

Product Grammars for Alignment and Folding

Supplemental Material

Christian Höner zu Siederdissen, Ivo L. Hofacker, Peter F. Stadler

PROOF OF ASSOCIATIVITY FOR 2-GNF

Since normal forms contain only a few different types of rules, it suffices to check associativity for every combination of types of rules. A grammar in Greibach 2-Normal Form (2-GNF) [1], [2] is of the form $A \rightarrow aBC \mid bD \mid c$ with a, b, c terminal and A, B, C, D non-terminal symbols. For the Greibach 2-NF the two products that need to be checked take the form:

$$\begin{aligned} & (\{A \rightarrow aBC|bD|c\} \odot \{A \rightarrow aBC|bD|c\}) \odot \\ & \quad \{A \rightarrow aBC|bD|c\} \\ = & \\ & \{A \rightarrow aBC|bD|c\} \odot \\ & (\{A \rightarrow aBC|bD|c\} \odot \{A \rightarrow aBC|bD|c\}) \end{aligned}$$

Computationally, we check that both grammar products yield the same set of 57 productions:

$$\begin{array}{ll} \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ a \end{matrix}\right) \left(\begin{matrix} B \\ B \end{matrix}\right) \left(\begin{matrix} C \\ C \end{matrix}\right) & \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} b \\ b \end{matrix}\right) \left(\begin{matrix} D \\ D \end{matrix}\right) \left(\begin{matrix} \epsilon \\ \epsilon \end{matrix}\right) \\ \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ b \end{matrix}\right) \left(\begin{matrix} B \\ D \end{matrix}\right) \left(\begin{matrix} C \\ \epsilon \end{matrix}\right) & \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} b \\ a \end{matrix}\right) \left(\begin{matrix} D \\ B \end{matrix}\right) \left(\begin{matrix} \epsilon \\ \epsilon \end{matrix}\right) \\ \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ b \end{matrix}\right) \left(\begin{matrix} B \\ \epsilon \\ D \end{matrix}\right) \left(\begin{matrix} C \\ \epsilon \end{matrix}\right) & \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} b \\ b \end{matrix}\right) \left(\begin{matrix} D \\ B \end{matrix}\right) \left(\begin{matrix} \epsilon \\ C \end{matrix}\right) \\ \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ c \\ \epsilon \end{matrix}\right) \left(\begin{matrix} B \\ C \\ \epsilon \end{matrix}\right) & \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} b \\ b \\ \epsilon \end{matrix}\right) \left(\begin{matrix} D \\ C \\ \epsilon \end{matrix}\right) \\ \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ b \\ a \end{matrix}\right) \left(\begin{matrix} B \\ D \\ B \end{matrix}\right) \left(\begin{matrix} C \\ \epsilon \\ C \end{matrix}\right) & \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} b \\ b \\ a \end{matrix}\right) \left(\begin{matrix} D \\ D \\ B \end{matrix}\right) \left(\begin{matrix} \epsilon \\ \epsilon \\ C \end{matrix}\right) \\ \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ b \\ b \end{matrix}\right) \left(\begin{matrix} B \\ D \\ D \end{matrix}\right) \left(\begin{matrix} C \\ \epsilon \\ \epsilon \end{matrix}\right) & \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} b \\ b \\ b \end{matrix}\right) \left(\begin{matrix} D \\ D \\ D \end{matrix}\right) \left(\begin{matrix} \epsilon \\ \epsilon \\ \epsilon \end{matrix}\right) \\ \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ b \\ c \\ \epsilon \end{matrix}\right) \left(\begin{matrix} B \\ D \\ D \\ C \end{matrix}\right) \left(\begin{matrix} C \\ \epsilon \\ \epsilon \\ \epsilon \end{matrix}\right) & \end{array}$$

$$\begin{array}{ll} \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ c \\ c \end{matrix}\right) \left(\begin{matrix} B \\ \epsilon \\ \epsilon \end{matrix}\right) \left(\begin{matrix} C \\ \epsilon \\ \epsilon \end{matrix}\right) & \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} c \\ a \\ a \end{matrix}\right) \left(\begin{matrix} B \\ B \\ B \end{matrix}\right) \left(\begin{matrix} C \\ C \\ C \end{matrix}\right) \\ \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} b \\ a \\ a \end{matrix}\right) \left(\begin{matrix} D \\ B \\ B \end{matrix}\right) \left(\begin{matrix} \epsilon \\ C \\ C \end{matrix}\right) & \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} c \\ b \\ b \end{matrix}\right) \left(\begin{matrix} B \\ D \\ D \end{matrix}\right) \left(\begin{matrix} \epsilon \\ C \\ C \end{matrix}\right) \\ \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ b \\ b \\ b \end{matrix}\right) \left(\begin{matrix} B \\ B \\ D \\ D \end{matrix}\right) \left(\begin{matrix} C \\ C \\ \epsilon \\ \epsilon \end{matrix}\right) & \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} c \\ b \\ b \\ b \end{matrix}\right) \left(\begin{matrix} B \\ D \\ D \\ D \end{matrix}\right) \left(\begin{matrix} C \\ C \\ C \\ \epsilon \end{matrix}\right) \\ \left(\begin{matrix} A \\ A \end{matrix}\right) \rightarrow \left(\begin{matrix} a \\ b \\ b \\ c \\ \epsilon \end{matrix}\right) \left(\begin{matrix} B \\ B \\ D \\ D \\ C \end{matrix}\right) \left(\begin{matrix} C \\ C \\ \epsilon \\ \epsilon \\ \epsilon \end{matrix}\right) & \end{array}$$

DNA-PROTEIN ALIGNMENT

The Complete Semi-Global Alignment Grammar

The full semi-global grammar contains a large number of symbols and rules. There are 5 non-terminal symbols $(\frac{F_0}{P}), (\frac{F_1}{P}), (\frac{F_2}{P})$ denoting three different reading frames for the DNA sequence F , as well as $, (\frac{L}{P}), (\frac{R}{P})$ for the left and right unaligned DNA sequence part. All non-terminals are of dimension two.

We use three terminal symbols. Terminal symbols act on one dimension only, but are combined to read from two tapes simultaneously. The c terminal reads a single DNA nucleotide while a reads a single amino acid. The $\$$ terminal acts as the sentinel parsing the empty substring. We *bind* terminal symbols rather late in the grammar construction process, as this allows us to switch between underlying representation types. This includes the sentinel character which therefore acts as a normal terminal symbol.

