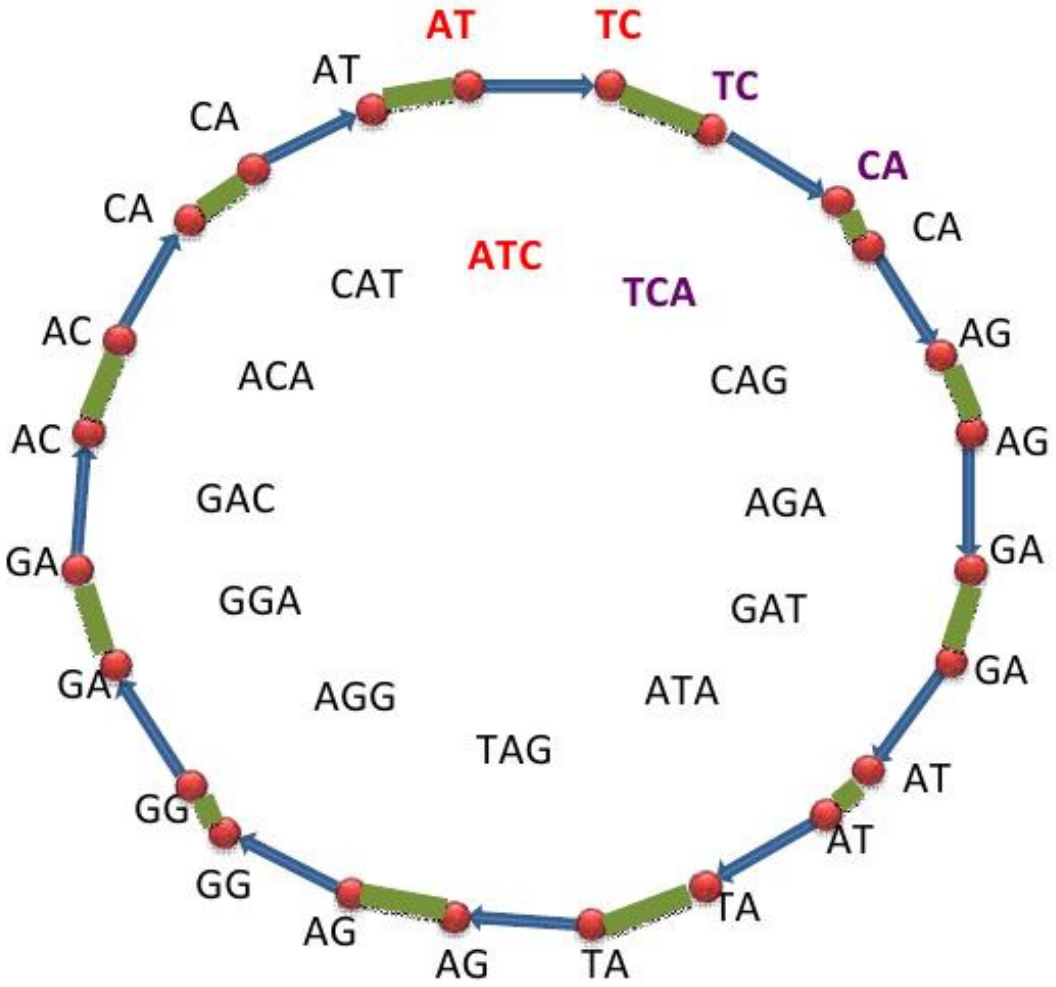
Multiple libraries pathsets

Student: Nadiya Sitdykova
Academic Adviser: Son Pham

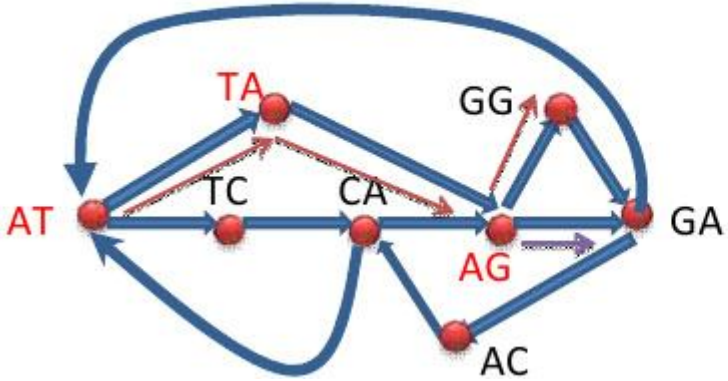
Problem statement

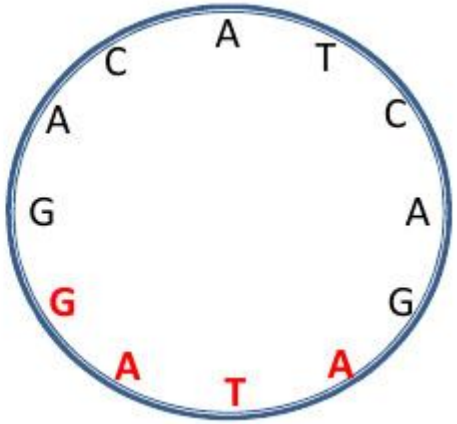


- Genome is circular string.
- Reads cover all k-mers in the genome.
- Two libraries of read-pairs with different insert sizes d_1 and d_2.

Problem statement

- Find the Eulerian cycle in the de Bruiijn graph, corresponded to Genome from fundamental way (without heuristics).

