Алгоритм разрешения повторов по графу де Брюйна

Пржибельский Андрей, гр. 604 (SE)

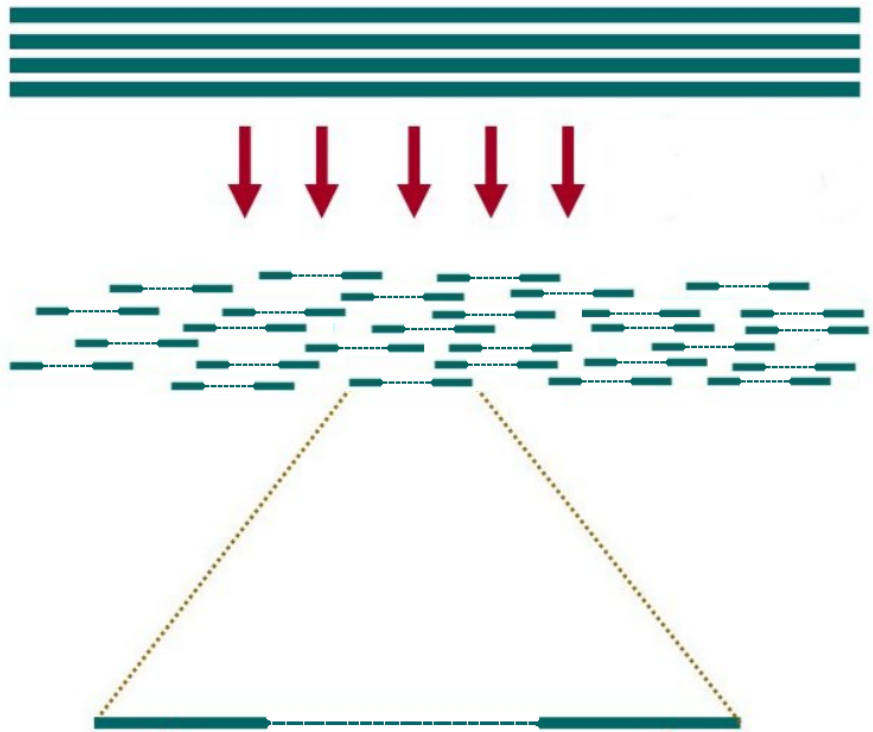
Кафедра математических и информационных технологий

Научные руководители:
Банкевич А.В.
Нурк С.Ю.

Рецензент:
Алексеев М.А.

Геном и секвенирование

- Для биологов геном — это длинные молекулы ДНК
- Для нас — длинные строки над алфавитом из 4 букв (A,T,G,C)
- Секвенирование — процесс "чтения" генома
- Результат — короткие парные фрагменты ($100 - 200$ символов)



Граф де Брюина

- По полученным фрагментам строится граф де Брюйна $\mathbf{G}$
- Геном соответствует некоторому пути $\mathbf{\Gamma}$ в графе
- За счет парности исходных фрагементов между ребрами можно задать связи (парная информация)
- Вес парной информации $W(e_1, e_2, d)$ определяет степень уверенности в том, что ребра e_1 и e_2 находятся в пути генома $\mathbf{\Gamma}$ на расстоянии d

Сложности сборки

- Многие ребра повторяются в пути генома
- Граф де Брюйна является очень запутанным
- Из-за неравномерности исходных данных путь может Γ иметь разрывы

Постановка задачи

Дано:

- Граф де Брюйна $\mathbf{G}$
- Парная информация $W(e_1, e_2, d), e_1, e_2 \in E(\mathbf{G}), d \in \mathbb{Z}$

Найти:

- Множество путей P, таких что:
  - $|\{p \in P | p \notin \mathbf{\Gamma}\}| \to min$
  - $LEN\{p \in P | p \in \mathbf{\Gamma}\} \to max$

