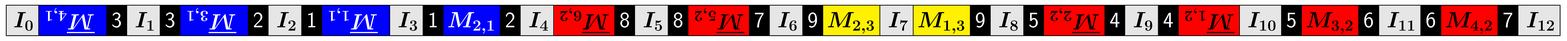
Extending the Assembly Graph Model to include Multiple Scrambled Genes

Jonathan Burns (jtburns@mail.usf.edu)

Top: Masahico Saito, Egor Dolzhenko, Tilahun Muche, Ryan Arredondo

Bottom: Jonathan Burns, Nataša Jonoska, Maja Milošević, Christeen Bisnath



A scrambled micronuclear gene consisting of pointers (black), IESs (gray), and MDSs (blue, red, and yellow).

Micronuclear Arrangement

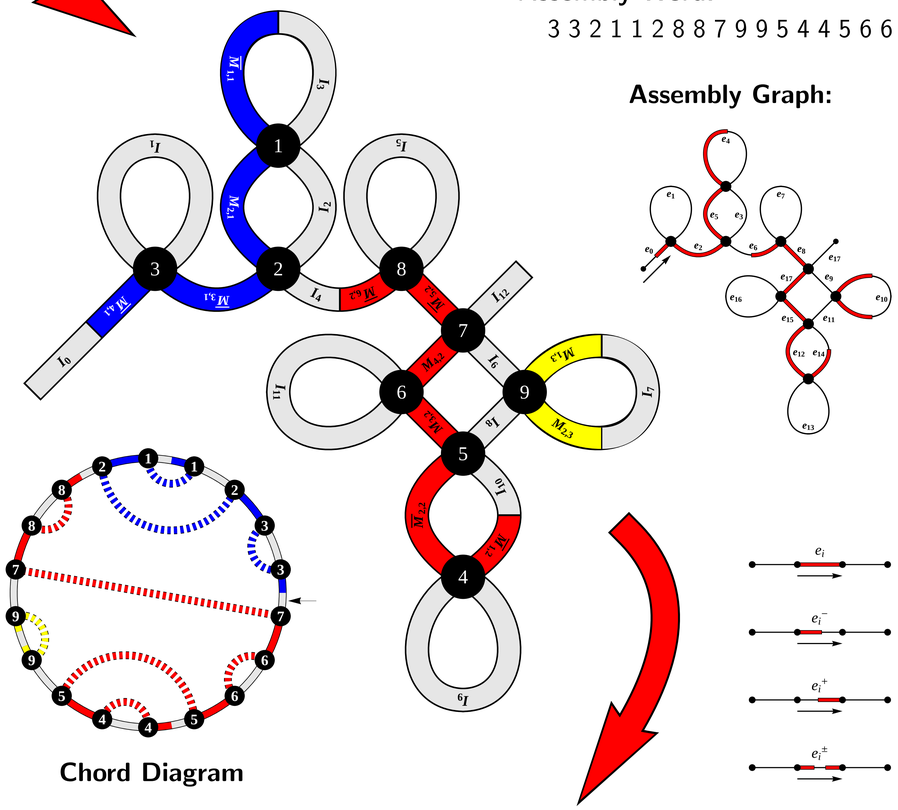
$$\alpha = \overline{M}_{4,1}\,\overline{M}_{3,1}\,\overline{M}_{1,1}\,M_{2,1}\,\overline{M}_{6,2}\,\overline{M}_{5,2}\,M_{2,3}\,M_{1,3}\,\overline{M}_{2,2}\,\overline{M}_{1,2}\,M_{3,2}\,M_{4,2}$$

HaSPoPs: $\Gamma = \{e_0^+, e_2, e_4^-, e_5, e_6^+, e_8, e_{10}^\pm, e_{12}, e_{14}^-, e_{15}, e_{17}\}$

Assembly Word:

3 3 2 1 1 2 8 8 7 9 9 5 4 4 5 6 6 7

Assembly Graph:



Chord Diagram

$M_{1,1}$ 1 $M_{2,1}$ 2 $M_{3,1}$ 3 $M_{4,1}$ | $M_{1,2}$ 4 $M_{2,2}$ 5 $M_{3,2}$ 6 $M_{4,2}$ 7 $M_{5,2}$ 8 $M_{6,2}$ | $M_{1,3}$ 9 $M_{2,3}$

Unscrambled macronuclear gene with MDSs aligned by pointers and IESs excised.