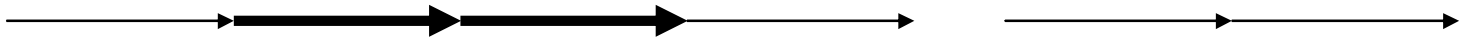




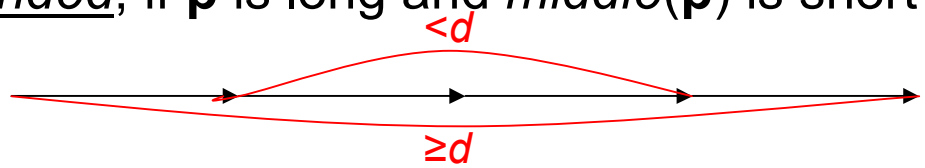
Pathset assembly

Definitions

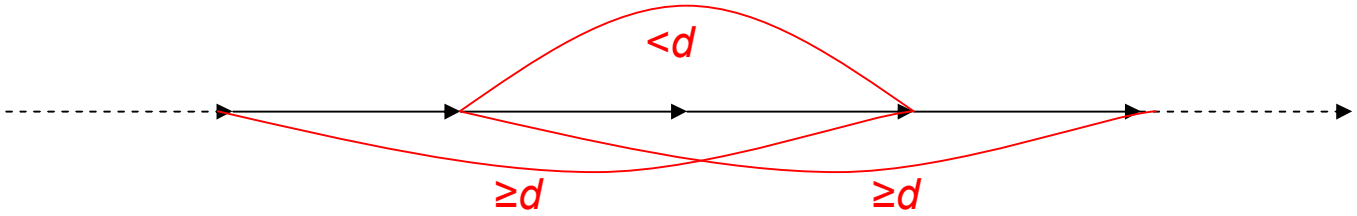
- Path **p** is *long*, if length(**p**)≥*d*
- Path **p** is *short*, if length(**p**)<*d*
- *middle*(**p**)



- Path **p** is *d-bounded*, if **p** is long and *middle*(**p**) is short



- Path **p** is *d-maxpath*, if **p** is maximal *d-bounded*



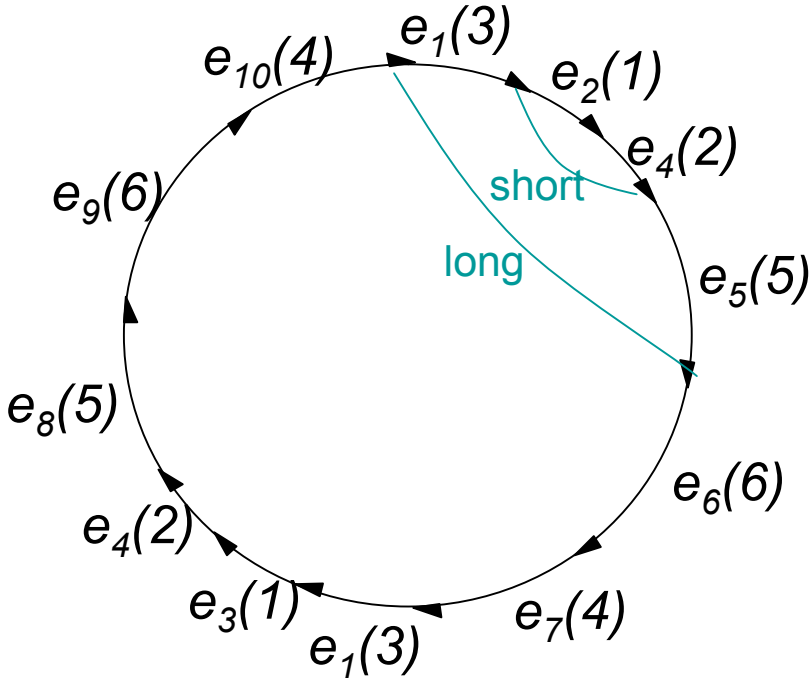
- First and last edges of *d-bounded* paths produce *edge-pairs* in the condensed de Bruijn graph.
- *Biedge(a|b,d)* defines *Pathset(a|b,d)* as all paths between *a* and *b* of length *d* in the condensed de Bruijn graph.

Pathset assembly

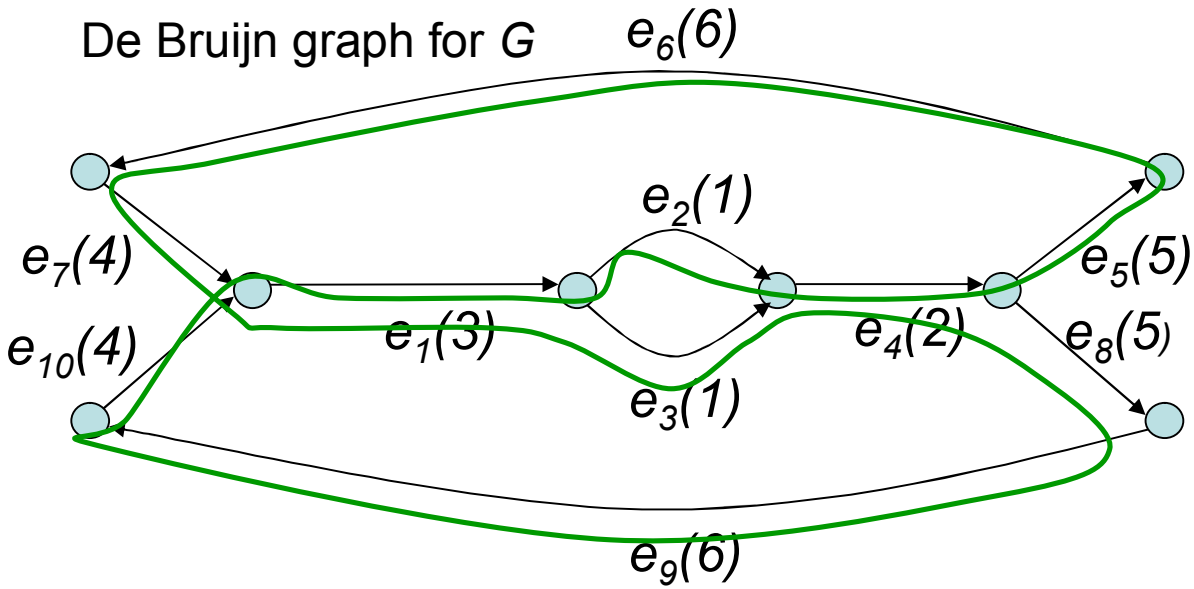
7-bounded paths

$e_1e_2e_4e_5$	$e_1e_3e_4e_8$
$e_2e_4e_5$	$e_3e_4e_8$
$e_4e_5e_6$	$e_4e_8e_9$
$e_5e_6e_7$	$e_8e_9e_{10}$
$e_6e_7e_1$	$e_9e_{10}e_1$
$e_7e_1e_3e_4e_8$	$e_{10}e_1e_2e_4e_5$