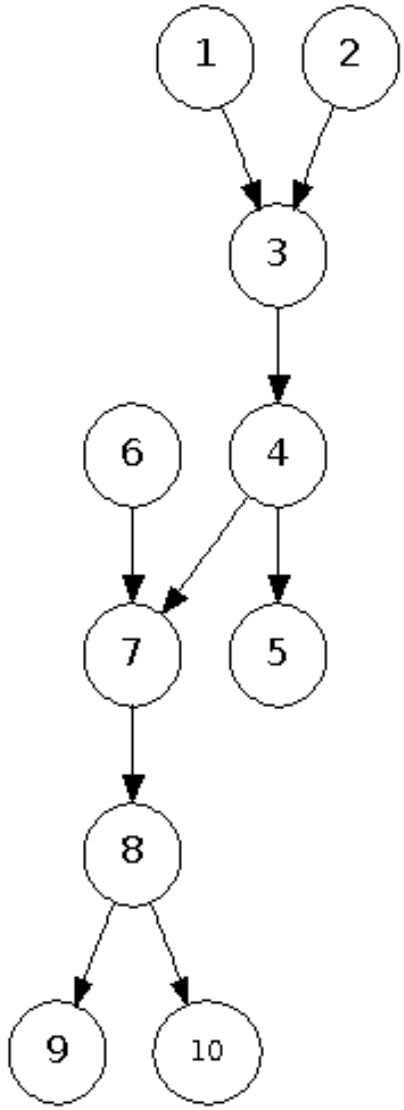
Предложенная стратегия

- Выбирем в графе $\mathbf{G}$ достоверный путь p
- Пусть этот путь имеет несколько возможных продолжений $h_1, \ldots, h_n$
- Добавим в p то, которое лучше остальных согласуется парной информацией с уже пройденным путем
- Продолжаем рост путей итеративно, пока это возможно

Выбор начальных путей

TRIVIAL_PATHS(G)

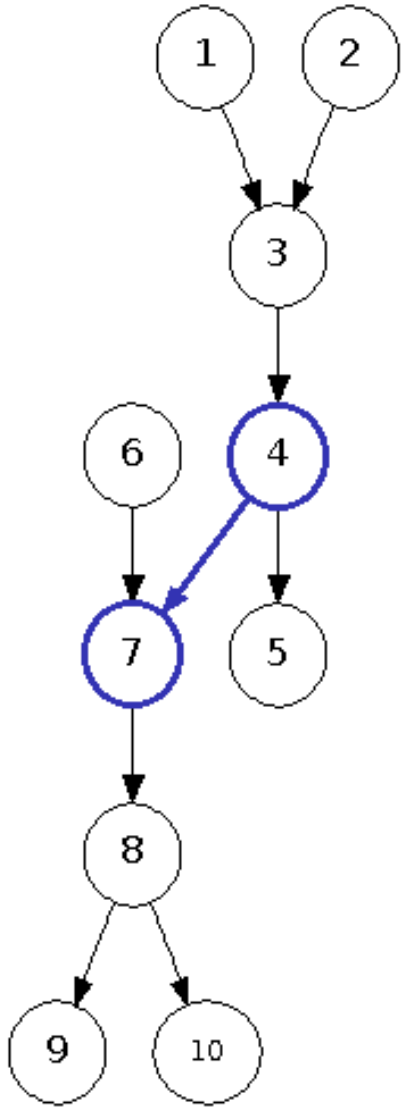
1. $P \leftarrow \{\}$
2. **for each** $e \in E(\mathbf{G})$:
3. $\quad p \leftarrow (e)$
4. $\quad (u, v) \leftarrow e$
5. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (v, w) \in E(\mathbf{G})$:
6. $\quad\quad p \leftarrow (p, (v, w))$
7. $\quad\quad (u, v) \leftarrow (v, w)$
8. $\quad (u, v) \leftarrow e$
9. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (w, u) \in E(\mathbf{G})$:
10. $\quad\quad p \leftarrow ((w, u), p)$
11. $\quad\quad (u, v) \leftarrow (w, u)$
12. $\quad P \leftarrow P \cup \{p\}$
13. **return** P



Выбор начальных путей

TRIVIAL_PATHS(G)