These non-terminal and terminal symbols are combined by 34 production rules. 3×9 of those are used to calculate the frameshift-dependent alignment. Two rules deal with the left unaligned DNA sequence and five rules with the right unaligned part.

The start symbol is (^R_P) .

$$\begin{array}{ll}
 (^F_0_P) \rightarrow (^F_0_P)(\varepsilon_a) & (^F_1_P) \rightarrow (^F_1_P)(c)(c)(c)(c) \\
 (^F_0_P) \rightarrow (^L_P) & (^F_2_P) \rightarrow (^F_2_P)(\varepsilon_a) \\
 (^F_0_P) \rightarrow (\$) & (^F_2_P) \rightarrow (^L_P) \\
 (^F_0_P) \rightarrow (^F_1_P)(c)(c) & (^F_2_P) \rightarrow (\$) \\
 (^F_0_P) \rightarrow (^F_1_P)(c)(\varepsilon) & (^F_2_P) \rightarrow (^F_0_P)(c)(c)(c) \\
 (^F_0_P) \rightarrow (^F_2_P)(c) & (^F_2_P) \rightarrow (^F_0_P)(c)(c)(c) \\
 (^F_0_P) \rightarrow (^F_2_P)(c)(\varepsilon) & (^F_2_P) \rightarrow (^F_1_P)(c)(a) \\
 (^F_0_P) \rightarrow (^F_0_P)(c)(c)(c) & (^F_2_P) \rightarrow (^F_1_P)(c)(\varepsilon) \\
 (^F_0_P) \rightarrow (^F_0_P)(c)(\varepsilon)(c) & (^F_2_P) \rightarrow (^F_2_P)(c)(c)(c)(c) \\
 (^F_1_P) \rightarrow (^F_1_P)(\varepsilon_a) & (^F_2_P) \rightarrow (^F_2_P)(c)(c)(c)(c) \\
 (^F_1_P) \rightarrow (^L_P) & (^L_P) \rightarrow (^L_P)(c) \\
 (^F_1_P) \rightarrow (\$) & (^L_P) \rightarrow (\$) \\
 (^F_1_P) \rightarrow (^F_2_P)(c)(c) & (^R_P) \rightarrow (^R_P)(c)(\varepsilon) \\
 (^F_1_P) \rightarrow (^F_2_P)(c)(\varepsilon) & (^R_P) \rightarrow (^F_0_P) \\
 (^F_1_P) \rightarrow (^F_0_P)(c) & (^R_P) \rightarrow (^F_1_P) \\
 (^F_1_P) \rightarrow (^F_1_P)(c)(c)(c) & (^R_P) \rightarrow (^F_2_P) \\
 (^F_1_P) \rightarrow (^F_1_P)(c)(c)(\varepsilon) & (^R_P) \rightarrow (\$)
 \end{array}$$

Alignments of *R. americana* proteins to *P. polycephalum* mitogenomic DNA

As a pilot study for the frameshift DNA-Protein alignment, we aligned the mitochondrial genome of *P. polycephalum* (62,862 nt [3]) against either the mitochondrial protein sequences, as determined from transcriptome sequencing [4], or against the 67 mitochondrial protein coding sequences from *Reclinomonas americana*. The former tests in how far the algorithm is able to predict RNA editing sites, while the comparison to *Reclinomonas* tests our ability to annotate the *Physarum* genome using remote homologs. The results of the *Reclinomonas* comparison are summarized in Table 1 and Fig. 3. In addition we present detailed results and alignments for the nad5 gene as an example below.

In REDBASE [4] the nad5 alignment is annotated at genomic position 17259–19152 with 69 C insertions, 7 U, and 3 A or G insertions¹.

Our algorithm produces a very high-scoring hit with an average score (using BLOSUM 50) of 4.03 per amino acid position at genomic position 17258–19149 with 76 one-nucleotide insertions. Hence, we are able to recover the alignment of the *Physarum* amino acid sequence to its mitogenome. Table 1 gives an overview of the results for all protein coding genes in the *Physarum* mitogenome.

1. <http://bioserv.mps.ohio-state.edu/redbase/> (nad5)

TABLE 1
Comparison of annotated *Physarum polycephalum* mitogenes from REDBASE with location predicted by aligning *Reclinomonas americana* genes to the *Physarum polycephalum* genome.

name	REDBASE	predicted	comment
nad5	17259-19152	17248-19152	
nadG	19300-20316	—	no homolog
rpS2	20278-21633	—	
rpS12	21746-22241	21735-22109	
rpS7	22246-23009	—	
rpL2	23009-23777	—	
rpS19	23774-24139	—	
php15	24416-25471	—	no homolog
cox1	27534-25816	27474-25950	
nad7	27670-27535	27674	
cox2	29666-28983	29698-28949	
php22	29776-30652	—	no homolog
nad2	30699-32105	30706-32133	
rpS16	34704-34988	—	no homolog
rpL19	34988-35540	—	
atp8	35567-35788	—	
nad4L	35788-36062	—	
atp6	36067-36774	36059-36771	
nad4	38315-36933	38344-36928	
nad3	38808-38435	38802-38450	
rpL14	38985-39338	38981-39336	
php23	39338-39822	—	no homolog
rpS14	39823-40087	39798-40084	
rpS8	40088-40509	—	
rpL6	40506-40972	—	
rpS13	40978-41517	—	
nad9	41520-41994	—	
rpS11	41997-42717	—	
php24	42721-43372	—	no homolog
rpS4	44229-44440	—	
php25	53565-53858	—	no homolog
atpA	53845-55380	53854-55319	
cox3	56278-55517	56287-55526	
nad6	56758-56280	56780-56187	
rpL16	57434-56910	57338-56948	
rpS3	58800-57431	—	
nad1	58893-59821	58889-59834	
cytb	61038-59903	61043-59921	
atp9	61225-61467	61250-61464	

Fig. 1 gives the alignment of the *Physarum* protein against the full *Physarum* mitogenome. The proposed alignment is quite close to the reference provided by the REDBASE ([4], <http://bioserv.mps.ohio-state.edu/redbase/>). In particular, the genomic start and end positions are within 1 resp. 3 nt of the proposed positions. The number of frameshift modifications is also very close (76 vs. 79 in the reference). This result shows that we can successfully align protein sequences to their genomes under the assumption that frameshift modifications are possible. In this example we do not, a priori, assume that C insertions are to be scored better than other insertions.