Genome *G*



De Bruijn graph for *G*

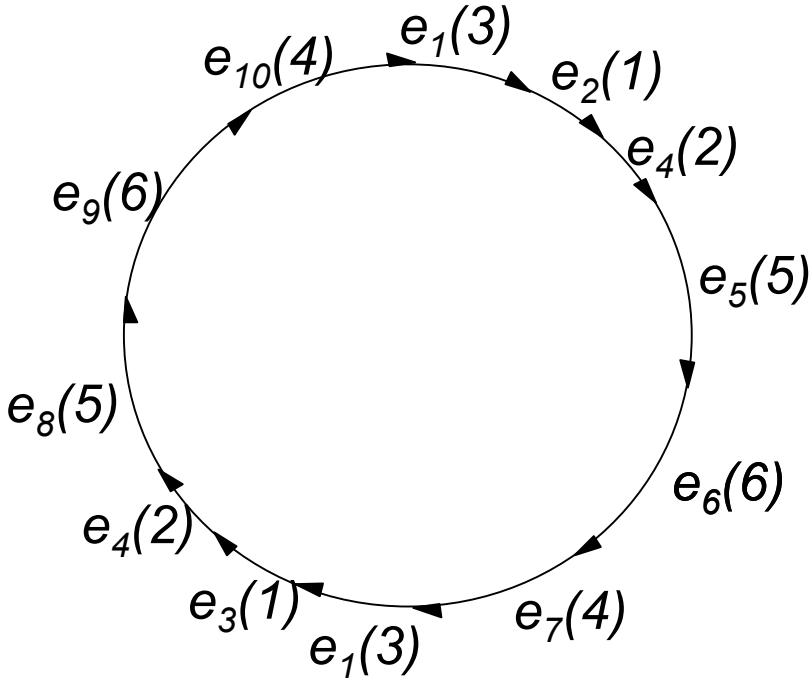


Pathset assembly

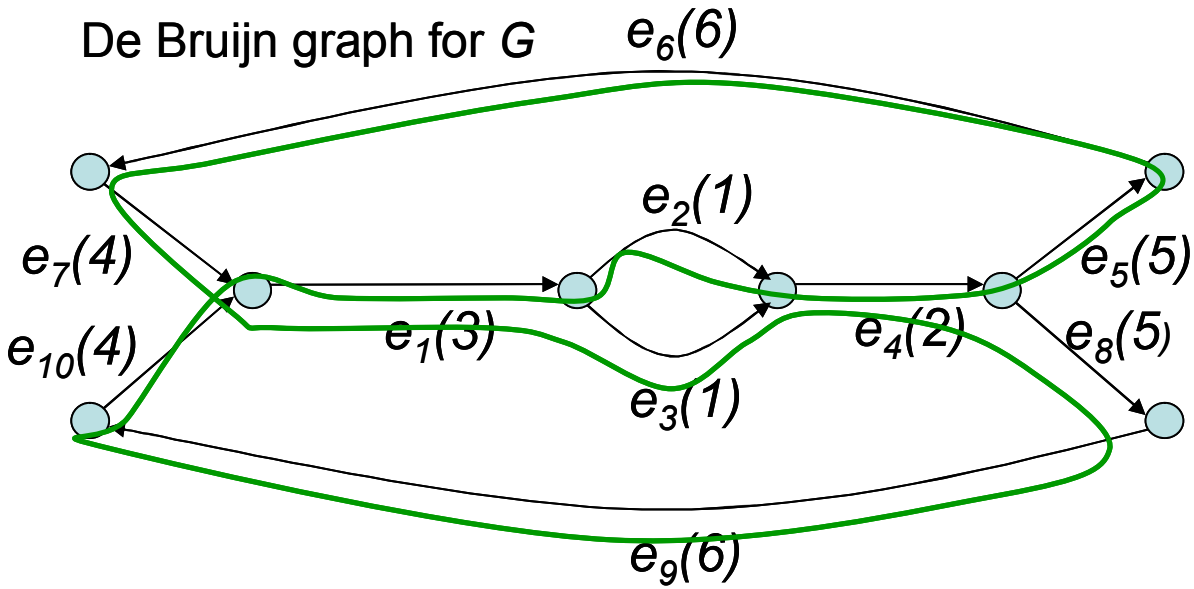
7-maxpaths

$e_4e_5e_6$	$e_4e_8e_9$
$e_5e_6e_7$	$e_8e_9e_{10}$
$e_6e_7e_1$	$e_9e_{10}e_1$
$e_7e_1e_3e_4e_8$	$e_{10}e_1e_2e_4e_5$

