**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:

\[
\{ A \to aBC | bD | c \} \odot \{ A \to aBC | bD | c \}
\]

The full semi-global grammar contains a large number of symbols and rules. There are 5 non-terminal symbols \((P_1), (P_2), (P_3)\) denoting three different reading frames for the DNA sequence \( F \), as well as \((P_1), (P_2)\) 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 \( \ell \) terminal acts as the sentinel parsing the empty subtring. 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. $3 \times 9$ 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 $(\delta)$. 

$$
(F_F, p) \rightarrow (F_F, p) (\epsilon) (\delta) \\
(F_F, p) \rightarrow (F_F, p) (\epsilon) (\delta) \\
(F_F, p) \rightarrow (F_F, p) (\epsilon) (\delta) \\
(F_F, p) \rightarrow (F_F, p) (\epsilon) (\delta) \\
(F_F, p) \rightarrow (F_F, p) (\epsilon) (\delta) \\
(F_F, p) \rightarrow (F_F, p) (\epsilon) (\delta) \\
(F_F, p) \rightarrow (F_F, p) (\epsilon) (\delta) \\
(F_F, p) \rightarrow (F_F, p) (\epsilon) (\delta) \\
(F_F, p) \rightarrow (F_F, p) (\epsilon) (\delta)
$$

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 $g$ 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.

<table>
<thead>
<tr>
<th>gene</th>
<th>REDBASE</th>
<th>predicted</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>nad5</td>
<td>17259-19152</td>
<td>17248-19152</td>
<td>no homolog</td>
</tr>
<tr>
<td>nad6</td>
<td>19300-20316</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps2</td>
<td>20278-21633</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps12</td>
<td>21746-22241</td>
<td>21735-22109</td>
<td></td>
</tr>
<tr>
<td>rps7</td>
<td>22246-23009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rpl2</td>
<td>23009-23777</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps19</td>
<td>23774-24139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>php15</td>
<td>24441-25471</td>
<td></td>
<td>no homolog</td>
</tr>
<tr>
<td>cox1</td>
<td>25734-25816</td>
<td>20743-25950</td>
<td></td>
</tr>
<tr>
<td>nad7</td>
<td>25760-27353</td>
<td>27674</td>
<td></td>
</tr>
<tr>
<td>cox2</td>
<td>29666-29893</td>
<td>29698-29894</td>
<td></td>
</tr>
<tr>
<td>php22</td>
<td>29776-30652</td>
<td></td>
<td>no homolog</td>
</tr>
<tr>
<td>nad1</td>
<td>30693-32105</td>
<td>30706-32133</td>
<td></td>
</tr>
<tr>
<td>rps16</td>
<td>34704-34988</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rpl19</td>
<td>34988-35540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atp6</td>
<td>35567-35788</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nad4L</td>
<td>35788-36062</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alp6</td>
<td>36067-36774</td>
<td>36059-36771</td>
<td></td>
</tr>
<tr>
<td>nad3</td>
<td>38808-38435</td>
<td>38802-38450</td>
<td></td>
</tr>
<tr>
<td>rpl14</td>
<td>38895-39338</td>
<td>38981-39336</td>
<td></td>
</tr>
<tr>
<td>php23</td>
<td>39328-39822</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps14</td>
<td>39823-40087</td>
<td>39798-40084</td>
<td></td>
</tr>
<tr>
<td>rps8</td>
<td>40088-40509</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rpl6</td>
<td>40506-40972</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps13</td>
<td>40978-41517</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nad9</td>
<td>41521-41994</td>
<td></td>
<td>me</td>
</tr>
<tr>
<td>rps11</td>
<td>41997-42717</td>
<td></td>
<td>me</td>
</tr>
<tr>
<td>php24</td>
<td>42721-43372</td>
<td></td>
<td>no homolog</td>
</tr>
<tr>
<td>rps4</td>
<td>44229-43440</td>
<td></td>
<td></td>
</tr>
<tr>
<td>php25</td>
<td>53565-53858</td>
<td></td>
<td>no homolog</td>
</tr>
<tr>
<td>atpA</td>
<td>53845-53850</td>
<td>53854-53919</td>
<td></td>
</tr>
<tr>
<td>cox3</td>
<td>56278-55517</td>
<td>56287-55526</td>
<td></td>
</tr>
<tr>
<td>nad6</td>
<td>56758-56280</td>
<td>56780-56187</td>
<td></td>
</tr>
<tr>
<td>rpl16</td>
<td>57434-56910</td>
<td>57338-56948</td>
<td></td>
</tr>
<tr>
<td>rps3</td>
<td>58800-57431</td>
<td></td>
<td>me</td>
</tr>
<tr>
<td>nad1</td>
<td>58893-59821</td>
<td>58889-59834</td>
<td></td>
</tr>
<tr>
<td>cytB</td>
<td>61038-59903</td>
<td>61043-59921</td>
<td></td>
</tr>
<tr>
<td>ap9</td>
<td>61225-61467</td>
<td>61250-61464</td>
<td></td>
</tr>
</tbody>
</table>

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 $(\epsilon) (\epsilon) (\epsilon)$ 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

**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 genomic positions.
SUPPLEMENTAL MATERIAL

DNA: gi|11466232|ref|NC_002508.1| Physarum polycephalum mitochondrial, complete genome @ Forward 17247

Protein: lcl|KC533536.1.cdsid.AGH24310.1| [gene=nad5] [protein=NADH dehydrogenase subunit 5] [protein_id=AGH24310.1] [location=54506..5651f]

DNA length: 4029 Protein length: 670

3 nt shifts: 35 35 35 2 2 nt shifts: 0

Score: 582 Length-adjusted: 0.87

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.