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 $A \to aBC \mid bD \mid c$ 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:

$$\begin{aligned}
&\left(\{A \to aBC|bD|c\} \circledcirc \{A \to aBC|bD|c\}\right) \circledcirc \\
&\{A \to aBC|bD|c\} \\
=& \\
&\{A \to aBC|bD|c\} \circledcirc \\
&\left(\{A \to aBC|bD|c\} \circledcirc \{A \to aBC|bD|c\}\right)
\end{aligned}$$

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

$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ a \\ a \end{pmatrix} \begin{pmatrix} B \\ B \\ B \end{pmatrix} \begin{pmatrix} C \\ C \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ b \end{pmatrix} \begin{pmatrix} D \\ D \\ \epsilon \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ D \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ a \\ b \end{pmatrix} \begin{pmatrix} B \\ B \\ D \end{pmatrix} \begin{pmatrix} C \\ C \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ b \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ D \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ b \\ b \end{pmatrix} \begin{pmatrix} B \\ B \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ C \\ D \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ b \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ \epsilon \end{pmatrix} \begin{pmatrix} D \\ D \\ D \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ a \\ c \end{pmatrix} \begin{pmatrix} B \\ B \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ C \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ \epsilon \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ b \\ a \end{pmatrix} \begin{pmatrix} B \\ D \\ B \end{pmatrix} \begin{pmatrix} C \\ \epsilon \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ \epsilon \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ D \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ b \\ a \end{pmatrix} \begin{pmatrix} B \\ \epsilon \\ B \end{pmatrix} \begin{pmatrix} C \\ D \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ D \end{pmatrix} \begin{pmatrix} D \\ D \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ b \\ b \end{pmatrix} \begin{pmatrix} B \\ D \\ D \end{pmatrix} \begin{pmatrix} C \\ \epsilon \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ c \end{pmatrix} \begin{pmatrix} D \\ D \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ b \\ b \end{pmatrix} \begin{pmatrix} B \\ D \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ \epsilon \\ D \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ c \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ \epsilon \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ b \\ b \end{pmatrix} \begin{pmatrix} B \\ \epsilon \\ D \end{pmatrix} \begin{pmatrix} C \\ D \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ c \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ \epsilon \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ b \\ b \end{pmatrix} \begin{pmatrix} B \\ \epsilon \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ D \\ D \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ c \\ a \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ B \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ C \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ b \\ c \end{pmatrix} \begin{pmatrix} B \\ D \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ \epsilon \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ c \\ a \end{pmatrix} \begin{pmatrix} \epsilon \\ B \\ \epsilon \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ C \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ b \\ c \end{pmatrix} \begin{pmatrix} B \\ \epsilon \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ D \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ c \\ b \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ D \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ c \\ a \end{pmatrix} \begin{pmatrix} B \\ \epsilon \\ B \end{pmatrix} \begin{pmatrix} C \\ \epsilon \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ c \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ D \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ c \\ b \end{pmatrix} \begin{pmatrix} B \\ \epsilon \\ D \end{pmatrix} \begin{pmatrix} C \\ \epsilon \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ c \\ c \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ c \\ b \end{pmatrix} \begin{pmatrix} B \\ \epsilon \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ \epsilon \\ D \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ c \\ c \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} a \\ c \\ c \end{pmatrix} \begin{pmatrix} B \\ \epsilon \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ \epsilon \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ a \\ a \end{pmatrix} \begin{pmatrix} \epsilon \\ B \\ B \end{pmatrix} \begin{pmatrix} \epsilon \\ C \\ C \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ a \\ a \end{pmatrix} \begin{pmatrix} D \\ B \\ B \end{pmatrix} \begin{pmatrix} \epsilon \\ C \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ a \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ B \\ D \end{pmatrix} \begin{pmatrix} \epsilon \\ C \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ a \\ a \end{pmatrix} \begin{pmatrix} \epsilon \\ B \\ B \end{pmatrix} \begin{pmatrix} D \\ C \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ a \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ B \\ \epsilon \end{pmatrix} \begin{pmatrix} \epsilon \\ C \\ D \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ a \\ b \end{pmatrix} \begin{pmatrix} D \\ B \\ D \end{pmatrix} \begin{pmatrix} \epsilon \\ C \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ b \\ a \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ B \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ C \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ a \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ B \\ D \end{pmatrix} \begin{pmatrix} D \\ C \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ b \\ a \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ B \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ C \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ a \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ B \\ \epsilon \end{pmatrix} \begin{pmatrix} D \\ C \\ D \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ b \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ D \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ a \\ c \end{pmatrix} \begin{pmatrix} D \\ B \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ C \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ b \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ \epsilon \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ D \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ a \\ c \end{pmatrix} \begin{pmatrix} \epsilon \\ B \\ \epsilon \end{pmatrix} \begin{pmatrix} C \\ C \\ \epsilon \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ b \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ \epsilon \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ D \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ a \end{pmatrix} \begin{pmatrix} D \\ D \\ B \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ b \\ c \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ \epsilon \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ a \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ B \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ c \\ a \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ B \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ C \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ a \end{pmatrix} \begin{pmatrix} \epsilon \\ D \\ B \end{pmatrix} \begin{pmatrix} D \\ \epsilon \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ c \\ b \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ D \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ a \end{pmatrix} \begin{pmatrix} \epsilon \\ \epsilon \\ B \end{pmatrix} \begin{pmatrix} D \\ D \\ C \end{pmatrix} \quad \begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} c \\ c \\ c \end{pmatrix}$$
$$\begin{pmatrix} A \\ A \\ A \end{pmatrix} \to \begin{pmatrix} b \\ b \\ b \end{pmatrix} \begin{pmatrix} D \\ D \\ D \end{pmatrix}$$