Genome *G*



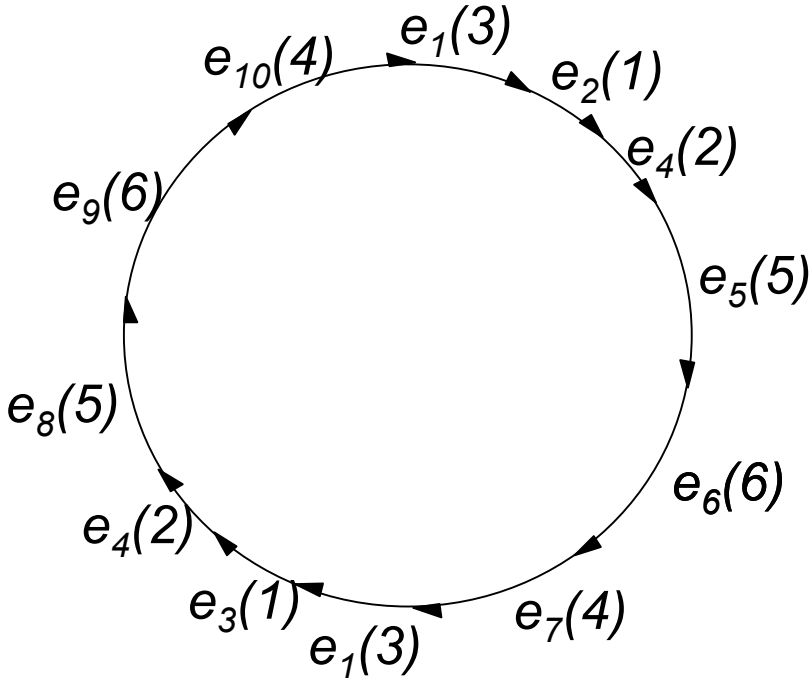
De Bruijn graph for *G*



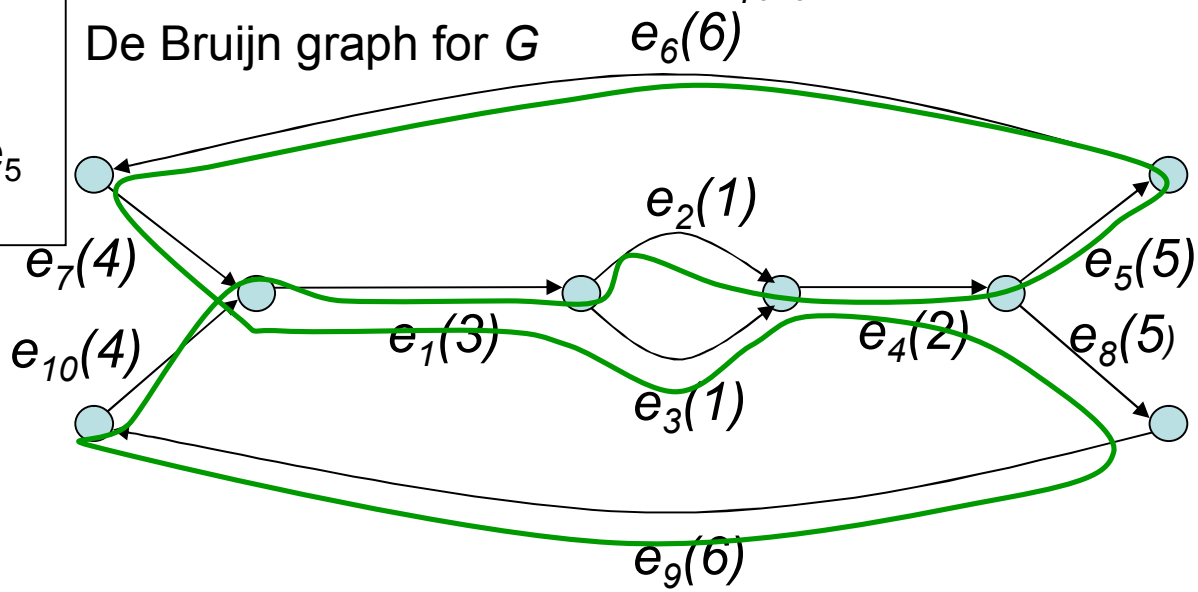
Pathset assembly

Biedges(a|b,d)	*7-maxpaths*
$(e_4\|e_6,7)$	$e_4e_5e_6$
$(e_5\|e_7,7)$	$e_5e_6e_7$
$(e_6\|e_1,7)$	$e_6e_7e_1$
$(e_7\|e_8,7)$	$e_7e_1e_3e_4e_8$
$(e_4\|e_9,7)$	$e_4e_8e_9$
$(e_8\|e_{10},7)$	$e_8e_9e_{10}$
$(e_9\|e_1,7)$	$e_9e_{10}e_1$
$(e_{10}\|e_5,7)$	$e_{10}e_1e_2e_4e_5$