Note that in the alignment representation, we currently denote all 1nt frameshift alignments by $(c_1)(c_2)(-)$ irrespective of where the actual insertion happens to form the full codon. If required, all co-optimal solutions can be extracted.

For a more challenging task, we extracted the nad5 protein sequence from *Reclinomonas americana* and

aligned the sequence against the *Physarum polycephalum* mitogenome. Due to their larger evolutionary distance, finding the correct alignment is not trivial, considering that frameshift modifications have to be observed as well. As Fig. 2 shows, we still recover the alignment with 100% overlap (with respect to the smaller sequence) between the two alignments of Fig. 1 and Fig. 2. The start and end positions differ by 12 nt and the number of proposed frameshifts drops to 35, however.

It is worth noting that both alignments have much better scores than the respective next-best candidates. For the self-alignment, the next-best solution has a score of 449 compared to 2642 (or a score of 0.68 vs 4.03 per amino acid), while for *reclinomonas* it is -150 to 582 (score of -0.22 vs 0.87 per amino acid).

REFERENCES

- [1] S. A. Greibach, "A new normal-form theorem for context-free phrase structure grammars," *J. ACM*, vol. 12, pp. 42–52, 1965.
- [2] N. Blum and R. Koch, "Greibach normal form transformation revisited," *Inform. Comput.*, vol. 150, pp. 112–118, 1999.
- [3] H. Takano, T. Abe, R. Sakurai, Y. Moriyama, Y. Miyazawa, H. Nozaki, S. Kawano, N. Sasaki, and T. Kuroiwa, "The complete DNA sequence of the mitochondrial genome of *Physarum polycephalum*," *Mol. Gen. Genet.*, vol. 264, pp. 539–545, 2001.
- [4] R. Bundschuh, J. Altmüller, C. Becker, P. Nürnberg, and J. M. Gott, "Complete characterization of the edited transcriptome of the mitochondrion of *Physarum polycephalum* using deep sequencing of RNA," *Nucleic Acids Res.*, vol. 39, pp. 6044–6055, 2011.

