Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome

(genomics/algorithm/inversions/edit distance/conserved segments)

DAVID SANKOFF*, GUILLAUME LEDUC*, NATALIE ANTOINE†, BRUNO PAQUIN†, B. FRANZ LANG†, AND ROBERT CEDERGREN†

Presentation by Julie Hudson
MAT5313
Definitions

- **Gene order**: permutation of genomic arrangement
- **Phylogenetics**: the study of evolutionary history and finding genetic connection between species
- **Mitochondrial genome**: complete set of genes specific to the mitochondria that guide its function

Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome
Research Question

Can we infer evolutionary history from the arrangement of genes in the mitochondria of various species?
Introduction

• Evolutionary inference traditionally done via comparison of homologous versions of a single gene
• mtDNA is susceptible to mutation (rapid nucleotide substitution) making homology difficult to differentiate from noise
• Genome level analysis will be more robust to this mutation
• Analytic tools different when comparing gene vs genome level similarity
Data

- Mitochondrial genome sequences from:
  - Fungi (8)
  - Animals (7)
  - Protists (1)
With 31-50 mitochondrial genes (humans have 37)

- Why mitochondrial?

  Complete genome consists of a small number of genes

<table>
<thead>
<tr>
<th>Genome</th>
<th>Genes, no.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fungi</strong></td>
<td></td>
</tr>
<tr>
<td>Fission yeast</td>
<td></td>
</tr>
<tr>
<td>Schizosaccharomyces pombe (6)*</td>
<td>35</td>
</tr>
<tr>
<td>Budding yeast</td>
<td></td>
</tr>
<tr>
<td>Torulopsis glabrata (7)</td>
<td>34</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (8)</td>
<td>39</td>
</tr>
<tr>
<td>Kluyveromyces lactis (9)</td>
<td>31</td>
</tr>
<tr>
<td>Filamentous Ascomycetes</td>
<td></td>
</tr>
<tr>
<td>Neurospora crassa (10)</td>
<td>50</td>
</tr>
<tr>
<td>Aspergillus nidulans (11)</td>
<td>44</td>
</tr>
<tr>
<td>Podospora anserina (12)</td>
<td>45</td>
</tr>
<tr>
<td>Chytridiomycetes</td>
<td></td>
</tr>
<tr>
<td>Allomyces macrogynus</td>
<td></td>
</tr>
<tr>
<td>(B.P. and B.F.L., unpublished data)</td>
<td>35</td>
</tr>
<tr>
<td><strong>Protist</strong></td>
<td></td>
</tr>
<tr>
<td>Phytophthora infestans</td>
<td></td>
</tr>
<tr>
<td>(B.F.L., unpublished data)</td>
<td>31</td>
</tr>
<tr>
<td><strong>Animals</strong></td>
<td></td>
</tr>
<tr>
<td>Vertebrates</td>
<td></td>
</tr>
<tr>
<td>Mammalia (13, 14)</td>
<td>37</td>
</tr>
<tr>
<td>Gallus gallus (chicken) (15)</td>
<td>37</td>
</tr>
<tr>
<td>Echinoderms</td>
<td></td>
</tr>
<tr>
<td>Strongylocentrotus purpuratus (sea urchin) (16)</td>
<td>37</td>
</tr>
<tr>
<td>Asterina pectinifera (star fish) (17)</td>
<td>37</td>
</tr>
<tr>
<td>Pisaster ochraceus (sea star) (18)</td>
<td>36</td>
</tr>
<tr>
<td>Insect</td>
<td></td>
</tr>
<tr>
<td>Drosophila yakuba (19)</td>
<td>37</td>
</tr>
<tr>
<td>Nematode</td>
<td></td>
</tr>
<tr>
<td>Ascaris suum (20)</td>
<td>36</td>
</tr>
</tbody>
</table>
Methods

- Edit Distance (noted as $E(a,b)$) is the number of elementary events necessary to change the gene order of one circular genome $a$ into that of $b$.
- This is how gene arrangement is measured.
- Elementary events: deletion, insertion, inversion, transposition.
- $E(a,b) = D(a,b) + R(a,b)$
Methods

• $D(a, b) = \text{total number of genes present in only one of } a \text{ or } b.$
• (D for deletion/insertion)
• Simple to determine
• $a : 1, 2, 3, 4, 5 ; \quad b : 1, 3, 4, 5, 6$
• $D(a, b) = 2$
Methods

- \( R(a, b) \) = minimal number of inversion and transposition events necessary to convert one to the other, ignoring missing genes (D)
- Not so straightforward
- Based on a “conserved chromosomal segment” counting technique by Nadeau and Taylor, C
- \( a : 1, 2, 3, 4, 5 ; \quad b : 1, 3, 2, 4, 5 \)
- \( C = 3 \) (what is expected from 1 inversion event)
- Differs if inversion occurs at the end or coincides with a previous inversion event.
- \( a : 1, 2, 3, 4, 5 ; \quad b: 1, 2, 3, 5, 4 \)
- \( C = 2 \)
- C is no greater than \( 2R \) in a circular genome
Methods

Alignment Reduction

- Rearrangement Distance (R) determined through a branch-and-bound search using the program DERANGE.
- A series of alignment reductions and inversion/transposition events until the alignment is completely reduced (one link)

When pairs of genes are adjacent in both genomes with same orientation and order, they can be combined because the minimum number of recombinatory events will be the same.
Conserved segments can be reduced
Methods

Three-inversion solution

DERANGE runs up to 5000 paths consecutively to determine which set of events returns the minimal rearrangement distance (where the bound-and-search is used). Paths are discarded if they cannot lead to a minimal value or if the intermediate genome is determined to be probabilistically unlikely.