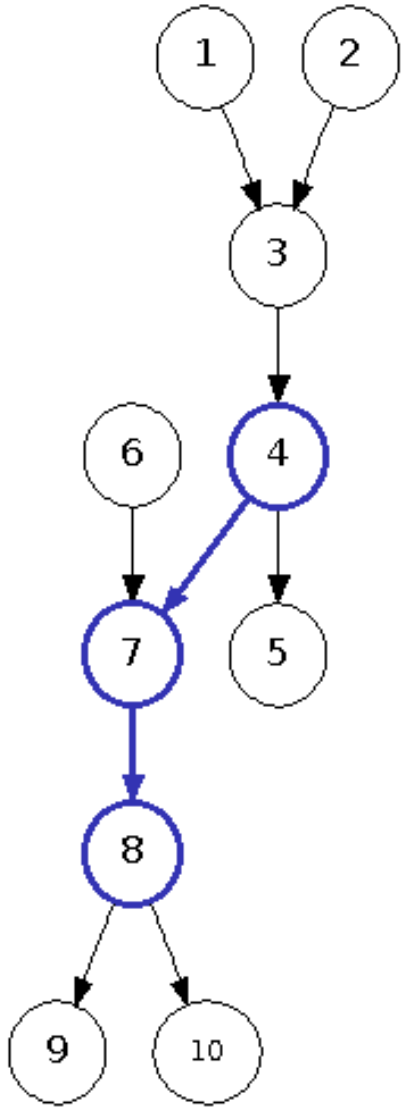
1. $P \leftarrow \{\}$
2. **for each** $e \in E(\mathbf{G})$:
3. $\quad p \leftarrow (e)$
4. $\quad (u, v) \leftarrow e$
5. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (v, w) \in E(\mathbf{G})$:
6. $\quad\quad p \leftarrow (p, (v, w))$
7. $\quad\quad (u, v) \leftarrow (v, w)$
8. $\quad (u, v) \leftarrow e$
9. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (w, u) \in E(\mathbf{G})$:
10. $\quad\quad p \leftarrow ((w, u), p)$
11. $\quad\quad (u, v) \leftarrow (w, u)$
12. $\quad P \leftarrow P \cup \{p\}$
13. **return** P



Выбор начальных путей

TRIVIAL_PATHS(G)

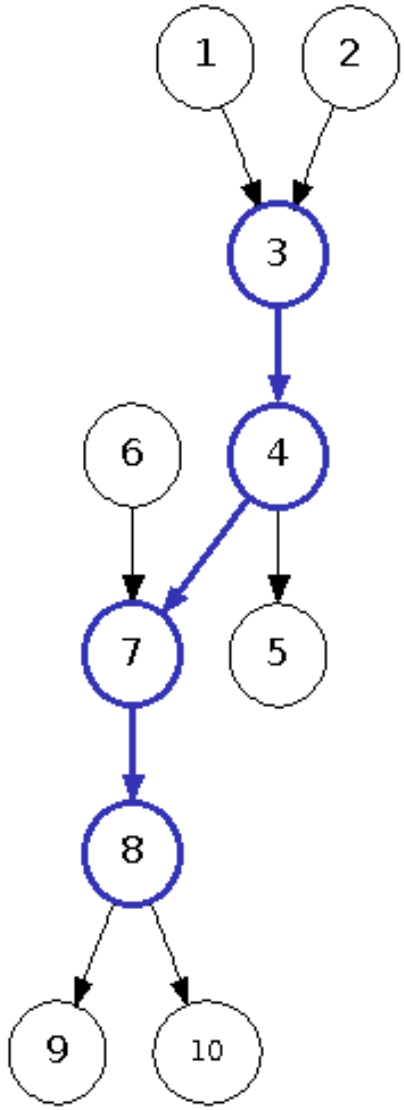
1. $P \leftarrow \{\}$
2. **for each** $e \in E(\mathbf{G})$:
3. $\quad p \leftarrow (e)$
4. $\quad (u, v) \leftarrow e$
5. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (v, w) \in E(\mathbf{G})$:
6. $\quad\quad p \leftarrow (p, (v, w))$
7. $\quad\quad (u, v) \leftarrow (v, w)$
8. $\quad (u, v) \leftarrow e$
9. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (w, u) \in E(\mathbf{G})$:
10. $\quad\quad p \leftarrow ((w, u), p)$
11. $\quad\quad (u, v) \leftarrow (w, u)$
12. $\quad P \leftarrow P \cup \{p\}$
13. **return** P



Выбор начальных путей

TRIVIAL_PATHS(G)

