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 scratch on the user's own computer.

1 PATACSDB - The Database of PolyA

2 Translational Attenuators in Coding

3 Sequences

5 Malgorzata Habich[1], Sergej Djuranovic[2], Pawel Szczesny[1,3#]

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

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.

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 a subject of continous

43   modification for both, direction of the flow and the number of players involved. Over decades of

44   research we keep accumulating evidences of several control points at different levels of these

45   processes. The past studies were focused on transcriptional regulation, but more recently

46   regulation of gene expression at the level of translation drew researchers' attention. Translational

47   regulation generally controls the amount of protein synthesised from a given mRNA through

48   several mechanisms, targeting recruitment of ribosomes to the transcript, elongation speed,

49   termination and as a proxy to all these processes mRNA stability. Ribosome stalling, that is

50   pausing of ribosome during translational cycle, is recognized by components of several mRNA

51   surveillance pathways. As a result of impeded rate of ribosome along the mRNA, the transcript is

52   endonucleolytically cleaved and nascent albeit incomplete protein product is degraded by

53   proteasome (Shoemaker & Green, 2012). Over the years we have got to know that certain

54   sequence features can trigger ribosome stalling. These are damaged bases (Cruz-Vera et al.,

55   2004), stable stem-loop structures (Doma & Parker, 2006), rare codons (Letzring, Dean &

56   Grayhack, 2010), mRNAs lacking stop codons (so called non-stop mRNAs) (Dimitrova et al.,

57   2009), runs of codons that encode consecutive basic aminoacids (Kuroha et al., 2010; Brandman

58   et al., 2012), or finally, runs of adenines encoding poly-lysine tracks (Koutmou et al., 2015;

59   Arthur et al., 2015).

60   We have recently shown that polyA tracks trigger a response in a different manner than runs of

61   basic aminoacids (Arthur et al., 2015). In addition to stalling, occasionally they lead to ribosome

62   sliding on mRNA transcript which results in production of additional frameshifted product next

63   to the known and well annotated gene protein product. As such polyA track sequences may

64   support programed translational frameshifts in such mRNA transcripts giving rise to alternative

65   protein products from those genes. This feature of polyA track genes resembles programmed

66 frameshifting observed in viral genes with slippery sequences however without a need for

67 additional mRNA structures that induces ribosome stalling in known viral transcripts (Chen et

68 al., 2014; Yan et al., 2015). The ultimate control over the production and stability of alternative

69 transcripts from polyA track genes in Eukaryotes would be based on mRNA surveillance

70 mechanisms, mainly non-sense mediated mRNA decay (NMD) or if the kinetic stall persists by

71 no-go mRNA decay (NGD). PolyA tracks are highly conserved in genes among Eukaryotes and

72 it is likely that they represent a universal translational attenuators or programed translational

73 frameshift signals. Intrinsically this novel RNA motif plays an important role in balancing gene

74 dosage and homeostasis of cellular environment. The level of attenuation, frameshifting and

75 exact role of polyA tracks in organisms homeostasis is still to be elucidated.

76 ## PATACSDB server

77 While there are several resources devoted to polyadenylation signals in genomic sequences,

78 these have different sequence signature and refer to the processing of mRNA, not translation. No

79 genomic database reports polyA tracks in coding sequences, therefore we have designed

80 PATACSDB (PolyA Translational Attenuators in Coding Sequences DataBase), a resource

81 devoted to collection of such features among eukaryotic organisms. In concordance with our

82 experimental data from the controlled expression of reporter sequences or natural gene

83 expression profiles we have designed a 12A-1 pattern, that is pattern of twelve adenines in

84 coding region allowing for one mismatch. Based on our experiments, this is a minimal pattern

85 that should result in reduction of expression by roughly 30%, a magnitude that can potentially

86 have a measurable biological impact in human cells (Arthur et al., 2015). We have extrapolated

87 this pattern to other organisms, because without further experimental work we have no way to

88 define the minimal polyA pattern in other organisms. We have analyzed eukaryotic Ensembl

89 genomes (Flicek et al., 2014) for the presence of this pattern in coding sequences, using only

90 these entries for which coding sequence matched reported translated sequence. This was done

91 not only on standard Ensembl genomes but its additional eukaryotic databases like Ensembl

92 Protists and Ensembl Metazoa. As a result, we have identified 197964 genes in 254 genomes that

93 carry 446206 polyA tracks.