Genome *G*



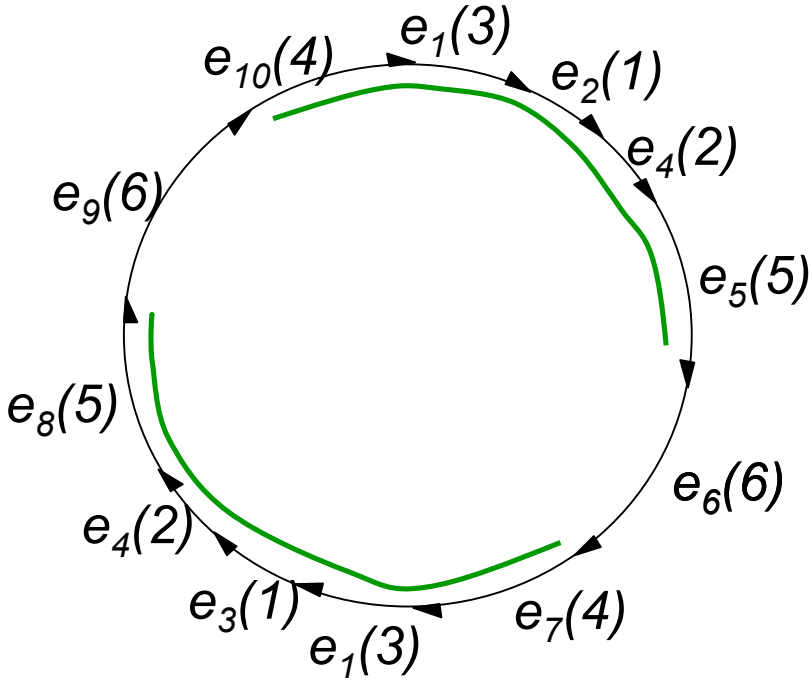
De Bruijn graph for *G*



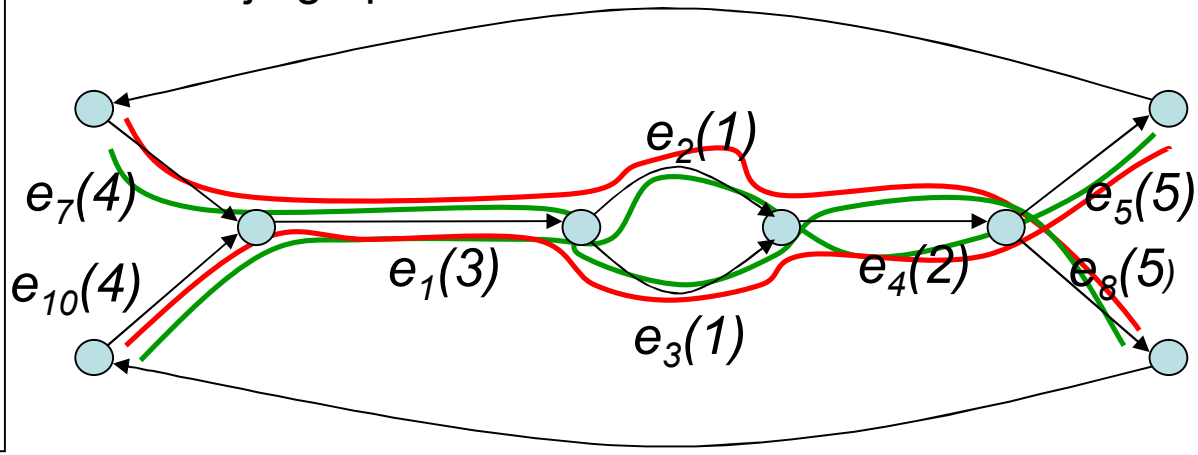
Pathset assembly

Biedges(a|b,d)	*Pathsets*
$(e_4\|e_6,7)$	$\{e_4e_5e_6\}$
$(e_5\|e_7,7)$	$\{e_5e_6e_7\}$
$(e_6\|e_1,7)$	$\{e_6e_7e_1\}$
$(e_7\|e_8,7)$	$\{e_7e_1e_3e_4e_8;$
	$e_7e_1e_2e_4e_8\}$
$(e_4\|e_9,7)$	$\{e_4e_8e_9\}$
$(e_8\|e_{10},7)$	$\{e_8e_9e_{10}\}$
$(e_9\|e_1,7)$	$\{e_9e_{10}e_1\}$
$(e_{10}\|e_5,7)$	$\{e_{10}e_1e_2e_4e_5;$
	$e_{10}e_1e_3e_4e_5\}$

Genome *G*



De Bruijn graph for *G*



Pathset assembly

Pathsets

{$e_4e_5e_6$}

{$e_5e_6e_7$}

{$e_6e_7e_1$}

{$e_7e_1e_3e_4e_8$;

$e_7e_1e_2e_4e_8$}

{$e_4e_8e_9$}

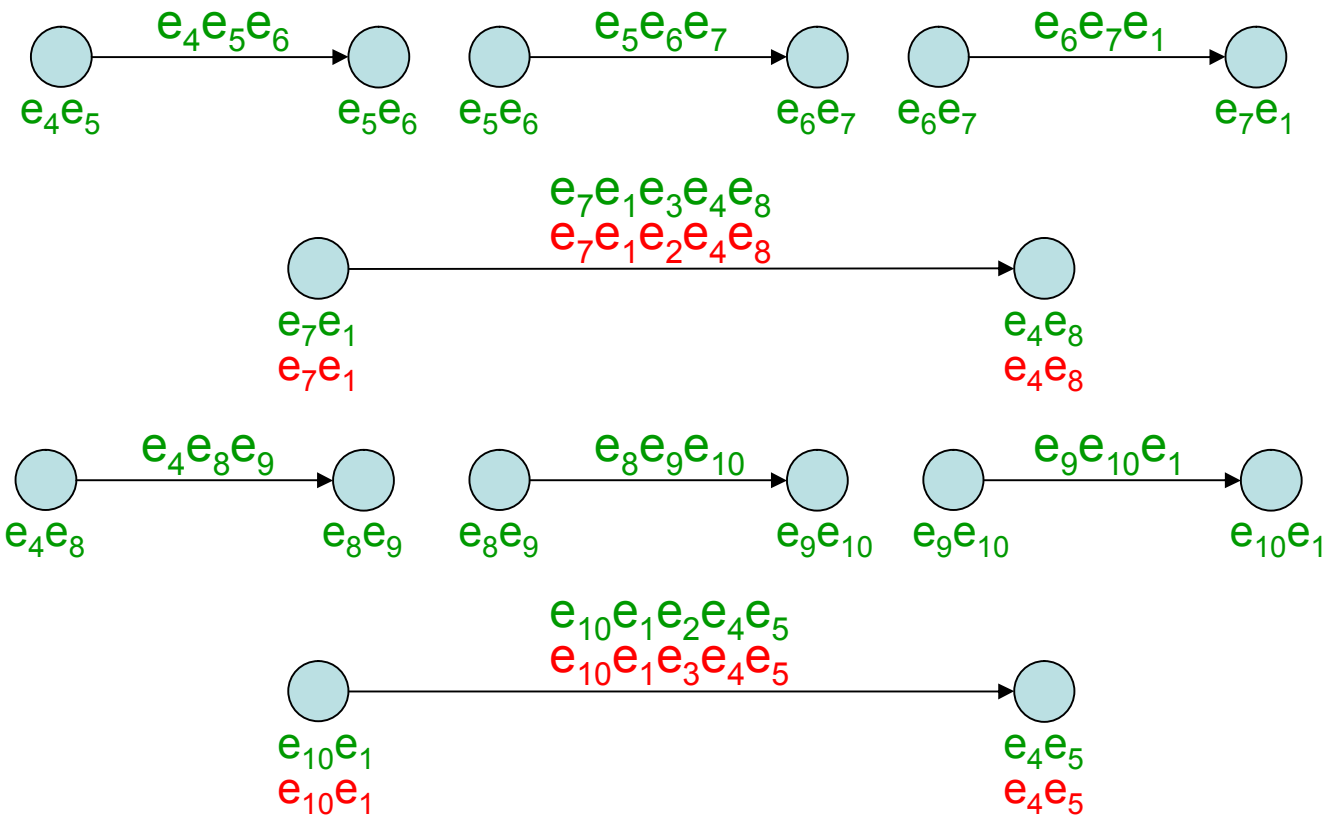
{$e_8e_9e_{10}$}

{$e_9e_{10}e_1$}

{$e_{10}e_1e_2e_4e_5$;