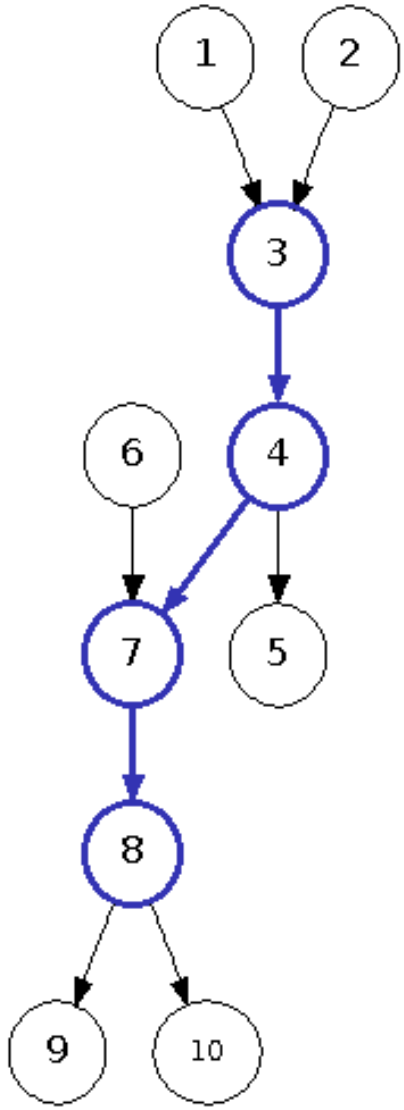
1. $P \leftarrow \{\}$
2. **for each** $e \in E(\mathbf{G})$:
3. $\quad p \leftarrow (e)$
4. $\quad (u, v) \leftarrow e$
5. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (v, w) \in E(\mathbf{G})$:
6. $\quad\quad p \leftarrow (p, (v, w))$
7. $\quad\quad (u, v) \leftarrow (v, w)$
8. $\quad (u, v) \leftarrow e$
9. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (w, u) \in E(\mathbf{G})$:
10. $\quad\quad p \leftarrow ((w, u), p)$
11. $\quad\quad (u, v) \leftarrow (w, u)$
12. $\quad P \leftarrow P \cup \{p\}$
13. **return** P



Выбор начальных путей

TRIVIAL_PATHS(G)

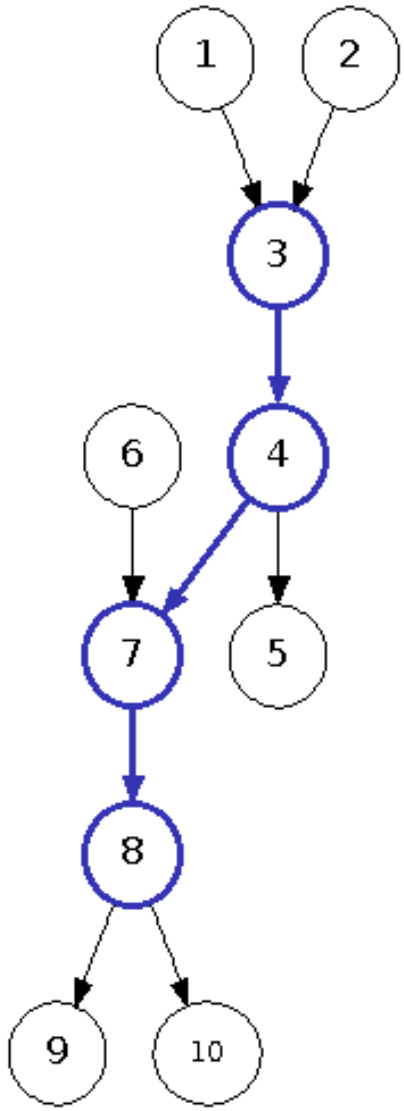
1. $P \leftarrow \{\}$
2. **for each** $e \in E(\mathbf{G})$:
3. $\quad p \leftarrow (e)$
4. $\quad (u, v) \leftarrow e$
5. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (v, w) \in E(\mathbf{G})$:
6. $\quad\quad p \leftarrow (p, (v, w))$
7. $\quad\quad (u, v) \leftarrow (v, w)$
8. $\quad (u, v) \leftarrow e$
9. $\quad$ **while** $\exists! w \in V(\mathbf{G}) : (w, u) \in E(\mathbf{G})$:
10. $\quad\quad p \leftarrow ((w, u), p)$
11. $\quad\quad (u, v) \leftarrow (w, u)$
12. $\quad P \leftarrow P \cup \{p\}$
13. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