PolyA tracks across eukaryotic organisms

95 In the previous studies (Koutmou et al., 2015; Arthur et al., 2015) we focused mainly on polyA

96 tracks from human and yeast genomes, using NCBI (Pruitt et al., 2014) database and SGD

97 (Cherry et al., 1998) as data sources, respectively. Overall there is a good agreement between our

98 previous analysis and this study for high eukaryotes, while we see some discrepancies for lower

99 eukaryotes, such as yeast. For example, in the previous study we have underestimated the

100 number of polyA-carrying genes in yeast by an order of magnitude (29 vs 369) - a result of

101 different data source.

102 The percentage of polyA carrying transcripts varies from organism to organism and exceeds 60%

103 for *Plasmodium* species, well known for their AT-rich genome (see Table 1 for summary).

104 However, the distribution of lengths of polyA tracks is quite similar across whole observed

105 spectrum of AT-content (Fig.1). It might be that the single *Plasmodium* genus is skewing the

106 distribution, as the species distribution of genomic databases is heavily biased. In human, around

107 only around 1% of transcripts coming from ca. 2% of genes carry polyA track and as such, are

108 subjects of translational attenuation. This is close to a median across all analyzed genomes.

109 Furthermore, we did not find any correlation between organismal complexity and number of

110   polyA-affected genes. This might indicate that such feature is a constituent element of

111   translational machinery, unrelated to external factors and regulatory mechanisms.

112   Software architecture

113   The main table consists of protein common name, gene and transcripts Ensembl ids, location of

114   the polyA track expressed as percentage (allows for quick identification of cases where polyA

115   track is either at the end or at the beginning of the protein) and finally, the identified polyA track

116   with a context of surrounding sequence. All columns are sortable. By default, the table is sorted

117   by protein name, alphabetically. Sorting gene and transcript ids is also alphabetical. Location is

118   sorted numerically. The rows with polyA sequences is sortable by polyA track length, so the user

119   can quickly identify sequences with the longest track in particular organism. Obviously, due to

120   used pattern, the shortest polyA tracks have length of 12 nucleotides. To facilitate quick

121   interaction with tables, we have used Bootstrap-table library that allows for easy and intuitive

122   sorting and searching through all fields in particular genome.

123   Project was created using Python 2.7. To parse biological data we used Biopython 1.65.

124   To compare protein and cdna sequences we used local version of NCBI blast+ software v.

125   2.2.31. To run the web service we used Flask v.0.10.1 . We used SQLite3 database engine and

126   SQLAlchemy for database access. To query Ensembl database we used mysql client. We also

127   used two other Python libraries: *xmltodict* and *requests*. The most difficult task was to ensure

128   short page load times given the large dataset we worked on. To solve this problem we have

129   created additional tables in database which contain metadata with the heaviest queries. This

130   solution decreased time of loading more than 20 times.

131   We have designed two step architecture. In the first step we analyse data from Ensembl

132   database and create our database with 12A-1 pattern. In the second step we use created database

133   to provide information to web service. This architecture allows to separate obtaining data and

134   running web service thus during analysis of new version of Ensembl data we still can provide

135   data about old version, and change between versions can be done in seconds without user

136   noticing. In the future we will work on parallelization process of Ensembl data analysis to

137   speedup first step. It is likely that polyA segments are not the only sequence determinants of

138   translation efficiency in coding sequences and further studies will discover more of such motifs

139   or different lengths of minimal polyA pattern for a particular organism. Design of the

140   PATACSDB engine allows for easy modification towards finding and cataloguing of novel

141   sequence patterns.

142

143

144

145

146

147   Table 1. Summary of the content of PATACSDB

Feature	Value
Total number of polyA-carrying transcripts	197964
Highest percentage of polyA-carrying transcripts (first 5)	Plasmodium berghei 68.259% Plasmodium yoelii 17x 64.957% Plasmodium falciparum 63.539% Plasmodium chabaudi 63.372% Plasmodium reichenowi 62.933%
Lowest percentage of polyA-carrying transcripts (first 5)	Pythium vexans 0.025% Saprolegnia diclina vs 20 0.038% Leishmania major 0.048% Phytophthora sojae 0.058% Salpingoeca rosetta 0.060%
Median and average percentage of polyA-carrying transcripts	2.0% and 7.6% respectively
The longest polyA tracks (first 10)	132 nt - CDO62875 [*Plasmodium reichenowi*] 131 nt - CDO63348 [*Plasmodium reichenowi*] 111 nt - ETW31025 [*Plasmodium falciparum fch 4*] 109 nt - ETW57402 [*Plasmodium falciparum palo alto uganda*] 107 nt - ETW41820 [*Plasmodium falciparum nf135 5 c10*] 107 nt - ETW15539 [*Plasmodium falciparum vietnam oak knoll fvo*] 97 nt - CDO66404 [*Plasmodium reichenowi*] 95 nt - EUT78604 [*Plasmodium falciparum santa lucia*] 89 nt - ETW44841 [*Plasmodium falciparum nf135 5 c10*] 88 nt - ETW48723 [*Plasmodium falciparum malips096 e11*]