DNA: gi 11466223 ref NC_002508.1 Physarum polycephalum mitochondrion, complete genome @ Forward 17258		Protein: tr F2Y9T4 F2Y9T4_PHYPO NADH-ubiquinone oxidoreductase chain 5 OS=Physarum polycephalum GN=nad5 PE=2 SV=1 @ 0	
DNA length: 3936 Protein length: 656		1 Nt shifts: 76 2 Nt shifts: 0	
Score: 2642 Length-adjusted: 4.03			
17258 AATGTTTCATGTTCTTA-ATAGCATTATAATTCTTTATGTTAGGTAGACATCTGGGA <u>AG</u> -CAAATT <u>GT</u> -CTCGGC <u>CTT</u> GCTATT	17344		
1 M S F M F P L I A F I I L F M L G R H L G K Q I A L G F A I	30		
17345 ACAATGTC <u>ATT</u> <u>TA</u> -TC <u>ACT</u> TATAATT <u>GT</u> -TTATATT <u>TTT</u> ATT <u>CAT</u> GTT <u>TTT</u> <u>TTA</u> -GGTCAA <u>AT</u> <u>TA</u> -AG <u>CTT</u> A <u>TT</u> TTAGGTTCT	17430		
31 T M S F L S L I I C L Y Y F I H V F F Y G Q I Y S F N L G S	60		
17431 TGGGTTCTGTAGGT <u>ACT</u> TTAG <u>AT</u> ACT <u>AC</u> -AA <u>ATT</u> T <u>TA</u> <u>AT</u> -GAT <u>CTT</u> TT <u>CC</u> <u>AT</u> ACT <u>TT</u> -GGTAC <u>AT</u> TA <u>AT</u> <u>CC</u> -TTT <u>AT</u> ACT	17516		
61 W V S V G T L D I T Y K F I I D P L S I T F G T L I S F I T	90		
17517 TT <u>AT</u> TA <u>AT</u> <u>CT</u> <u>TA</u> -AT <u>TT</u> AT <u>TT</u> CT <u>AT</u> G <u>AT</u> <u>AT</u> -TT <u>AC</u> T <u>GA</u> AG <u>AT</u> C <u>CT</u> <u>TA</u> <u>AT</u> <u>TA</u> -G <u>TT</u> AA <u>TT</u> TT <u>TT</u> G <u>CT</u> <u>T</u> TT <u>AT</u> <u>TA</u> <u>GT</u> -TTT <u>TC</u> TT <u>TT</u> CT	17602		
91 L L I L I Y S Y D Y L H E D P N L V K F F A Y L S F F S F S	120		
17603 AT <u>GT</u> <u>TT</u> -TG <u>T</u> TT <u>GT</u> TT <u>TT</u> G <u>CT</u> GG <u>TA</u> TT <u>AC</u> TT <u>CT</u> <u>TT</u> -AT <u>AT</u> TT <u>GT</u> TT <u>TT</u> AG <u>AT</u> GG <u>GG</u> AG <u>G</u> C <u>GT</u> GG <u>GA</u> -TT <u>AG</u> CT <u>CT</u> TT <u>AT</u> <u>TT</u> -CTT <u>AT</u> TT <u>AT</u> TT <u>TC</u>	17688		
121 M S C L V F A G N Y F I M F L G W E A V G G L A S Y L L I N F	150		
17689 TGG <u>AT</u> -ACA <u>AG</u> AA <u>AT</u> CA <u>AG</u> CA <u>AT</u> -C <u>AG</u> T <u>CT</u> G <u>CT</u> TTAA <u>AG</u> CA <u>TT</u> TT <u>TA</u> AT <u>CG</u> <u>TA</u> -GG <u>GT</u> G <u>AT</u> G <u>CA</u> G <u>CT</u> TT <u>CT</u> <u>TA</u> AG <u>GT</u> G <u>CT</u> AT <u>GG</u> GT	17775		
151 W T T R N Q A N Q S A I K A I I F N R V G D A A F I S A M G	180		
17776 CT <u>TT</u> TT <u>TT</u> TT <u>TT</u> <u>TT</u> -TT <u>AA</u> TT <u>CT</u> <u>TT</u> -G <u>AT</u> TT <u>AG</u> AG <u>AT</u> TT <u>AG</u> AA <u>TT</u> ACT <u>CT</u> G <u>TT</u> <u>CA</u> -CA <u>AT</u> TA <u>AT</u> GA <u>CA</u> ACT <u>AC</u> TA <u>CT</u> TT <u>TT</u> AG <u>CT</u> <u>TT</u> -TTT <u>TC</u> A	17861		
181 L I Y Y L F N S F D L E D L E L L V P Q Y E H T T F S L F S	210		
17862 T <u>AT</u> T <u>CT</u> TT <u>TT</u> <u>CA</u> -ACA <u>AT</u> TT <u>GA</u> <u>AT</u> <u>GT</u> -AT <u>AG</u> C <u>T</u> TT <u>CT</u> TT <u>GT</u> <u>CT</u> -G <u>CT</u> <u>GT</u> -G <u>CT</u> AA <u>AT</u> C <u>AG</u> C <u>AA</u> <u>AT</u> -TTT <u>TT</u> AC <u>AT</u> CC <u>TT</u> GG <u>TT</u> A	17946		
211 Y S F H T I E L I A L F L F L A A A A K S A Q L F L H P W L	240		
17947 CCT <u>GAT</u> G <u>CT</u> AT <u>GAA</u> -GG <u>AC</u> CT <u>AC</u> CC <u>AG</u> TT <u>TC</u> AG <u>CA</u> TT <u>TA</u> CA <u>TT</u> CT <u>G</u> C <u>TA</u> <u>AT</u> <u>GT</u> -G <u>TA</u> AC <u>AG</u> C <u>AG</u> G <u>GT</u> <u>TT</u> <u>TC</u> -TT <u>AT</u> TT <u>AT</u> TA <u>AG</u> AT <u>CT</u>	18033		
241 P D A M E G P T P V S A L L H S A T M V T A G V F L I L R S	270		
18034 CT <u>GT</u> -G <u>TT</u> TT <u>TT</u> CT <u>CA</u> AT <u>G</u> C <u>CT</u> TT <u>AT</u> TT <u>CA</u> TT <u>AT</u> <u>GT</u> <u>TA</u> -G <u>CT</u> <u>GT</u> -G <u>TT</u> GG <u>CT</u> AA <u>AT</u> <u>CA</u> <u>TA</u> -G <u>CT</u> AA <u>AT</u> TT <u>CT</u> TT <u>CA</u> AC <u>AG</u> GT <u>TA</u> <u>AT</u>	18119		
271 S V I F S H A P Y I S L L V A C I G L I T A N I S S L T G L	300		
18120 TT <u>AC</u> AA <u>AT</u> AT <u>GA</u> CA <u>AA</u> AC <u>GT</u> <u>AT</u> <u>TT</u> -G <u>C</u> AT <u>TT</u> CA <u>AC</u> CT <u>GT</u> AG <u>CC</u> AA <u>CT</u> TT <u>GG</u> <u>TT</u> -AT <u>GT</u> TT <u>GT</u> <u>CT</u> <u>AT</u> -GG <u>T</u> TT <u>GG</u> TA <u>AT</u> TA <u>CT</u>	18206		