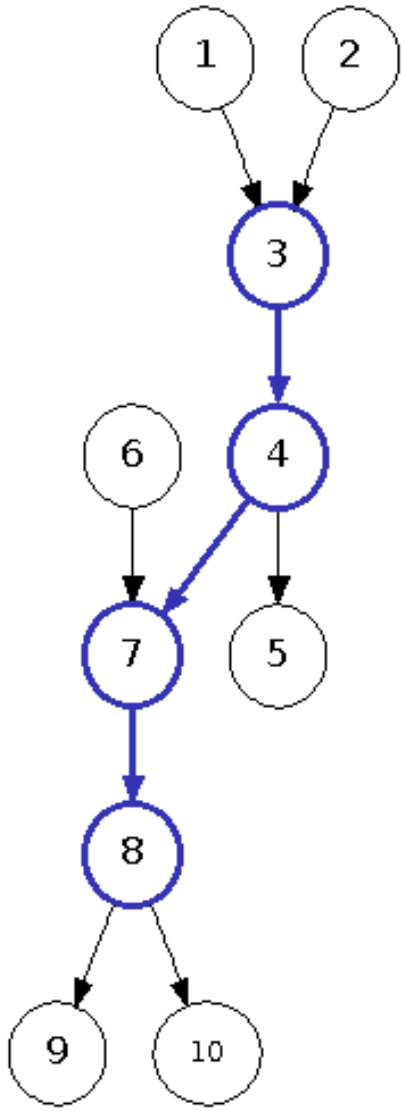
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

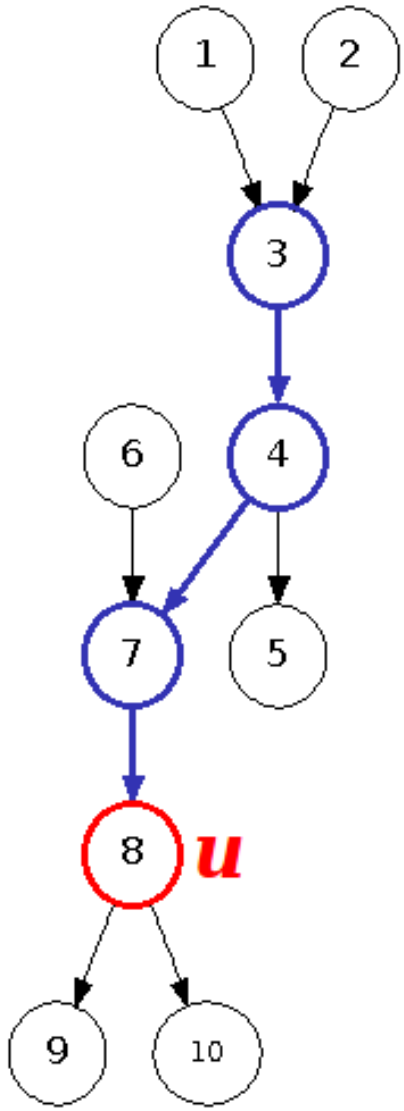
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u,v)|(u,v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

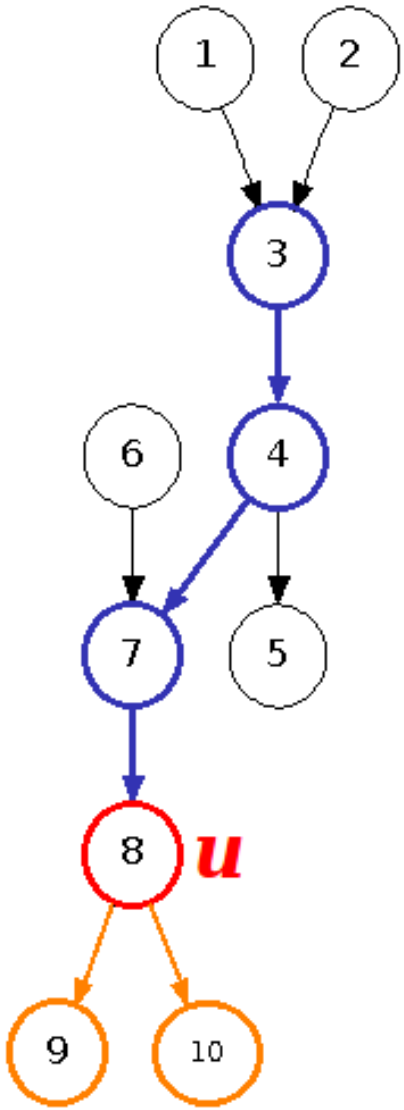
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u,v) | (u,v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