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 $\binom{F_0}{P}, \binom{F_1}{P}, \binom{F_2}{P}$ denoting three different reading frames for the DNA sequence F, as well as $, \binom{L}{P}, \binom{R}{P}$ 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 $\binom{R}{P}$.

$$\binom{F_0}{P} \to \binom{F_0}{P}\binom{\varepsilon}{a} \qquad \binom{F_1}{P} \to \binom{F_1}{P}\binom{c}{\varepsilon}\binom{c}{\varepsilon}\binom{c}{\varepsilon}$$
$$\binom{F_0}{P} \to \binom{L}{P} \qquad \binom{F_2}{P} \to \binom{F_2}{P}\binom{\varepsilon}{a}$$
$$\binom{F_0}{P} \to \binom{\$}{\$} \qquad \binom{F_2}{P} \to \binom{L}{P}$$
$$\binom{F_0}{P} \to \binom{F_1}{P}\binom{c}{a}\binom{c}{\varepsilon} \qquad \binom{F_2}{P} \to \binom{\$}{\$}$$
$$\binom{F_0}{P} \to \binom{F_1}{P}\binom{c}{\varepsilon}\binom{c}{\varepsilon} \qquad \binom{F_2}{P} \to \binom{F_0}{P}\binom{c}{a}\binom{c}{\varepsilon}$$
$$\binom{F_0}{P} \to \binom{F_2}{P}\binom{c}{a} \qquad \binom{F_2}{P} \to \binom{F_0}{P}\binom{c}{\varepsilon}\binom{c}{\varepsilon}$$
$$\binom{F_0}{P} \to \binom{F_2}{P}\binom{c}{\varepsilon} \qquad \binom{F_2}{P} \to \binom{F_1}{P}\binom{c}{a}$$
$$\binom{F_0}{P} \to \binom{F_0}{P}\binom{c}{a}\binom{c}{\varepsilon}\binom{c}{\varepsilon} \qquad \binom{F_2}{P} \to \binom{F_1}{P}\binom{c}{\varepsilon}$$
$$\binom{F_0}{P} \to \binom{F_0}{P}\binom{c}{\varepsilon}\binom{c}{\varepsilon}\binom{c}{\varepsilon} \qquad \binom{F_2}{P} \to \binom{F_2}{P}\binom{c}{a}\binom{c}{\varepsilon}\binom{c}{\varepsilon}$$
$$\binom{F_1}{P} \to \binom{F_1}{P}\binom{\varepsilon}{a} \qquad \binom{F_2}{P} \to \binom{F_2}{P}\binom{c}{\varepsilon}\binom{c}{\varepsilon}\binom{c}{\varepsilon}$$
$$\binom{F_1}{P} \to \binom{L}{P} \qquad \binom{L}{P} \to \binom{L}{P}\binom{c}{\varepsilon}$$
$$\binom{F_1}{P} \to \binom{\$}{\$} \qquad \binom{L}{P} \to \binom{\$}{\$}$$
$$\binom{F_1}{P} \to \binom{F_2}{P}\binom{c}{a}\binom{c}{\varepsilon} \qquad \binom{R}{P} \to \binom{R}{P}\binom{c}{\varepsilon}$$
$$\binom{F_1}{P} \to \binom{F_2}{P}\binom{c}{\varepsilon}\binom{c}{\varepsilon} \qquad \binom{R}{P} \to \binom{F_0}{P}$$
$$\binom{F_1}{P} \to \binom{F_0}{P}\binom{c}{a} \qquad \binom{R}{P} \to \binom{F_1}{P}$$
$$\binom{F_1}{P} \to \binom{F_0}{P}\binom{c}{\varepsilon} \qquad \binom{R}{P} \to \binom{F_2}{P}$$
$$\binom{F_1}{P} \to \binom{F_1}{P}\binom{c}{a}\binom{c}{\varepsilon}\binom{c}{\varepsilon} \qquad \binom{R}{P} \to \binom{\$}{\$}$$