Example determining the minimal event distance through inversions and alignment reductions
## Results

### Table 2 from paper: Distance between genome pairs

<table>
<thead>
<tr>
<th></th>
<th>Mam</th>
<th>Gal</th>
<th>Str</th>
<th>Ast</th>
<th>Pis</th>
<th>Dro</th>
<th>Asc</th>
<th>Phy</th>
<th>All</th>
<th>Sch</th>
<th>Tor</th>
<th>Klu</th>
<th>Sac</th>
<th>Asp</th>
<th>Neu</th>
<th>Pod</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mam</td>
<td>19</td>
<td>16</td>
<td>13</td>
<td>25</td>
<td>12</td>
<td>18</td>
<td>21</td>
<td>17</td>
<td>16</td>
<td>19</td>
<td>23</td>
<td>26</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gal</td>
<td>0</td>
<td>19</td>
<td>17</td>
<td>12</td>
<td>6</td>
<td>22</td>
<td>21</td>
<td>17</td>
<td>17</td>
<td>19</td>
<td>23</td>
<td>24</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Str</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>26</td>
<td>27</td>
<td>13</td>
<td>21</td>
<td>19</td>
<td>16</td>
<td>20</td>
<td>24</td>
<td>27</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ast</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>22</td>
<td>25</td>
<td>13</td>
<td>20</td>
<td>16</td>
<td>14</td>
<td>18</td>
<td>18</td>
<td>24</td>
<td>25</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pis</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>23</td>
<td>24</td>
<td>12</td>
<td>17</td>
<td>20</td>
<td>17</td>
<td>16</td>
<td>19</td>
<td>24</td>
<td>24</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dro</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>28</td>
<td>11</td>
<td>19</td>
<td>21</td>
<td>17</td>
<td>19</td>
<td>17</td>
<td>26</td>
<td>26</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asc</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>15</td>
<td>14</td>
<td>13</td>
<td>12</td>
<td>16</td>
<td>20</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phy</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>28</td>
<td>25</td>
<td>26</td>
<td>25</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>11</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>14</td>
<td>13</td>
<td>12</td>
<td>13</td>
<td>30</td>
<td>15</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>17</td>
<td>17</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Sch</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>18</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>28</td>
<td>18</td>
<td>15</td>
<td>15</td>
<td>18</td>
<td>18</td>
<td>19</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Tor</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>19</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>29</td>
<td>19</td>
<td>11</td>
<td>10</td>
<td>15</td>
<td>12</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klu</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>20</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>28</td>
<td>20</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sac</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>32</td>
<td>24</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>13</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td>12</td>
<td>31</td>
<td>13</td>
<td>16</td>
<td>17</td>
<td>21</td>
<td>10</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neu</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>19</td>
<td>16</td>
<td>15</td>
<td>16</td>
<td>35</td>
<td>19</td>
<td>17</td>
<td>22</td>
<td>23</td>
<td>25</td>
<td>14</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Pod</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>14</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>30</td>
<td>16</td>
<td>14</td>
<td>19</td>
<td>20</td>
<td>22</td>
<td>11</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