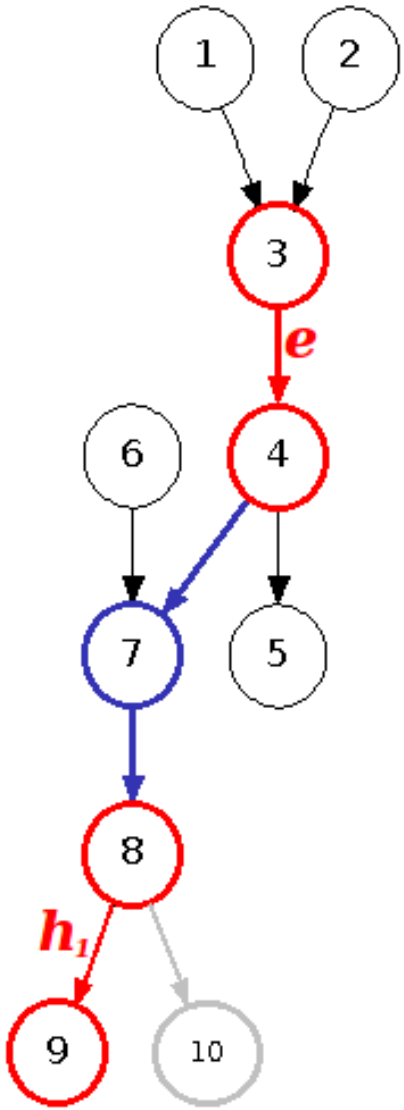
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u,v) | (u,v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

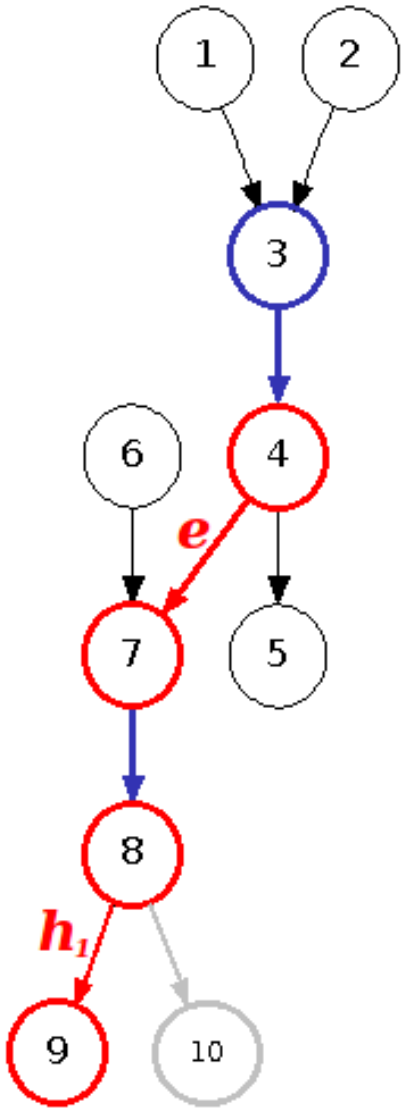
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j, \forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

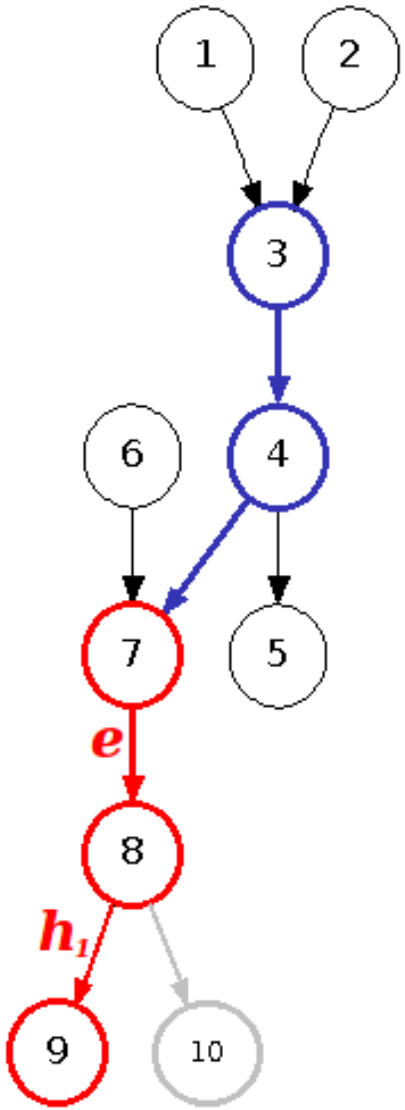
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u,v) | (u,v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