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

As a pilot study for the frameshift DNA-Protein alignment, we aligned the mitochiondrial 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` insertions[1].

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	44229-43440	—	
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 refence 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 $\binom{c_1}{a}\binom{c_2}{}\binom{-}{}$ 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 17257
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 AATGTTTTCATGTTTCCTTA-ATAGCATTTATAATTCTTTTTATGTTAGGTAGACATCTGGGAAG-CAAATTGT-CTCGGCTTTGCTATT    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 ACAATGTCATTTTA-TCACTTATAATTGT-TTATATTATTTTATTCATGTTTTTTTTAT-GGTCAAATATA-AGCTTTAATTTAGGTTCT    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 TGGGTTTCTGTAGGTACTTTAGATATTACTAC-AAATTTATAAT-GATCCTTTATCCATTACTTT-GGTACATTAATTCC-TTTATTACT    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 TTATTAATCTA-ATTTATTCTTATGATAT-TTACATGAAGATCCTAATTA-GTTAAATTTTTTGCTTATTTAGT-TTTTTCTCTTTTTCT    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 ATGTT-TGTCTTGTTTTTGCTGGTAATTACTT-ATTATGTTTTTAGGATGGGAAGCTGTGGA-TTAGCTTCTTATTT-CTTATTAATTTC    17688
     121 M  S  C  L  V  F  A  G  N  Y  F  I  M  F  L  G  W  E  A  V  G  L  A  S  Y  L  L  I  N  F      150

   17689 TGGAT-ACAAGAAATCAAGCAAT-CAGTCTGCTATTAAAGCAATTATTTTTAATCGTGA-GGTGATGCAGCTTTCATAAGTGCTATGGGT    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 CTTATTTATTATTT-TTTAATTCTTT-GATTTAGAAGATTTAGAATTACTTGTTCA-CAATATGAACATACTACTTTTAGCTT-TTTTCA    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 TATTCTTTTCA-ACAATTGAATG-ATAGCTCTTTTCCTATTTTTGCT-GCTGT-GCTAAATCAGCACAATC-TTTTTACATCCTTGGTTA    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 CCTGATGCTATGAA-GGACCTACACCAGTTTCAGCATTATTACATTCTGCTACATG-GTAACAGCAGGTGTTTC-TTAATATTAAGATCT    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-GTTATTTTCTCACATGCTCCTTATATTTCATTATTGTA-GCTTGTATTGGCCTAATACA-GCTAATATTTCTTCTTTAACAGGTTA-    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 TTACAATATGACATAAAACGTATTAT-GCATTTTCAACCTGTAGCCAACTTGGTTT-ATGATGTTTGCTAT-GGTATTGGTAATTATACT    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-GCTTTATTTCATTTAGTAAATAT-GCTTTCTTTAAAGCACTCTTATTTTA-TGTGCCGGATCCGTTATCCACGCTACCGGCCATCAG    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 GATATTCGGCGTATGGGAGT-TTATTTAAAGA-TTACCCATAACTTATGTTGCAATGCTCTTGCT-TCTTTATCTTTAATTGGTTTTCT-    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 TTCTTAAGTGGTTTTTATAG-AAGGATTTTCTTCTAGAAGCTACTTACAATATATTTGGTATGTC-TCTTATGTTATTTATTTTATATCC    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-ATTTCTACTGCTGTTAGTTA-TTTTACTCTTTCCGATTAATTTTCTTTGTGTTCTTGGT-GATCAAGCATA-TCAATAAAAACATTG    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 AAAACTATTCT-GAAAGCTCCTATTTTCTATATCTACA-TTAATTATTTTAACTATACTTTCTATTTTTCT-GGATTTTATTTAAAAGAT    18640
     451 K  T  I  S  E  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-ATGACTATTCAAACATCTTA-TATAATTTTAGTGCCTCTCT-TTCAGTGATACAGCTACTGATCAAGATTC-TTTAATGATTTATTT    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 AAAGTTTA-CCCACAATCTTTTCTTTATCGGT-TTACTTTTAGTATATATAATTTATTG-AAATTAAACTTAAAAAGTCAACTATTATAT    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 AGACAATAC-TTATTACCTTATTT-TCTTGTAAAAAATTCTTTGA-GATGCTTTTAATAGTTTTTATATTTC-TTACCATCTGCAACT    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 TTTCT-TTAAATATACT-TATAAAATTATAGATAA-GGTGTTTTAGAACATTTAGGTTCAACGGGTATTAT-GCTTTTCTTGAATT-ATT    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 TTTGAAAGTATTGTTAATGTGAA-ACAACCTTAATAATTAT-CGTTCTTTTCTTTTCATAATCTTTGTCCTCGGAGCTACATTATA-ATT    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 TTTTATGGTTTATACGCTTTTCTTATAGCAATTTTCTTTATTAGTACTCTTAATATTTTTAATACCGAAATCACTTCA    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.