301 L Q Y D I K R I I A F S T C S Q L G F M M F A T G I G N Y T	330		
18207 TT <u>GT</u> -G <u>TT</u> TT <u>TT</u> CT <u>AT</u> TT <u>GA</u> AA <u>AT</u> <u>AT</u> -G <u>CT</u> TT <u>CT</u> TT <u>AA</u> AG <u>CA</u> CT <u>CT</u> TT <u>AT</u> <u>TA</u> <u>TA</u> -T <u>GT</u> GC <u>GG</u> G <u>AT</u> CC <u>GT</u> TT <u>AC</u> GG <u>CC</u> AT <u>CA</u> G	18293		
331 F A L F H L V N H A F F K A L L F L C A G S V I H A T G H Q	360		
18294 GAT <u>AT</u> TC <u>GG</u> GT <u>AT</u> GG <u>GA</u> -TT <u>AT</u> TT <u>AA</u> GA <u>AT</u> -TT <u>AC</u> CC <u>CA</u> AA <u>CT</u> TT <u>AT</u> GT <u>TC</u> GA <u>AT</u> GT <u>CT</u> TT <u>GT</u> <u>CT</u> -T <u>CT</u> TT <u>AT</u> CT <u>TT</u> AA <u>TT</u> GG <u>TT</u> <u>CT</u> -	18379		
361 D I R R M G A L F K A L P I T Y V A M L L A S L S L I G F P	390		
18380 TT <u>CT</u> TT <u>AA</u> GT <u>GG</u> TT <u>TT</u> AT <u>AG</u> -A <u>GG</u> AT <u>TT</u> TT <u>CT</u> CT <u>AG</u> AA <u>AG</u> CT <u>AC</u> TT <u>AC</u> AA <u>AT</u> AT <u>TT</u> GG <u>AT</u> GT <u>TC</u> -T <u>CT</u> TT <u>AT</u> GT <u>TT</u> TT <u>AT</u> TT <u>AT</u> AC <u>CC</u>	18467		
391 F L S G F Y S K D F L L E A T Y N I F G M F S Y V I Y F I S	420		
18468 AT <u>AT</u> TT <u>TC</u> TA <u>CT</u> GT <u>TT</u> AG <u>TA</u> -TT <u>TT</u> AC <u>TC</u> TT <u>CC</u> G <u>AT</u> TT <u>TC</u> TT <u>GT</u> <u>CT</u> -G <u>AT</u> CA <u>AG</u> C <u>AT</u> <u>TA</u> -T <u>CA</u> AA <u>AA</u> AC <u>AT</u> <u>TT</u>	18553		
421 T I S T A V S S F Y S F R L I F F V F L G D Q A S S I K T L	450		
18554 AAA <u>AC</u> AT <u>AT</u> <u>CT</u> -GAA <u>AG</u> CT <u>CC</u> TT <u>AT</u> <u>TT</u> CT <u>AT</u> <u>TA</u> <u>CT</u> <u>CA</u> <u>TA</u> -T <u>TA</u> AT <u>TT</u> TT <u>TA</u> ACT <u>AT</u> AC <u>TC</u> TT <u>CT</u> TT <u>AT</u> <u>TT</u> <u>CT</u> -GG <u>AT</u> TT <u>TT</u> AA <u>AG</u> AT	18640		
451 K T I S E S S S Y F L Y L P L I I L T I L S I F S G F Y L K D	480		
18641 TG <u>AT</u> GA <u>CT</u> TT <u>AC</u> AC <u>AT</u> <u>CT</u> <u>TA</u> -T <u>AT</u> AA <u>TT</u> TT <u>AG</u> TC <u>GT</u> CT <u>CT</u> <u>CT</u> <u>AT</u> -T <u>TC</u> AG <u>T</u> G <u>AT</u> AC <u>G</u> C <u>TA</u> <u>CT</u> <u>GT</u> <u>CT</u> -TT <u>TA</u> AT <u>GA</u> TT <u>TT</u> TT	18726		
481 L M T I Q T S L Y N F S A S P F S D T A T D Q D F F N D L F	510		
18727 AA <u>AG</u> TT <u>TA</u> -CCC <u>AC</u> AT <u>CT</u> TT <u>CT</u> TT <u>AT</u> <u>CG</u> <u>GT</u> -T <u>TA</u> CT <u>TT</u> AG <u>TA</u> AT <u>TA</u> AT <u>TA</u> TT <u>AT</u> <u>TT</u> <u>GT</u> -AA <u>TT</u> AA <u>AC</u> TT <u>AA</u> AG <u>CT</u> AA <u>CT</u> TT <u>AT</u> <u>AT</u>	18813		
511 K V L P T I F S L S G L L L V Y I I Y L K L N L K S Q L L Y	540		
18814 AG <u>AC</u> AA <u>CA</u> TA <u>CT</u> -TT <u>AT</u> AC <u>CT</u> TT <u>TT</u> <u>TT</u> <u>GA</u> -G <u>AT</u> GC <u>TT</u> TT <u>TA</u> AG <u>GT</u> TT <u>TT</u> AT <u>AT</u> <u>TT</u> <u>TC</u> -TT <u>AC</u> CA <u>CT</u> GT <u>CA</u> CT	18899		
541 R Q Y L L L P Y L S C K K F F A D A F N S F Y I F L P S A T	570		
18900 T <u>TT</u> <u>CT</u> -TT <u>AA</u> AT <u>AT</u> <u>CT</u> -T <u>AT</u> AA <u>TT</u> AT <u>AG</u> <u>AT</u> <u>AA</u> -G <u>GT</u> TT <u>TT</u> AG <u>AC</u> AT <u>TT</u> AG <u>GT</u> TC <u>AC</u> GG <u>GT</u> <u>AT</u> <u>AT</u> -G <u>CT</u> TT <u>CT</u> GA <u>AT</u> <u>TT</u> -AT	18984		
571 F S L N I T Y K I I D Q G V L E H L G S T G I Y A F L E S I	600		
18985 T <u>TT</u> GA <u>AA</u> GT <u>TT</u> TT <u>AG</u> TA <u>AT</u> <u>GT</u> <u>GA</u> -AC <u>AA</u> CC <u>CT</u> TA <u>AT</u> <u>AA</u> TT <u>AT</u> <u>AT</u> <u>CT</u> -C <u>G</u> TT <u>CT</u> TT <u>CT</u> TA <u>AT</u> <u>AT</u> <u>TT</u> GT <u>CC</u> TC <u>GG</u> AG <u>GT</u> CA <u>CT</u> TT <u>AT</u> <u>TA</u> <u>AT</u> -AT	19071		
601 F E S I V N V E T T L I I Y R S F L F I I F V L G A T L S I	630		
19072 T <u>TT</u> TT <u>AG</u> TT <u>TT</u> AC <u>G</u> CT <u>TT</u> TT <u>CT</u> TA <u>AG</u> CA <u>TT</u> <u>TT</u> TT <u>AG</u> TA <u>CT</u> CT <u>TA</u> AT <u>TT</u> TT <u>TA</u> AT <u>AC</u> GG <u>AA</u> AT <u>CA</u> CT <u>TC</u> TA	19149		
631 F Y G L Y A F L I A I F F I S T L N I F N T E I T S	656		