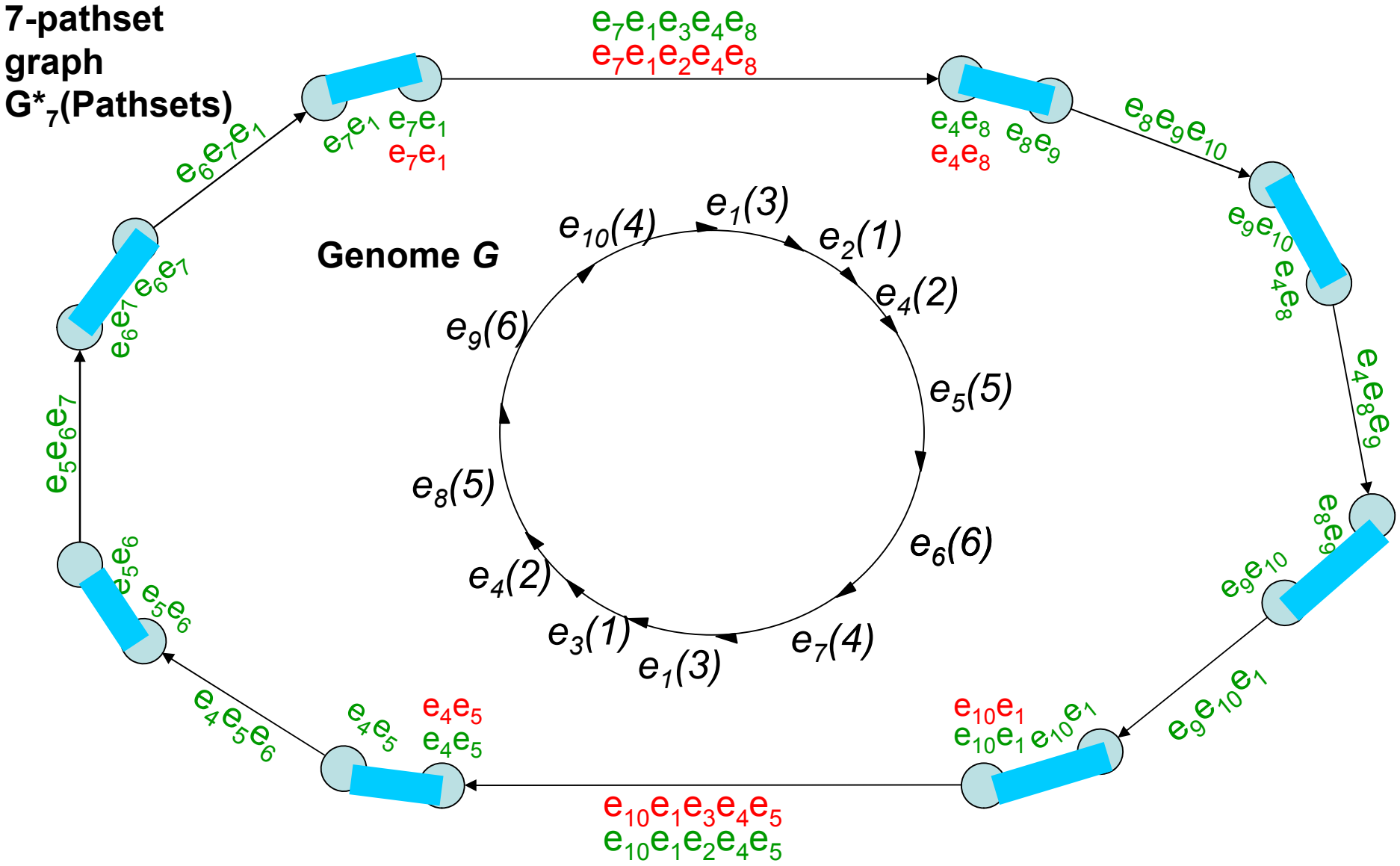
$e_{10}e_1e_3e_4e_5$}

1. For each pathset **P** introduce new vertices *u*, *v* and form edge $u \rightarrow v$.
   - Label edge by **P**.
   - Label *u* by set of all minimal long prefixes of paths in **P**.
   - Label *v* by set of all minimal long suffixes of paths in **P**.