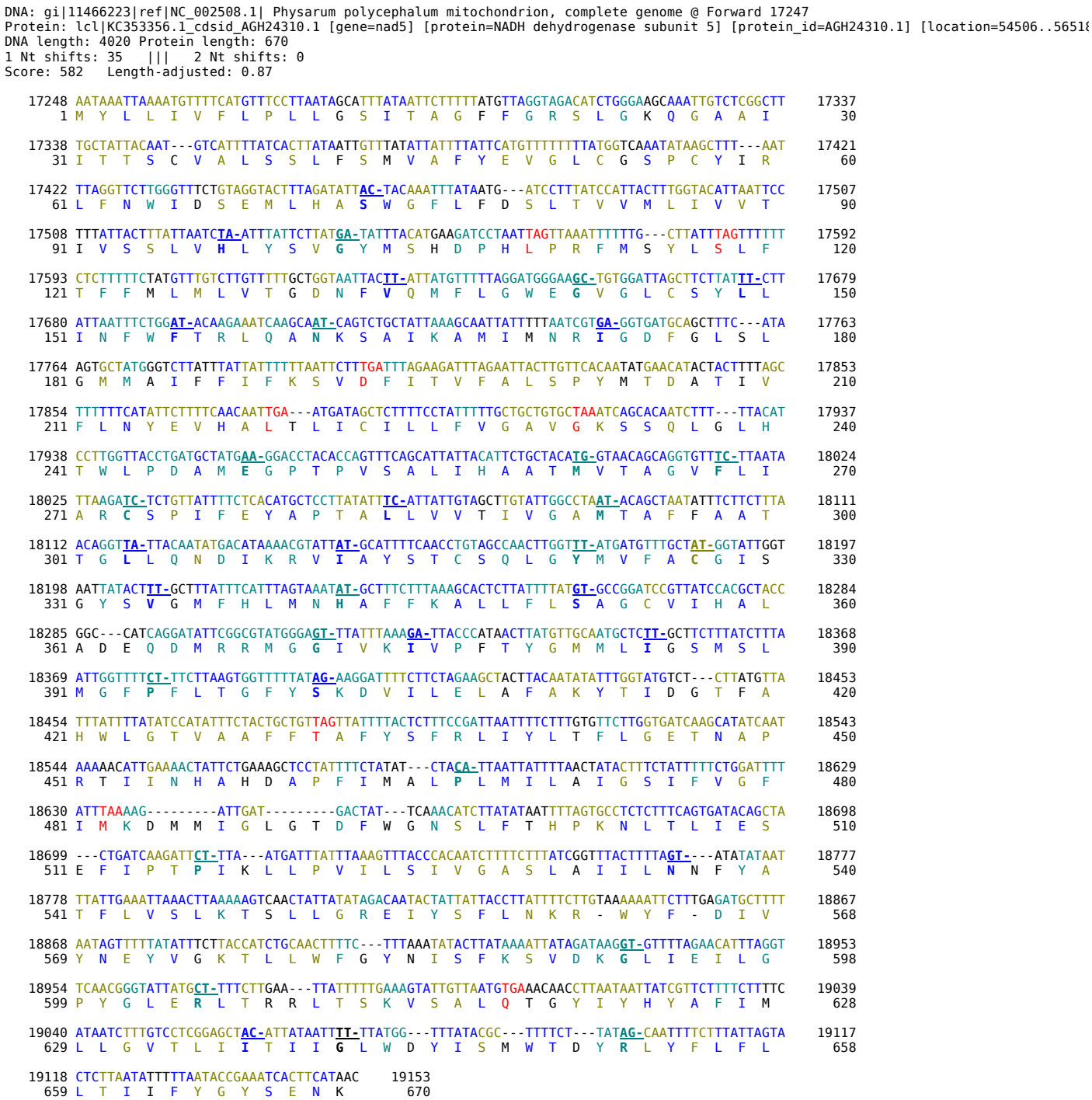

Fig. 2. **nad5 protein sequence of Reclinomonas aligned to Physarum mitogenome:** Due to the large evolutionary distance between the two species, a number of amino acids are aligned to dissimilar (yellow or red) codons. Due to the possibility of frameshifts (of which there are 35), the alignment is still 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.