Fig. 1. nad5 protein sequence of Physarum aligned to Physarum mitogenome: Output of the frameshift-aware DNA-Protein alignment tool. The protein sequence is aligned locally to the DNA sequence, creating a semi-global alignment. Individual codon / amino acid combinations are colored according to their similarity. Scores range from very similar (> 5, cyan), similar (> 0, blue), neutral (0, white), dissimilar (< 0, yellow), to very dissimilar (< 5, red). Full in/del's are not colored. Combinations of frameshifts and alignments are colored, bold, and underlined (only the DNA sequence). The color is only determined by the similarity scores, not the additional in/del malus. For this example, a BLOSUM 50 matrix was used. The proposed genomic position start and end of the alignment are within 1nt and 2nt of the REDBASE [4] alignment.

DNA: gi 11466223 ref NC_002508.1 Physarum polycephalum mitochondrion, complete genome @ Forward 17247
Protein: lcl KC353356.1_cdsid_AGH24310.1 [gene=nad5] [protein=NADH dehydrogenase subunit 5] [protein_id=AGH24310.1] [location=54506..56518]
DNA length: 4020 Protein length: 670
1 Nt shifts: 35 2 Nt shifts: 0
Score: 582 Length-adjusted: 0.87
17248 AATAAA <u>T</u> AAATGTTT <u>C</u> AT <u>G</u> TT <u>C</u> CTTA <u>A</u> T <u>G</u> C <u>A</u> TT <u>T</u> AT <u>G</u> TT <u>AG</u> GT <u>A</u> CA <u>T</u> CTGG <u>G</u> AG <u>C</u> AA <u>AT</u> GT <u>T</u> CT <u>CG</u> CT <u>T</u> 17337 1 M Y L L I V F L P L L G S I T A G F F G R S L G K Q G A A I 30
17338 TGCTATTAC <u>A</u> AT---GTC <u>A</u> TT <u>T</u> AT <u>C</u> ACT <u>T</u> ATA <u>A</u> TT <u>G</u> TT <u>T</u> AT <u>A</u> TT <u>T</u> TT <u>A</u> TT <u>C</u> AT <u>G</u> TT <u>T</u> TT <u>T</u> AT <u>G</u> GT <u>C</u> AA <u>A</u> AT <u>A</u> AG <u>C</u> TT---AAT 17421 31 I T T S C V A L S S L F S M V A F Y E V G L C G S P C Y I R 60
17422 TT <u>AG</u> GT <u>T</u> CT <u>GG</u> TT <u>T</u> CT <u>G</u> TAG <u>G</u> ACT <u>T</u> TA <u>G</u> AT <u>A</u> TT <u>A</u> C <u>-</u> T <u>A</u> CA <u>A</u> TT <u>T</u> AT <u>A</u> AT <u>G</u> ---AT <u>C</u> TT <u>T</u> AT <u>C</u> AT <u>A</u> TT <u>T</u> GG <u>T</u> AC <u>A</u> TT <u>A</u> AT <u>T</u> CC 17507 61 L F N W I D S E M L H A S W G F L F D S L T V V M L I V V T 90
17508 TTT <u>AT</u> ACT <u>T</u> TT <u>A</u> AT <u>T</u> TA <u>A</u> AT <u>T</u> GA <u>-</u> AT <u>T</u> TT <u>A</u> AT <u>G</u> GA <u>A</u> GT <u>C</u> CT <u>A</u> AT <u>T</u> AG <u>T</u> AA <u>AT</u> TT <u>TT</u> GT <u>-</u> CT <u>T</u> TT <u>A</u> AT <u>T</u> TT 17592 91 I V S S L V H L Y S V G Y M S H D P H L P R F M S Y L S L F 120
17593 CT <u>T</u> TT <u>T</u> CT <u>T</u> AT <u>G</u> TT <u>T</u> GT <u>T</u> CT <u>G</u> TT <u>T</u> GT <u>G</u> TA <u>A</u> TT <u>A</u> CT <u>T</u> ---AT <u>T</u> AT <u>G</u> TT <u>T</u> TT <u>AG</u> GT <u>GG</u> AA <u>G</u> C <u>-</u> T <u>G</u> T <u>G</u> AT <u>T</u> AG <u>C</u> T <u>T</u> TT <u>A</u> TT <u>T</u> CT 17679 121 T F F M L M L V T G D N F V Q M F L G W E G V G L C S Y L L 150
17680 AT <u>T</u> AA <u>TT</u> CT <u>GG</u> AT <u>-</u> AC <u>A</u> AG <u>A</u> AT <u>C</u> A <u>G</u> CA <u>A</u> U <u>-</u> CA <u>G</u> CT <u>G</u> T <u>A</u> TT <u>A</u> AA <u>GC</u> AA <u>T</u> TT <u>T</u> TT <u>A</u> AT <u>C</u> G <u>T</u> GA <u>-</u> GG <u>T</u> G <u>A</u> T <u>G</u> C <u>A</u> G <u>C</u> T <u>T</u> TC <u>-</u> ATA 17763 151 I N F W F T R L Q A N K S A I K A M I M N R I G D F G L S L 180
17764 AG <u>T</u> G <u>C</u> T <u>A</u> GG <u>G</u> T <u>T</u> TT <u>A</u> TT <u>T</u> TA <u>A</u> TT <u>T</u> CT <u>T</u> GT <u>A</u> TT <u>T</u> AG <u>A</u> AG <u>A</u> TT <u>T</u> AG <u>A</u> TT <u>T</u> GT <u>T</u> CA <u>A</u> AT <u>A</u> GA <u>A</u> CA <u>T</u> ACT <u>T</u> TT <u>A</u> GC 17853 181 G M M A I F F I F K S V D F I T V F A L S P Y M T D A T I V 210
17854 TTT <u>T</u> CA <u>T</u> AT <u>T</u> TT <u>T</u> CA <u>A</u> CA <u>A</u> AT <u>T</u> GA <u>-</u> --AT <u>G</u> A <u>T</u> AG <u>C</u> T <u>T</u> TT <u>T</u> CT <u>A</u> TT <u>T</u> GT <u>G</u> T <u>G</u> T <u>G</u> CT <u>A</u> TA <u>A</u> TC <u>A</u> CG <u>C</u> AA <u>A</u> CT <u>T</u> TT <u>-</u> TT <u>A</u> CA <u>T</u> 17937 211 F L N Y E V H A L T L I C I L L F V G A V G K S S Q L G L H 240
17938 CCT <u>T</u> GG <u>T</u> AC <u>T</u> CT <u>G</u> AT <u>G</u> CT <u>A</u> TA <u>G</u> A <u>-</u> GG <u>A</u> CC <u>T</u> AC <u>A</u> CC <u>G</u> AT <u>T</u> GT <u>C</u> AG <u>C</u> AT <u>A</u> TT <u>A</u> CT <u>T</u> CG <u>T</u> CA <u>T</u> AT <u>G</u> -GT <u>A</u> AC <u>A</u> CG <u>C</u> GG <u>T</u> GT <u>T</u> TC <u>-</u> TT <u>A</u> TA <u>A</u> 18024 241 T W L P D A M E G P T P V S A L I H A A T M V T A G V F L I 270
18025 T <u>TA</u> AG <u>A</u> TC <u>-</u> T <u>C</u> T <u>G</u> T <u>T</u> AT <u>T</u> TC <u>A</u> CA <u>G</u> T <u>C</u> CT <u>T</u> TA <u>T</u> AT <u>T</u> TC <u>-</u> AT <u>C</u> -AT <u>T</u> AT <u>T</u> GT <u>A</u> CT <u>T</u> GT <u>T</u> AT <u>G</u> GC <u>C</u> TA <u>A</u> U <u>-</u> AC <u>G</u> CT <u>A</u> AT <u>A</u> TT <u>T</u> CT <u>T</u> CT <u>T</u> TA 18111 271 A R C S P I F E Y A P T A L L V V T I V G A M T A F F A A T 300