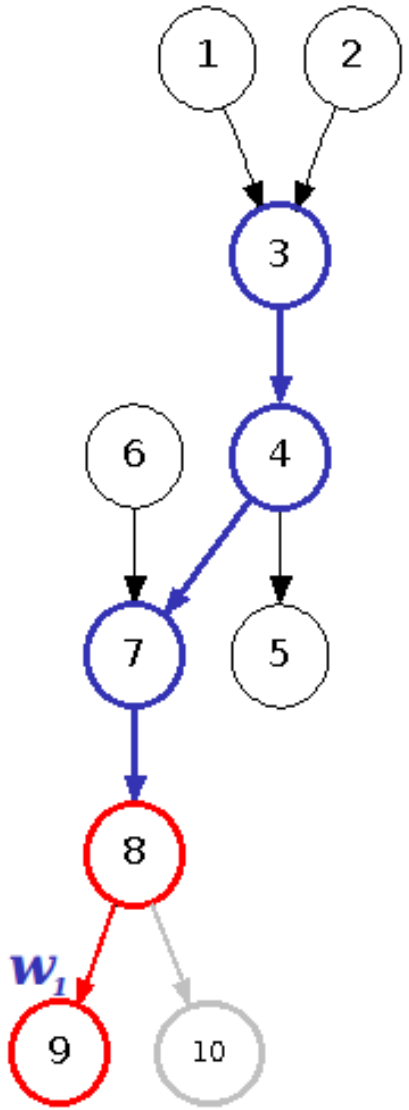
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j, \forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

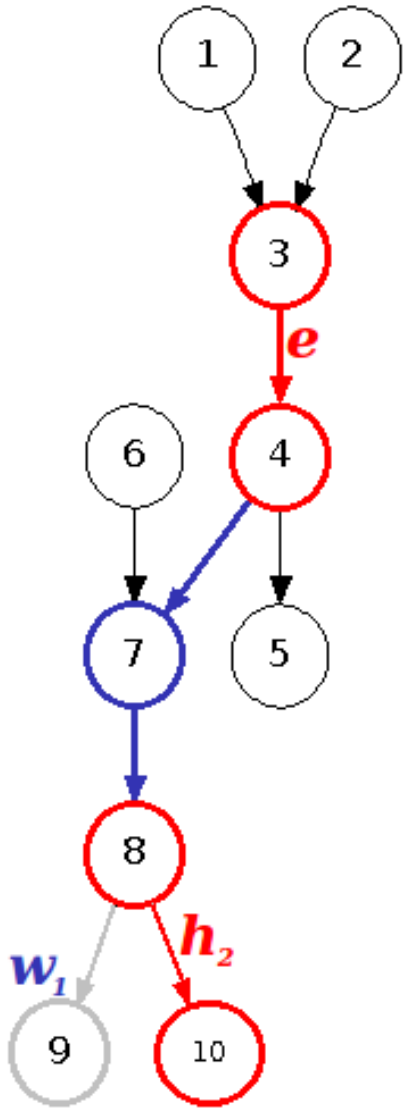
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j, \forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

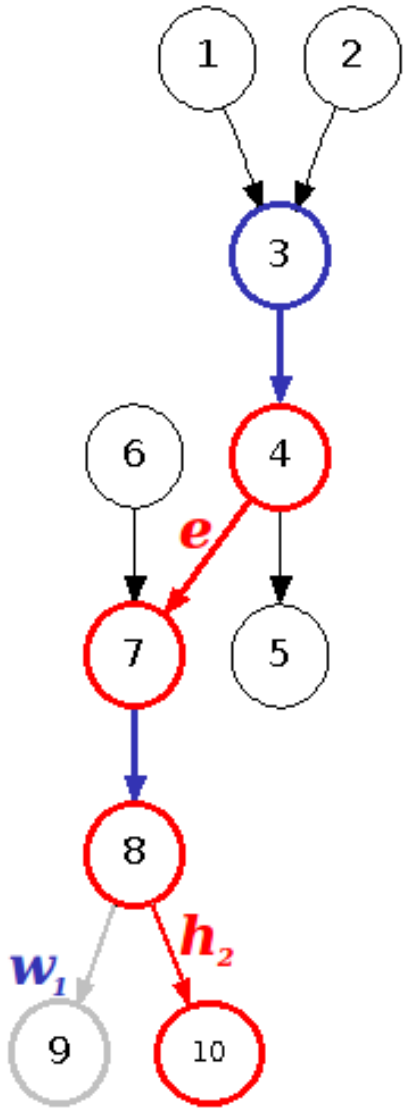
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