148

149

150

151

152

153   Fig.1. Distribution of polyA lengths vs AT-ratio of analyzed genomes. Data for lengths of polyA were

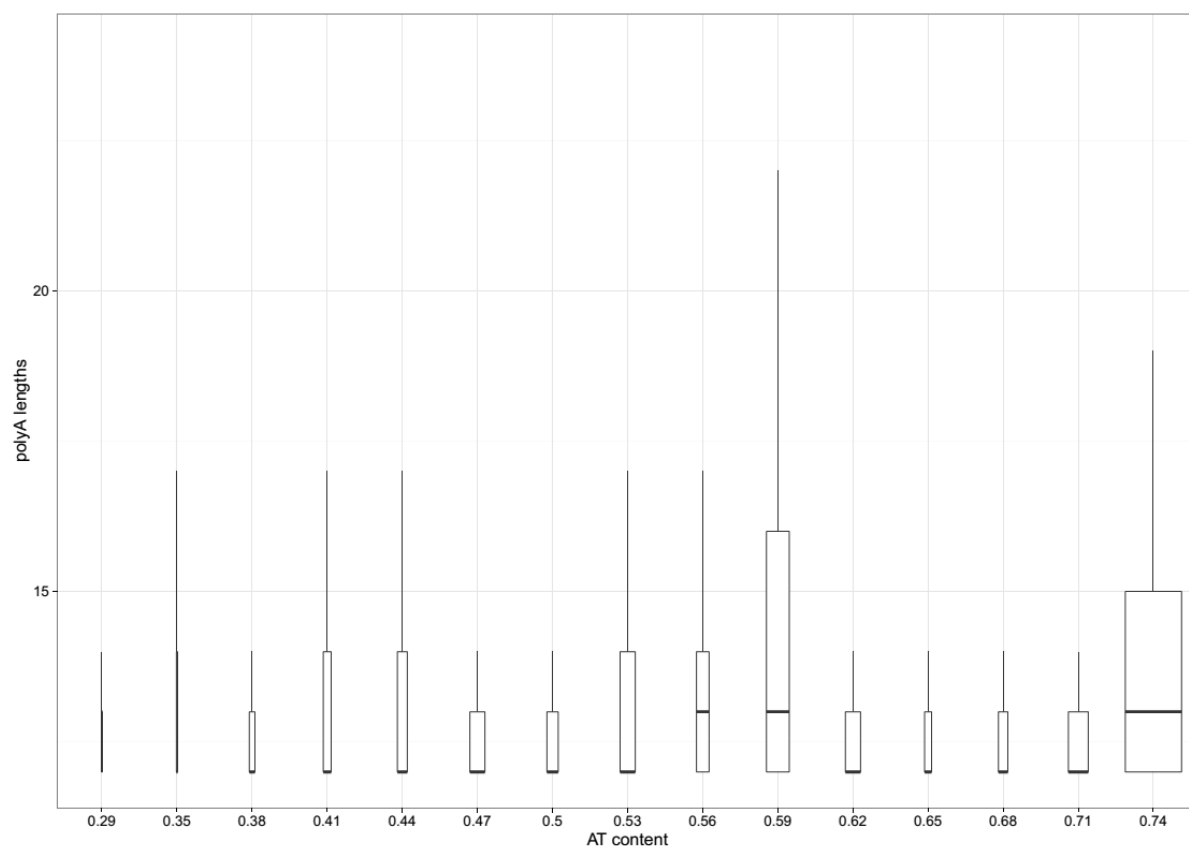
154   divided into 16 bins distributed evenly across spectrum of AT-richness of analyzed genomes. Width of a

155   box is proportional to the number of observances in a particular bin. Lines denote 1.5*IQR (interquantile

156   range). Outliers were removed for clarity. Data start at length of 12, as this was the length of the minimal

157   pattern used.