Orthodox Arrangement

$$\text{orth}(\alpha) = M_{1,1}\,M_{2,1}\,M_{3,1}\,M_{4,1} \mid M_{1,2}\,M_{2,2}\,M_{3,2}\,M_{4,2}\,M_{5,2}\,M_{6,2} \mid M_{1,3}\,M_{2,3}$$

Micronuclear Arrangement to Assembly Word

Algorithm:

- **Input**: $\alpha = M_{\sigma_1,\tau_1}^{\epsilon_1} M_{\sigma_2,\tau_2}^{\epsilon_2} \dots M_{\sigma_n,\tau_n}^{\epsilon_n}$
  (d genes with lengths $k_1, k_2, \dots, k_d$)
- **Output**: $w(\alpha) = a_1 a_2 \dots a_{n-d}$ (Assembly Word)

- **Proceedure**: Define ϱ such that
  1. (**Ends**) For each $i \in [d]$,
  $$\varrho(M_{1,i}^1) = \varrho(M_{1,i}^{-1}) = \left(\left[1 + \sum_{j=1}^{i-1}(k_j - 1)\right]\right) \text{ and}$$
  $$\varrho(M_{k_i,i}^1) = \varrho(M_{k_i,i}^{-1}) = \left(\left[\sum_{j=1}^{i}(k_j - 1)\right]\right).$$
  2. (**Non-ends**) For each $i \in [d]$ and $1 < p < k_i$,
  $$\varrho(M_{p,i}^1) = \left(\left[(p-1) + \sum_{j=1}^{i-1}(k_j - 1)\right]\left[p + \sum_{j=1}^{i-1}(k_j - 1)\right]\right)$$
  $$\varrho(M_{p,i}^{-1}) = \left(\left[p + \sum_{j=1}^{i-1}(k_j - 1)\right]\left[(p-1) + \sum_{j=1}^{i-1}(k_j - 1)\right]\right).$$

Example:

$$\alpha = M_{2,1}\overline{M}_{1,1}M_{3,1}M_{2,2}\overline{M}_{3,2}M_{4,2}M_{1,2}M_{4,1}\overline{M}_{6,1}M_{5,1}$$
$$d = 2: \quad k_1 = 6,\ k_2 = 4$$

$$\varrho(\alpha) = ([1][2])\,([1])\,([2][3])\,([6][7])\,([8][7])\,([8])\,([6])\,([3][4])\,([5])\,([4][5])$$

$$w(\alpha) = 1\,2\,1\,2\,3\,6\,7\,8\,7\,8\,6\,3\,4\,5\,4\,5$$

Micronuclear Arrangement to HaSPoPs

Algorithm:

- **Input**: $\alpha = M_{\sigma_1,\tau_1}^{\epsilon_1} M_{\sigma_2,\tau_2}^{\epsilon_2} \dots M_{\sigma_n,\tau_n}^{\epsilon_n}$
  (genes with lengths $k_1, k_2, \dots, k_d$)
- **Output**: $\Gamma = \{\gamma_1, \gamma_2, \dots, \gamma_d\}$ (Hamiltonian set of polygonal paths)

- **Proceedure**:
  1. **Set** all
     - $\overline{M}_{1,j}, M_{k_j,j} \to A$ (End)
     - $M_{1,j}, \overline{M}_{k_j,j} \to B$ (End)
     - $M_{\sigma_i,\tau_i}^{\epsilon_i} \to M$ (Non-ends)
  2. **Change** any $AB \to C$
  3. **Insert** Is before, after, and between As, Bs, Cs, and Ms
  4. **Remove** Is after Bs and before As
  5. **Index** letters sequentially from 0.
  6. **Remove** all Is.
  7. **Map** Letters to Edges:
     - $M_i \to e_i$
     - $B_i \to e_i^-$
     - $A_i \to e_i^+$
     - $C_i \to e_i^\pm$

Example:

$$\alpha = M_{2,1}\overline{M}_{1,1}M_{3,1}M_{2,2}\overline{M}_{3,2}M_{4,2}M_{1,2}M_{4,1}\overline{M}_{6,1}M_{5,1}$$

1. $M\,A\,M\,M\,M\,A\,B\,M\,B\,M$
3. $I\,M\,I\,A\,I\,M\,I\,M\,I\,M\,I\,C\,I\,M\,I\,B\,I\,M\,I$
5. $I_0\,M_1\,A_2\,I_3\,M_4\,I_5\,M_6\,I_7\,M_8\,I_9\,C_{10}\,I_{11}\,M_{12}\,I_{13}\,B_{14}\,M_{15}\,I_{16}$
7. $\Gamma = \{e_1, e_2^+, e_4, e_6, e_8, e_{10}^\pm, e_{12}, e_{14}^-, e_{15}\}$

