesny P., Djuranovic S. 2015. Translational control by lysine-encoding A-rich sequences. *Science advances* 1.
- Brandman O., Stewart-Ornstein J., Wong D., Larson A., Williams CC., Li G-W., Zhou S., King D., Shen PS., Weibezahn J., Dunn JG., Rouskin S., Inada T., Frost A., Weissman JS. 2012. A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* 151:1042–1054.
- Chen J., Petrov A., Johansson M., Tsai A., O’Leary SE., Puglisi J. 2014. Dynamic pathways of -1 translational frameshifting. *Nature* 512:328–332.
- Cruz-Vera LR., Mago-Castro MA., Zamora-Romo E., Guarneros G. 2004. Ribosome stalling and peptidyl-tRNA drop-off during translational delay at AGA codons. *Nucleic acids research* 32:4462–4468.
- Dimitrova LN., Kuroha K., Tatematsu T., Inada T. 2009. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *The Journal of biological chemistry* 284:10343–10352.
- Doma MK., Parker R. 2006. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature* 440:561–564.
- Flicek P., Amodè MR., Barrell D., Beal K., Billis K., Brent S., Carvalho-Silva D., Clapham P., Coates G., Fitzgerald S., Gil L., Girón CG., Gordon L., Hourlier T., Hunt S., Johnson N., Juettemann T., Kähäri

- AK., Keenan S., Kulesha E., Martin FJ., Maurel T., McLaren WM., Murphy DN., Nag R., Overduin
B., Pignatelli M., Pritchard B., Pritchard E., Riat HS., Ruffier M., Sheppard D., Taylor K., Thormann
A., Trevanion SJ., Vullo A., Wilder SP., Wilson M., Zadissa A., Aken BL., Birney E., Cunningham
F., Harrow J., Herrero J., Hubbard TJP., Kinsella R., Muffato M., Parker A., Spudich G., Yates A.,
Zerbino DR., Searle SMJ. 2014. Ensembl 2014. *Nucleic acids research* 42:D749–55.
- Koutmou KS., Schuller AP., Brunelle JL., Radhakrishnan A., Djuranovic S., Green R. 2015. Ribosomes
slide on lysine-encoding homopolymeric A stretches. *eLife* 4.
- Kuroha K., Akamatsu M., Dimitrova L., Ito T., Kato Y., Shirahige K., Inada T. 2010. Receptor for
activated C kinase 1 stimulates nascent polypeptide-dependent translation arrest. *EMBO reports*
11:956–961.
- Letzring DP., Dean KM., Grayhack EJ. 2010. Control of translation efficiency in yeast by codon-
anticodon interactions. *RNA* 16:2516–2528.
- Shoemaker CJ., Green R. 2012. Translation drives mRNA quality control. *Nature structural & molecular
biology* 19:594–601.
- Yan S., Wen J-D., Bustamante C., Tinoco I Jr. 2015. Ribosome excursions during mRNA translocation
mediate broad branching of frameshift pathways. *Cell* 160:870–881.