18112 AC <u>A</u> GG <u>T</u> TA <u>-</u> TT <u>A</u> CA <u>A</u> AT <u>G</u> AC <u>A</u> AA <u>AC</u> G <u>T</u> AT <u>T</u> AT <u>-</u> GC <u>A</u> TT <u>T</u> CA <u>AC</u> CT <u>G</u> T <u>A</u> GC <u>C</u> AA <u>C</u> TT <u>G</u> TT <u>T</u> AT <u>G</u> T <u>T</u> AT <u>G</u> C <u>T</u> AT <u>-</u> GG <u>T</u> AT <u>T</u> GG <u>T</u> 18197 301 T G L L Q N D I K R V I A Y S T C S Q L G Y M V F A C G I S 330
18198 A <u>T</u> AT <u>A</u> ACT <u>T</u> IT <u>-</u> GT <u>T</u> TT <u>T</u> AT <u>T</u> CA <u>T</u> TT <u>A</u> GT <u>A</u> AA <u>A</u> AT <u>A</u> AT <u>T</u> GT <u>-</u> GT <u>T</u> TT <u>T</u> CT <u>A</u> TT <u>A</u> AA <u>G</u> CA <u>T</u> CT <u>T</u> TT <u>A</u> AT <u>T</u> GT <u>T</u> AT <u>G</u> C <u>T</u> ACC 18284 331 G Y S V G M F H L M N H A F F K A L L F L S A G C V I H A L 360
18285 GGC <u>-</u> CAT <u>C</u> AG <u>A</u> GT <u>T</u> AT <u>T</u> CG <u>G</u> GT <u>T</u> GG <u>G</u> GT <u>-</u> TT <u>T</u> TT <u>A</u> AA <u>G</u> A <u>-</u> TT <u>A</u> CC <u>C</u> ATA <u>A</u> CT <u>T</u> AT <u>T</u> GT <u>G</u> CA <u>A</u> GT <u>G</u> C <u>T</u> IT <u>-</u> G <u>C</u> T <u>T</u> TT <u>T</u> AT <u>C</u> TT <u>A</u> 18368 361 A D E Q D M R R M G G I V K I V P F T Y G M M L I G S M S L 390
18369 AT <u>T</u> GG <u>T</u> TT <u>T</u> CT <u>-</u> TC <u>T</u> TA <u>G</u> T <u>G</u> TT <u>T</u> TT <u>A</u> AG <u>-</u> AG <u>G</u> AT <u>T</u> TT <u>C</u> T <u>T</u> CA <u>G</u> A <u>AG</u> CT <u>A</u> CT <u>T</u> AC <u>A</u> AT <u>A</u> AT <u>T</u> TT <u>G</u> GT <u>T</u> AT <u>G</u> T <u>C</u> T <u>-</u> CT <u>T</u> AT <u>G</u> T <u>A</u> 18453 391 M G F P F L T G F Y S K D V I L E L A F A K Y T I D G T F A 420
18454 T <u>TT</u> AT <u>T</u> TA <u>T</u> AT <u>C</u> CA <u>T</u> AT <u>T</u> CT <u>A</u> CT <u>G</u> T <u>G</u> T <u>T</u> AG <u>T</u> TA <u>T</u> TT <u>A</u> CT <u>T</u> CT <u>T</u> CC <u>G</u> AT <u>A</u> TT <u>T</u> TT <u>T</u> TT <u>G</u> T <u>G</u> T <u>T</u> GT <u>T</u> CT <u>T</u> GG <u>T</u> G <u>A</u> TC <u>A</u> AT <u>C</u> AT <u>T</u> CA <u>A</u> 18543 421 H W L G T V A A F T A F Y S F R L I Y L T F L G E T N A P 450
18544 AAA <u>A</u> AC <u>A</u> TT <u>G</u> AA <u>A</u> CT <u>T</u> TT <u>C</u> T <u>A</u> GA <u>A</u> GC <u>T</u> C <u>T</u> AT <u>T</u> TT <u>C</u> T <u>A</u> T <u>A</u> ---CT <u>A</u> CA <u>-</u> TT <u>A</u> TT <u>T</u> TT <u>A</u> ACT <u>A</u> CT <u>T</u> CT <u>T</u> TT <u>C</u> T <u>G</u> AT <u>T</u> TT 18629 451 R T I I N H A H D A P F I M A L P L M I L A I G S I F V G F 480
18630 AT <u>T</u> TT <u>A</u> A <u>A</u> G <u>-</u> ---AT <u>G</u> A <u>T</u> ---G <u>A</u> CT <u>T</u> AT <u>-</u> TC <u>AA</u> AC <u>A</u> CT <u>T</u> T <u>A</u> T <u>A</u> AT <u>T</u> TT <u>A</u> TT <u>T</u> AG <u>G</u> C <u>T</u> CT <u>T</u> CT <u>T</u> TT <u>C</u> AG <u>T</u> G <u>A</u> T <u>A</u> CG <u>C</u> TA 18698 481 I M K D M M I G L G T D F W G N S L F T H P K N L T L I E S 510
18699 --CT <u>G</u> A <u>T</u> CA <u>A</u> G <u>A</u> TT <u>C</u> T <u>-</u> TT <u>A</u> ---AT <u>G</u> A <u>T</u> TT <u>A</u> AG <u>T</u> TT <u>A</u> CC <u>C</u> AC <u>A</u> AT <u>T</u> TT <u>T</u> CT <u>T</u> TT <u>A</u> TC <u>G</u> G <u>T</u> TT <u>A</u> CT <u>T</u> TT <u>T</u> AG <u>T</u> ---AT <u>A</u> T <u>A</u> AT <u>A</u> 18777 511 E F I P T P I K L L P V I L S I V G A S L A I I L N N F Y A 540
18778 TT <u>A</u> TT <u>G</u> AA <u>-</u> TT <u>A</u> AA <u>A</u> CT <u>T</u> AA <u>A</u> AG <u>T</u> CA <u>A</u> CT <u>T</u> AT <u>A</u> TA <u>G</u> AC <u>A</u> AT <u>T</u> CT <u>A</u> TT <u>T</u> CT <u>T</u> GT <u>A</u> AA <u>AA</u> AT <u>T</u> CT <u>T</u> GT <u>A</u> AT <u>G</u> C <u>T</u> TT <u>A</u> 18867 541 T F L V S L K T S L L G R E I Y S F L N K R - W Y F - D I V 568
18868 A <u>A</u> T <u>A</u> GT <u>T</u> TT <u>T</u> AT <u>T</u> TT <u>C</u> T <u>AC</u> CA <u>T</u> CT <u>G</u> CA <u>A</u> CT <u>T</u> TT <u>C</u> ---TT <u>A</u> AA <u>A</u> TA <u>A</u> CT <u>T</u> AA <u>A</u> AT <u>T</u> AT <u>A</u> GA <u>A</u> GT <u>G</u> T <u>-</u> G <u>T</u> TT <u>A</u> GA <u>A</u> CA <u>T</u> TT <u>T</u> AG <u>T</u> A 18953 569 Y N E Y V G K T L L W F G Y N I S F K S V D K G L I E I L G 598
18954 T <u>C</u> AC <u>G</u> GG <u>T</u> AT <u>T</u> AT <u>G</u> T <u>-</u> TT <u>T</u> CT <u>G</u> AA <u>-</u> TT <u>A</u> TT <u>T</u> GT <u>A</u> AA <u>G</u> T <u>T</u> TT <u>G</u> TA <u>A</u> GT <u>G</u> A <u>A</u> AC <u>A</u> CT <u>T</u> AA <u>A</u> AT <u>T</u> AT <u>C</u> G <u>T</u> TT <u>T</u> CT <u>T</u> TT <u>A</u> 19039 599 P Y G L E R L T R R L T S K V S A L Q T G Y I Y H Y A F I M 628
19040 A <u>T</u> AT <u>A</u> CT <u>T</u> TT <u>G</u> T <u>C</u> T <u>CG</u> GG <u>G</u> CT <u>A</u> C <u>-</u> TT <u>A</u> TA <u>A</u> TT <u>T</u> TT <u>A</u> TT <u>T</u> GT <u>G</u> ---TT <u>A</u> AT <u>A</u> CT <u>G</u> C <u>-</u> TT <u>T</u> T <u>C</u> T <u>-</u> TT <u>A</u> TT <u>T</u> AG <u>T</u> A 19117 629 L L G V T L I I T I I G L W D Y I S M W T D Y R L Y F L F L 658
19118 CT <u>T</u> TA <u>A</u> TA <u>T</u> TT <u>T</u> TA <u>A</u> TC <u>CG</u> AA <u>A</u> TC <u>A</u> CT <u>T</u> CA <u>A</u> AC 19153 659 L T I I F Y G Y S E N K 670