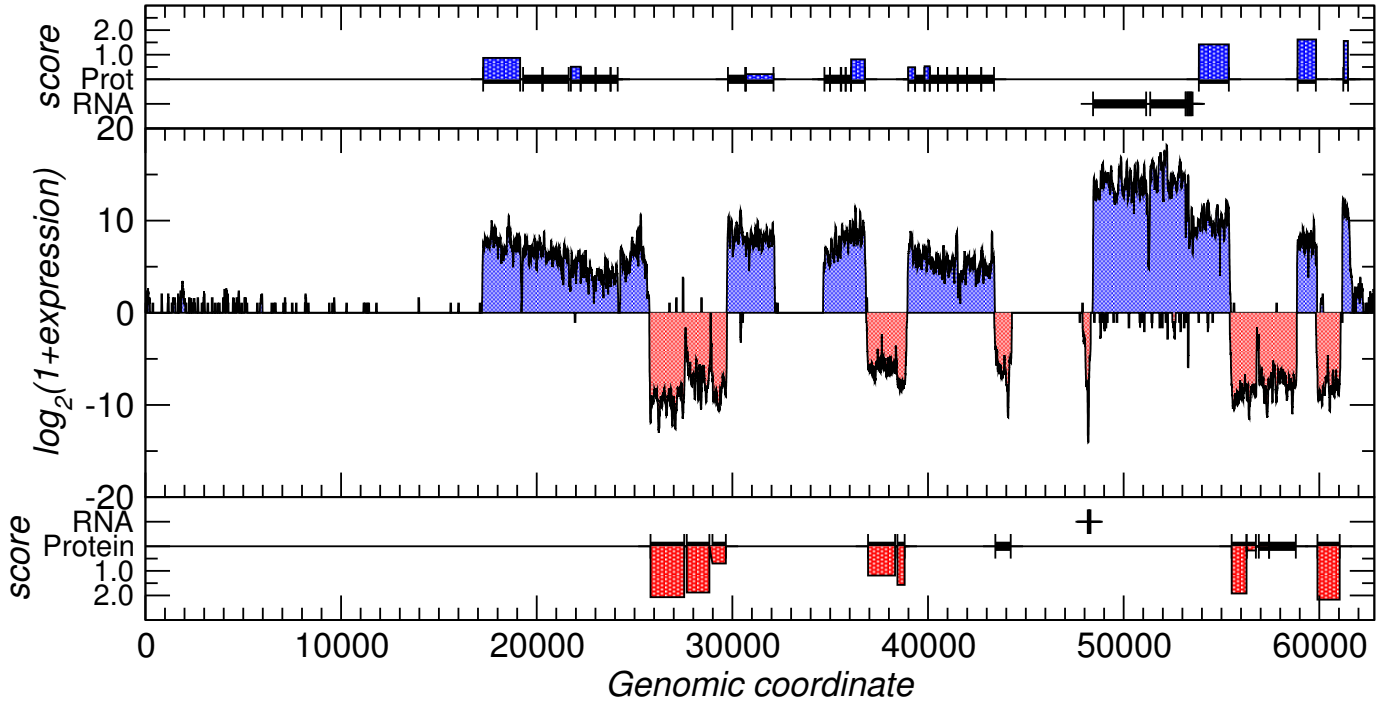

Fig. 3. Alignment of *R. americana* proteins to the *P. polycephalum* mitogenome. The central panel displays expression data from [4]. Above and below the known protein-coding (P) and ncRNA (R) genes are shown (thick black lines with delimiters for each gene) together with the alignment scores (normalized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