158

159

References

161   Arthur L., Pavlovic-Djuranovic S., Smith-Koutmou K., Green R., Szczesny P., Djuranovic S. 2015.

162   Translational control by lysine-encoding A-rich sequences. *Science advances* 1.

163   Brandman O., Stewart-Ornstein J., Wong D., Larson A., Williams CC., Li G-W., Zhou S., King D., Shen

164   PS., Weibezahn J., Dunn JG., Rouskin S., Inada T., Frost A., Weissman JS. 2012. A ribosome-bound

165   quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell*

166   151:1042–1054.

167   Chen J., Petrov A., Johansson M., Tsai A., O'Leary SE., Puglisi J. 2014. Dynamic pathways of -1

168   translational frameshifting. *Nature* 512:328–332.

169   Cherry JM., Adler C., Ball C., Chervitz SA., Dwight SS., Hester ET., Jia Y., Juvik G., Roe T., Schroeder

170   M., Others. 1998. SGD: Saccharomyces genome database. *Nucleic acids research* 26:73–79.

171   Cruz-Vera LR., Magos-Castro MA., Zamora-Romo E., Guarneros G. 2004. Ribosome stalling and

172   peptidyl-tRNA drop-off during translational delay at AGA codons. *Nucleic acids research* 32:4462–

173   4468.

174   Dimitrova LN., Kuroha K., Tatematsu T., Inada T. 2009. Nascent peptide-dependent translation arrest

175   leads to Not4p-mediated protein degradation by the proteasome. *The Journal of biological chemistry*

176   284:10343–10352.

177   Doma MK., Parker R. 2006. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation

178   elongation. *Nature* 440:561–564.

179   Flicek P., Amode MR., Barrell D., Beal K., Billis K., Brent S., Carvalho-Silva D., Clapham P., Coates G.,

180   Fitzgerald S., Gil L., Girón CG., Gordon L., Hourlier T., Hunt S., Johnson N., Juettemann T., Kähäri

181   AK., Keenan S., Kulesha E., Martin FJ., Maurel T., McLaren WM., Murphy DN., Nag R., Overduin

182   B., Pignatelli M., Pritchard B., Pritchard E., Riat HS., Ruffier M., Sheppard D., Taylor K., Thormann

183   A., Trevanion SJ., Vullo A., Wilder SP., Wilson M., Zadissa A., Aken BL., Birney E., Cunningham

184   F., Harrow J., Herrero J., Hubbard TJP., Kinsella R., Muffato M., Parker A., Spudich G., Yates A.,

185   Zerbino DR., Searle SMJ. 2014. Ensembl 2014. *Nucleic acids research* 42:D749–55.

186   Koutmou KS., Schuller AP., Brunelle JL., Radhakrishnan A., Djuranovic S., Green R. 2015. Ribosomes

187   slide on lysine-encoding homopolymeric A stretches. *eLife* 4.

188   Kuroha K., Akamatsu M., Dimitrova L., Ito T., Kato Y., Shirahige K., Inada T. 2010. Receptor for

189   activated C kinase 1 stimulates nascent polypeptide-dependent translation arrest. *EMBO reports*

190   11:956–961.

191   Letzring DP., Dean KM., Grayhack EJ. 2010. Control of translation efficiency in yeast by codon-

192   anticodon interactions. *RNA* 16:2516–2528.

193   Pruitt KD., Brown GR., Hiatt SM., Thibaud-Nissen F., Astashyn A., Ermolaeva O., Farrell CM., Hart J.,

194   Landrum MJ., McGarvey KM., Murphy MR., O'Leary NA., Pujar S., Rajput B., Rangwala SH.,

195   Riddick LD., Shkeda A., Sun H., Tamez P., Tully RE., Wallin C., Webb D., Weber J., Wu W.,

196   DiCuccio M., Kitts P., Maglott DR., Murphy TD., Ostell JM. 2014. RefSeq: an update on

197   mammalian reference sequences. *Nucleic acids research* 42:D756–763.

198   Shoemaker CJ., Green R. 2012. Translation drives mRNA quality control. *Nature structural & molecular*

199   *biology* 19:594–601.

200   Yan S., Wen J-D., Bustamante C., Tinoco I Jr. 2015. Ribosome excursions during mRNA translocation

201   mediate broad branching of frameshift pathways. *Cell* 160:870–881.

202

203