Fig. 2. nad5 protein sequence of Reclinomonas aligned to Physarum mitogenome: Due to the large evolutionary distance between the two species, a number of amino acids are aligned to dissimilar (yellow or red) codons. Due to the possibility of frameshifts (of which there are 35), the alignment is still very good. Note that the alignment matches alignment positions given in Fig. 1 and the reference in the REDBASE [4] (<http://bioserv.mps.ohio-state.edu/redbase/>) for nad5.

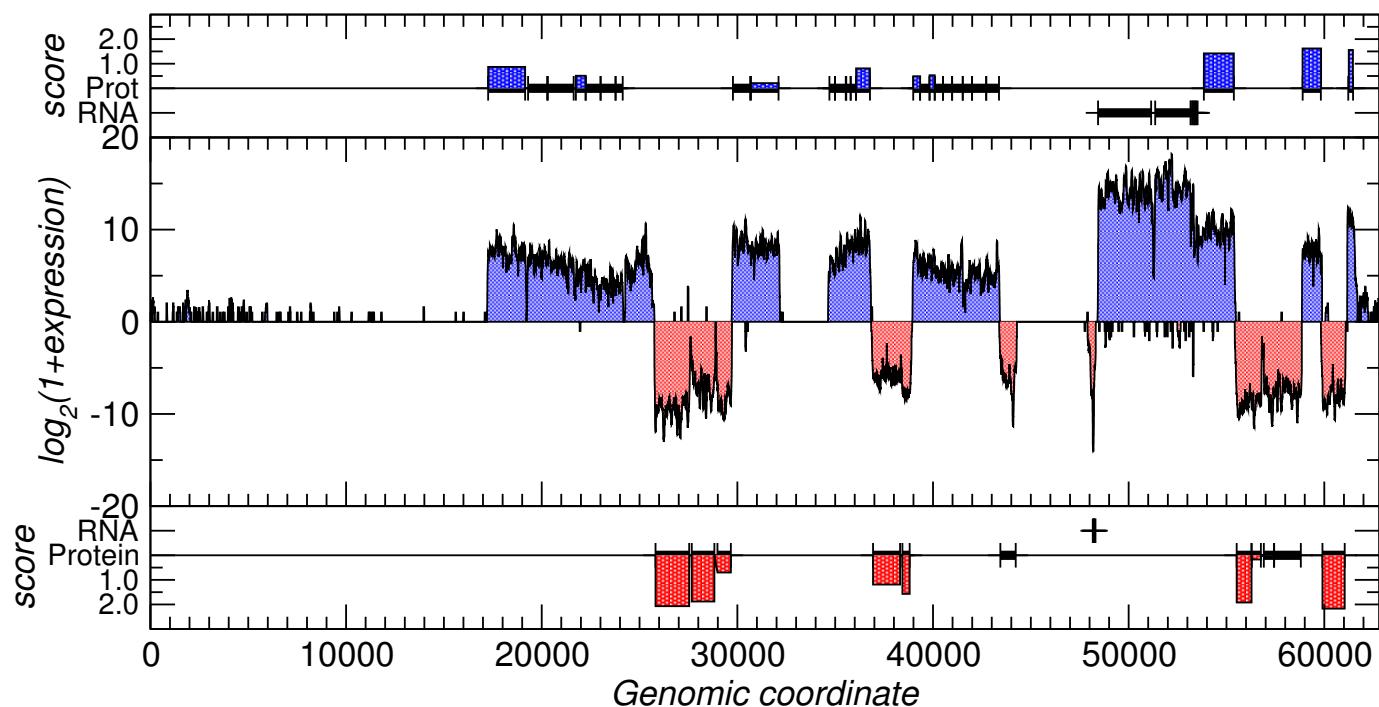


Fig. 3. Alignment of *R. americana* proteins to the *P. polycephalum* mitogenome. The central panel displays expression data from [4]. Above and below the known protein-coding (P) and ncRNA (R) genes are shown (thick black lines with delimiters for each gene) together with the alignment scores (normalized per nucleotide) for the *R. americana* proteins.