- **Animal**
- **Protist**
- **Fungus**
## Results

### Edit Distance (D + R)

<table>
<thead>
<tr>
<th></th>
<th>Mam</th>
<th>Gal</th>
<th>Str</th>
<th>Ast</th>
<th>Pis</th>
<th>Dro</th>
<th>Asc</th>
<th>Phy</th>
<th>All</th>
<th>Sch</th>
<th>Tor</th>
<th>Klu</th>
<th>Sac</th>
<th>Asp</th>
<th>Neu</th>
<th>Pod</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mam</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gal</td>
<td>18</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Str</td>
<td>20</td>
<td>21</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ast</td>
<td>20</td>
<td>18</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pis</td>
<td>13</td>
<td>12</td>
<td>26</td>
<td>26</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dro</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>30</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asc</td>
<td>38</td>
<td>39</td>
<td>39</td>
<td>41</td>
<td>37</td>
<td>37</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phy</td>
<td>30</td>
<td>34</td>
<td>33</td>
<td>34</td>
<td>30</td>
<td>31</td>
<td>28</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>35</td>
<td>35</td>
<td>33</td>
<td>34</td>
<td>30</td>
<td>35</td>
<td>29</td>
<td>38</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sch</td>
<td>34</td>
<td>34</td>
<td>36</td>
<td>33</td>
<td>34</td>
<td>31</td>
<td>39</td>
<td>33</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tor</td>
<td>34</td>
<td>35</td>
<td>34</td>
<td>38</td>
<td>33</td>
<td>37</td>
<td>32</td>
<td>36</td>
<td>33</td>
<td>23</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klu</td>
<td>39</td>
<td>39</td>
<td>40</td>
<td>38</td>
<td>37</td>
<td>33</td>
<td>42</td>
<td>38</td>
<td>28</td>
<td>15</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sac</td>
<td>24</td>
<td>34</td>
<td>35</td>
<td>37</td>
<td>36</td>
<td>37</td>
<td>28</td>
<td>43</td>
<td>30</td>
<td>31</td>
<td>28</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>41</td>
<td>39</td>
<td>42</td>
<td>44</td>
<td>40</td>
<td>41</td>
<td>26</td>
<td>46</td>
<td>36</td>
<td>34</td>
<td>38</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neu</td>
<td>37</td>
<td>36</td>
<td>39</td>
<td>33</td>
<td>37</td>
<td>27</td>
<td>45</td>
<td>32</td>
<td>34</td>
<td>33</td>
<td>37</td>
<td>21</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pod</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Animal**
- **Protist**
- **Fungus**
### Results

#### Edit Distance (D + R)

<table>
<thead>
<tr>
<th></th>
<th>Mam</th>
<th>Gal</th>
<th>Str</th>
<th>Ast</th>
<th>Pis</th>
<th>Dro</th>
<th>Asc</th>
<th>Phy</th>
<th>All</th>
<th>Sch</th>
<th>Tor</th>
<th>Klu</th>
<th>Sac</th>
<th>Asp</th>
<th>Neu</th>
<th>Pod</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gal</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Str</td>
<td>18</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ast</td>
<td>20</td>
<td>21</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pis</td>
<td>20</td>
<td>18</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dro</td>
<td>13</td>
<td>12</td>
<td>26</td>
<td>26</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asc</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>30</td>
<td>26</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phy</td>
<td>38</td>
<td>39</td>
<td>39</td>
<td>41</td>
<td>37</td>
<td>37</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>30</td>
<td>34</td>
<td>33</td>
<td>34</td>
<td>30</td>
<td>31</td>
<td>28</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sch</td>
<td>35</td>
<td>35</td>
<td>33</td>
<td>34</td>
<td>30</td>
<td>35</td>
<td>29</td>
<td>38</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tor</td>
<td>34</td>
<td>34</td>
<td>36</td>
<td>33</td>
<td>33</td>
<td>34</td>
<td>31</td>
<td>39</td>
<td>33</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klu</td>
<td>34</td>
<td>35</td>
<td>34</td>
<td>38</td>
<td>33</td>
<td>37</td>
<td>32</td>
<td>36</td>
<td>33</td>
<td>23</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sac</td>
<td>39</td>
<td>39</td>
<td>40</td>
<td>42</td>
<td>38</td>
<td>37</td>
<td>33</td>
<td>42</td>
<td>38</td>
<td>28</td>
<td>15</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>24</td>
<td>34</td>
<td>35</td>
<td>37</td>
<td>36</td>
<td>37</td>
<td>28</td>
<td>43</td>
<td>30</td>
<td>31</td>
<td>31</td>
<td>28</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neu</td>
<td>41</td>
<td>39</td>
<td>42</td>
<td>44</td>
<td>40</td>
<td>41</td>
<td>26</td>
<td>46</td>
<td>36</td>
<td>36</td>
<td>34</td>
<td>34</td>
<td>38</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pod</td>
<td>37</td>
<td>36</td>
<td>36</td>
<td>39</td>
<td>33</td>
<td>37</td>
<td>27</td>
<td>45</td>
<td>32</td>
<td>32</td>
<td>34</td>
<td>33</td>
<td>37</td>
<td>21</td>
<td>24</td>
<td></td>
</tr>
</tbody>
</table>

- **Animal**
- **Protist**
- **Fungus**

Dark to light -> Close to far
Results

Edit distances fitted to an additive tree model using a weighted least-squares criterion generated this tree.
Results

- **Validation**: to show calculated $R(a,b)$ significantly different from random noise
- **Method**: random circular permutations created and tested like the genome to determine noise level
- **Result**: Within animal and fungi group the values of $R$ are non-random; between the groups is random
- Shows that $R$ contains phylogenetic information
Discussion

- **Overall assessment:** coherence of phylogeny indicates that the macrostructures of genomes contain quantitatively meaningful information for phylogenetic reconstruction.

- **Assumptions:**
  - All inferred rearrangement events contribute the same amount to $E$ (solution: weighting).
  - Each deletion is a separate event (solution: slow-growing convex function to determine $D$ if events simultaneous).
  - No back mutation (solution: simulation study determining a correction for this underestimation).