Pathset assembly

2. Glue vertices together when they share a label.

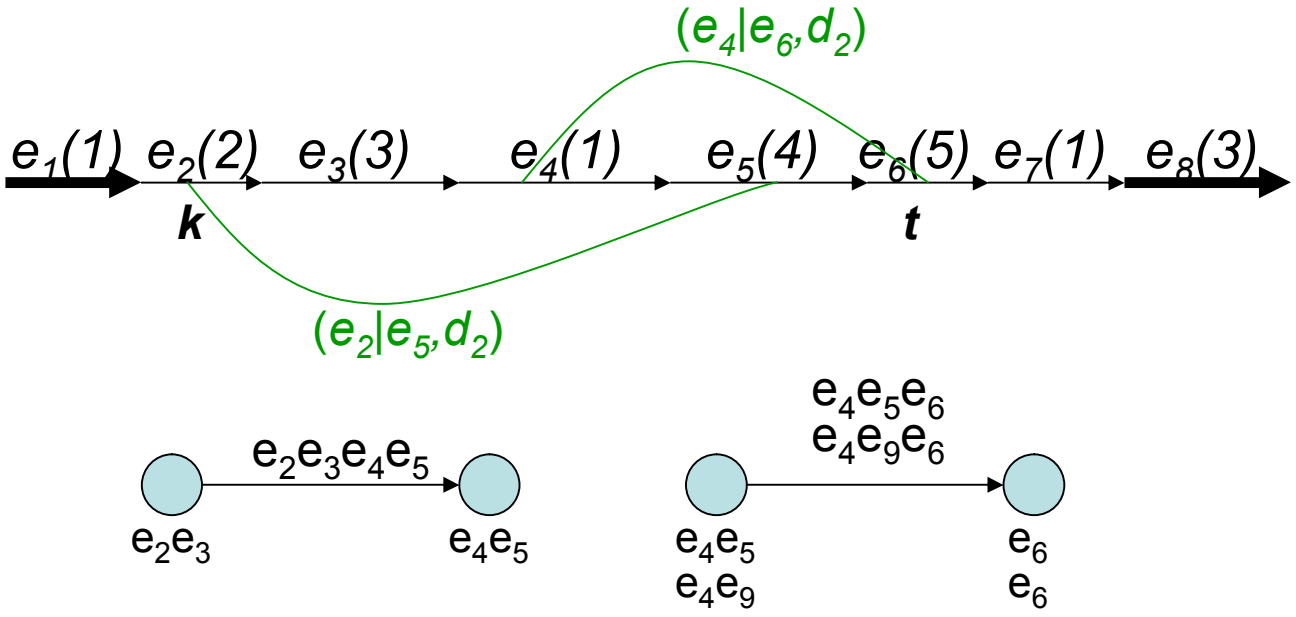


Multiple library pathsets

- Given two libraries of biedges HB_1 and HB_2 ($d_1 > d_2$)
- Constructed $Pathset(HB_1)$ for all $(a|b,d_1) \in HB_1$ and $Pathset(HB_2)$ for all $(a|b,d_2) \in HB_2$
- For each path **p** = *e1,…,en* from *Pathset(a|b, d1)* define set of biedges $HB_{corr}(\mathbf{p})$ and $Pathset(HB_{corr}(\mathbf{p}))$*:*
  1. Initiate $HB_{corr}(\mathbf{p})$ as set of all existing $(e_i|e_j,d_2) \in HB_2$, $i < j$

     Find $\boldsymbol{k} = min\{\, i : (e_i|e_j,d_2) \in HB_{corr}(\mathbf{p}) \,\}$

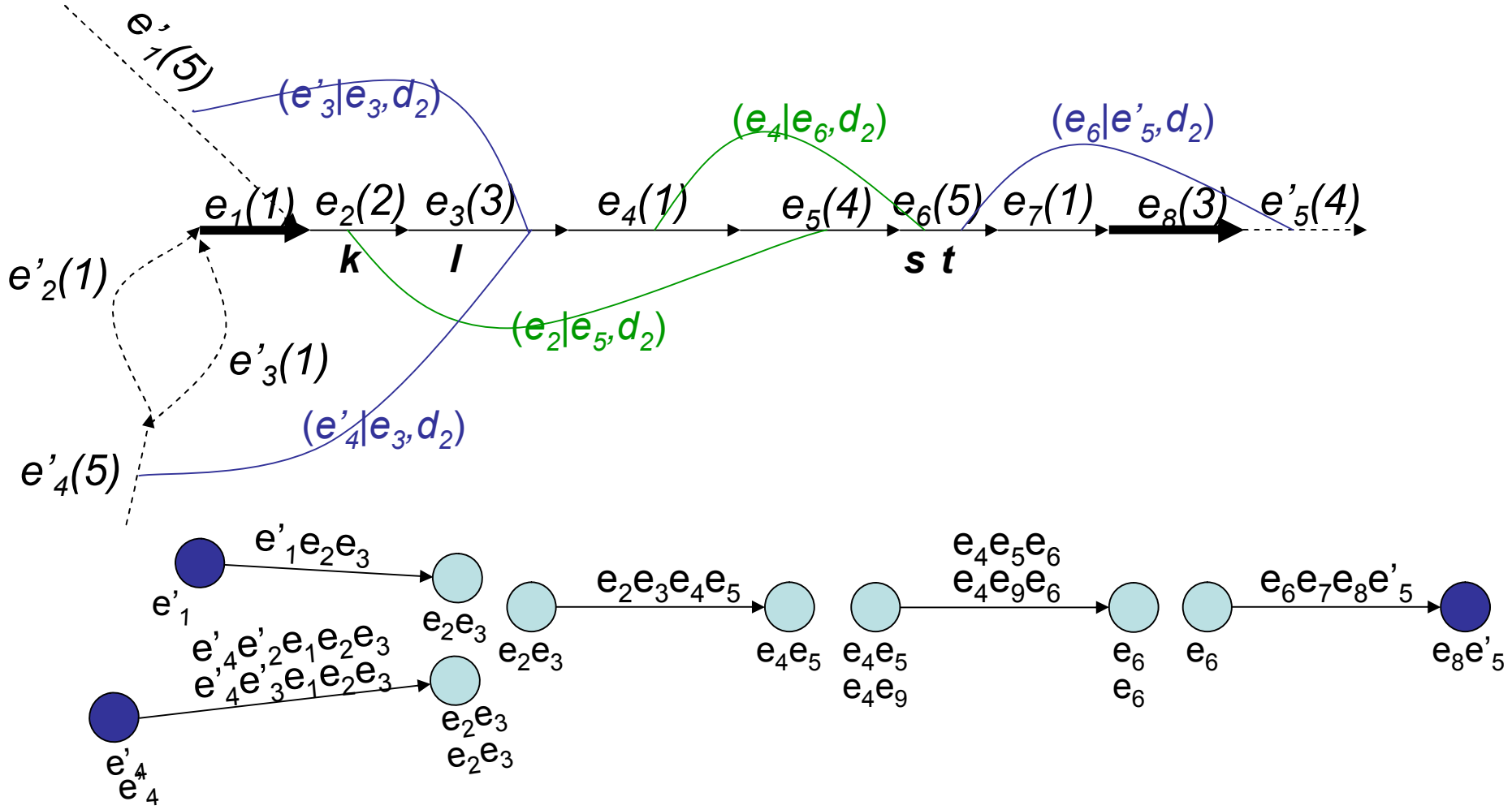
     Find $\boldsymbol{t} = max\{\, j : (e_i|e_j,d_2) \in HB_{corr}(\mathbf{p}) \,\}$



Multiple library pathsets

2. If $e_k \neq e_1$ then add to $HB_{corr}(\mathbf{p})$ all existing $(e'|e_l, d_2) \in HB_2$, where e_l is last edge of minimal long prefix in $e_k \ldots e_n$