Biological Motivation

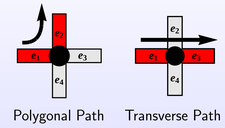
- Several species of ciliates, such as *Oxytricha* and *Stylonychia*, can undergo massive genome rearrangement while mating. See [Angeleska, 2007] and [Ehrenfeucht, 2005] for further details.
- Ciliates have both Macronuclei (Somatic) and Micronuclei (Germline).
  - During mating, micronuclear genes recombine to form macronuclear genes.
- Micronuclear genes contain
  - Coding Segements - **Macronuclear Destined Sequences** (MDSs)
  - Non-Coding Segments - **Internal Eliminated Sequences** (IESs)

  The MDSs are expressed in the final unscrambled gene while the IESs are excised during the recombination process.
- **Pointers** are short segments which flank the MDSs, guiding the recombination process.
- In relation to an unscrambled micronuclear gene, a scrambled macronuclear gene may have permuted or inverted MDS segments separated by IESs.
- Formation of macronuclear genes may require any combination of the following three events:
  - Unscrambling of segment order
  - DNA inversion
  - IES removal

Image: Cilliate undergoing the last processes of binary fission [http://en.wikipedia.org/wiki/File:Unk.cilliate.jpg]

Assembly Graph Model

- A 4-valent **rigid vertex** has is a vertex with degree 4 and cyclically ordered adjacent edges.
- An **assembly graph** is a finite connected graph where all vertices are rigid vertices of valency 1 or 4.
- Given a 4-valent rigid vertex with cyclic order $\{e_1, e_2, e_3, e_4\}$, the pairs
  - e_1, e_2, e_2, e_3 e_3, e_4 e_1, e_4 are **neighbors**
  - e_1, e_3, e_2, e_4 are **non-neighbors**.
- Two types of paths:
  - **Polygonal** - Consecutive path edges are neighbors. (Make $90°$ turns)
  - **Transverse** - Consecutive path edges are non-neighbors. (Continue on $180°$)
- A transverse path (with distinct edges) that *spans* an entire assembly graph is called a **transversal**.
- A set of polygonal paths (with distinct edges) that *cover* all the vertices of an assembly graph is called a **Hamiltonian Set of Polygonal Paths** (**HaSPoPs**).
- With this convention:
  - Scrambled micronuclear sequences correspond to the transversal of an assembly graph G.
  - Unscrambled macronuclear sequences correspond to a HaSPoPs $\Gamma \subset G$.



Polygonal Path Transverse Path

Micronuclear Arrangement Notation

- Each **micronuclear arrangement** α may be expressed in the form
  $$\alpha(\sigma, \tau, \epsilon) = M_{\sigma_1,\tau_1}^{\epsilon_1} M_{\sigma_2,\tau_2}^{\epsilon_2} \dots M_{\sigma_k,\tau_k}^{\epsilon_k}.$$
  Relative to the **orthodox arrangement** $\text{orth}(\alpha)$,
  - σ_i: the order of the MDS within gene,
  - τ_i: the gene the MDS belongs to,
  - ϵ_i: the orientation of the MDS.
- We write
  - M_{σ_i,τ_i} when $\epsilon_i = 1$ (unscrambled orientation)
  - $\overline{M}_{\sigma_i,\tau_i}$ when $\epsilon_i = -1$ (inverted orientation).

Fun Facts

1. There are
   $$\frac{1}{2}\left[(2n-1)!! + \sum_{k=0}^{\lfloor n/2 \rfloor} \frac{n!}{(n-2k)!k!}\right]$$
   distinct assembly graphs with n rigid vertices.

---

2. There are exactly
   $$2^n \frac{n!}{k!}\binom{n-1}{k-1}$$
   distinct micronuclear arrangements with n MDSs among k genes.

References

- A. Angeleska, N. Jonoska, M. Saito, L.F. Landweber, RNA-guided DNA assembly, *J. of Theo. Bio.* **248** (4) (2007) 706-720.
- J. Burns, T. Muche, Counting Irreducible Double Occurrence Words, *Cong. Num.* Accepted (2011)
- J. Burns, E. Dolzhenko, N. Jonoska, T. Muche, M. Saito, Four-Regular Graphs with Rigid Vertices Associated to DNA Recombination. Preprint (2011)
- A. Ehrenfeucht, T. Harju, I. Petre, D.M. Prescott, G. Rozenberg, *Computing in Living Cells*, Springer (2005)

Supported by: NSF Grant DUE-0756847

http://jtburns.myweb.usf.edu/assembly