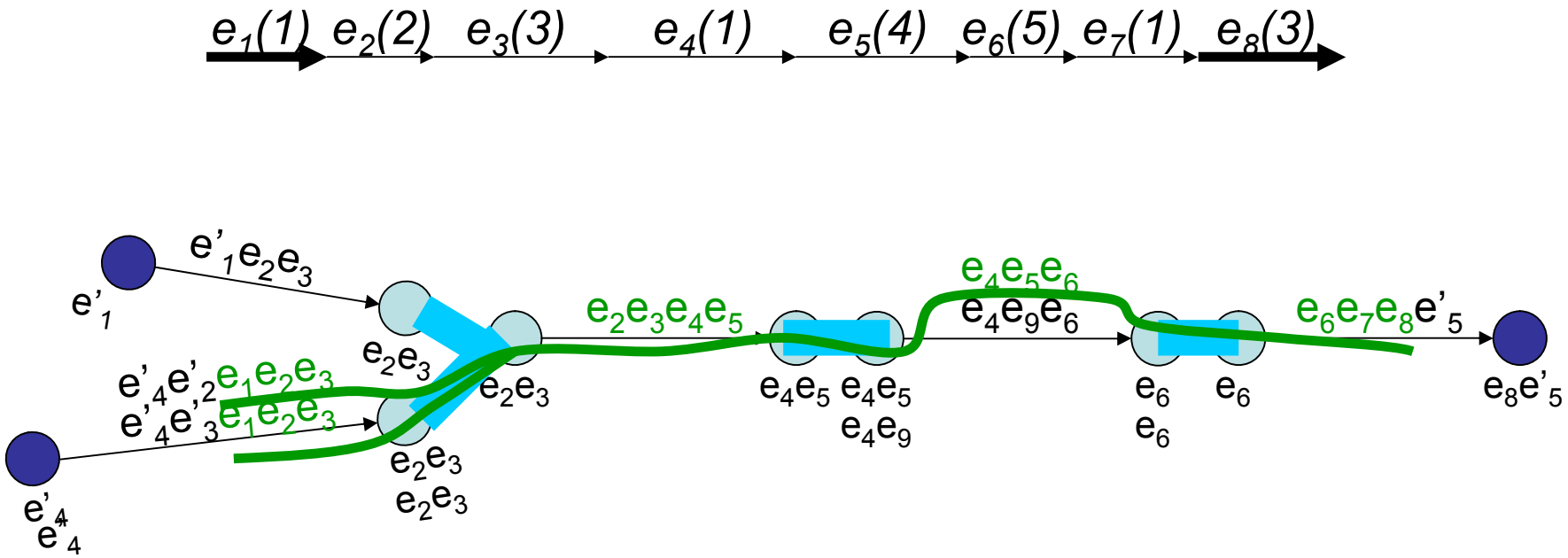
   If $e_t \neq e_n$ then add to $HB_{corr}(\mathbf{p})$ all existing $(e_s|e', d_2) \in HB_2$, where e_s is first edge of minimal long suffix in $e_1 \ldots e_t$



Multiple library pathsets

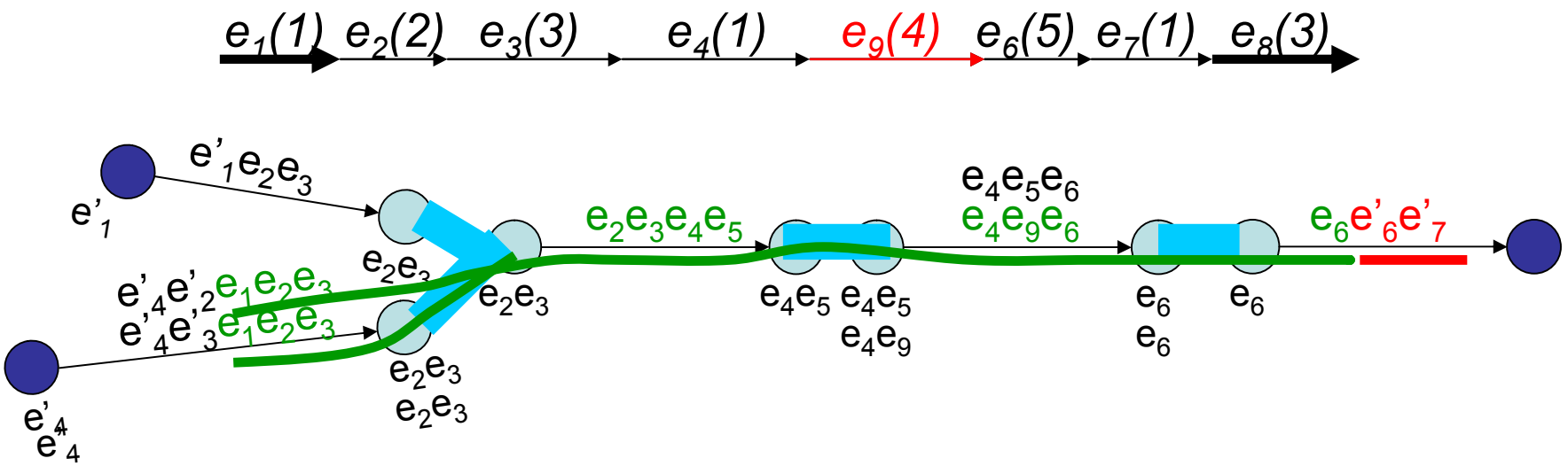
- Construct pathset graph $G^*_{d2}(HB_{corr}(\mathbf{p}))$

  If path **p** is genomic, it will be represented in $G^*_{d2}(HB_{corr})$



Multiple library pathsets

So, If path **p** doesn't exist in $G^*_{d2}(HB_{corr}(\mathbf{p}))$, it's <u>not genomic</u> path



- Remove corrupting paths from each *Pathset(a|b,d_1).*

  Removal of such paths reduce some pathsets from *Pathset(HB_1),* it makes pathset graph *G^*_{d1}(Pathset(HB_1))* less ambiguous
- Construct *G^*_{d1}(Pathset(HB_1))* and find Eulerian cycle.