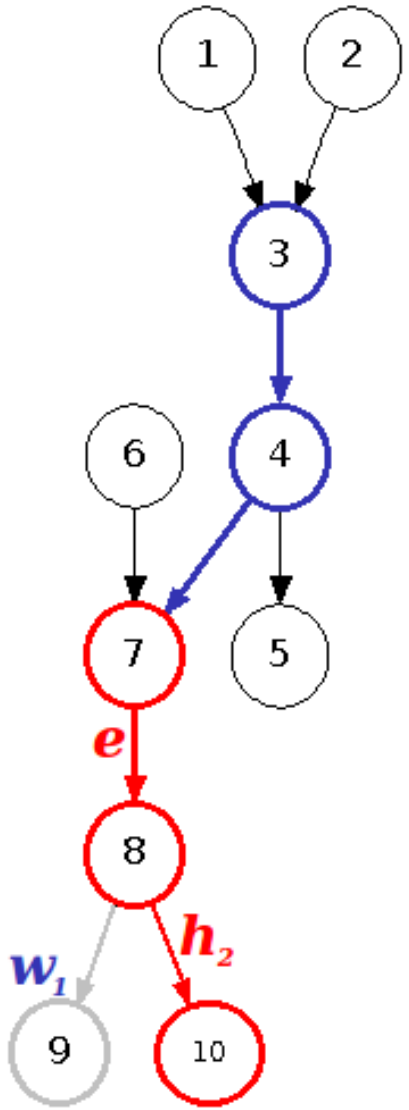
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

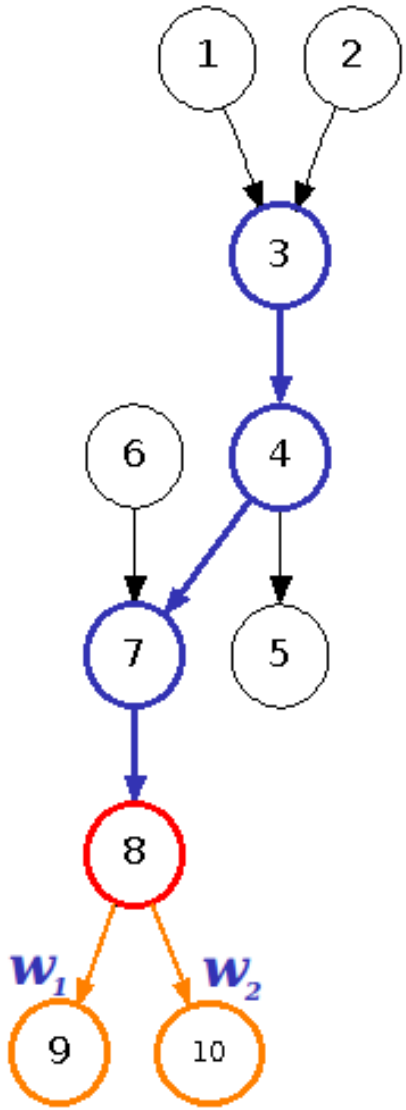
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

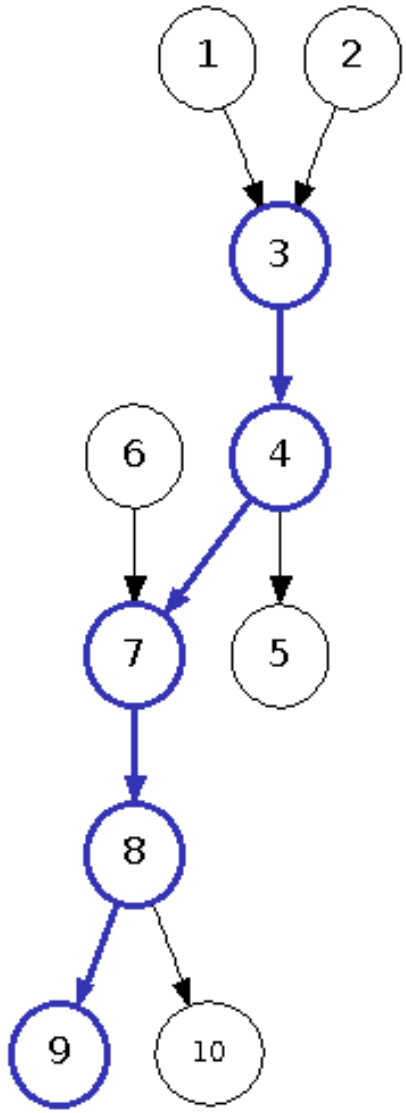
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

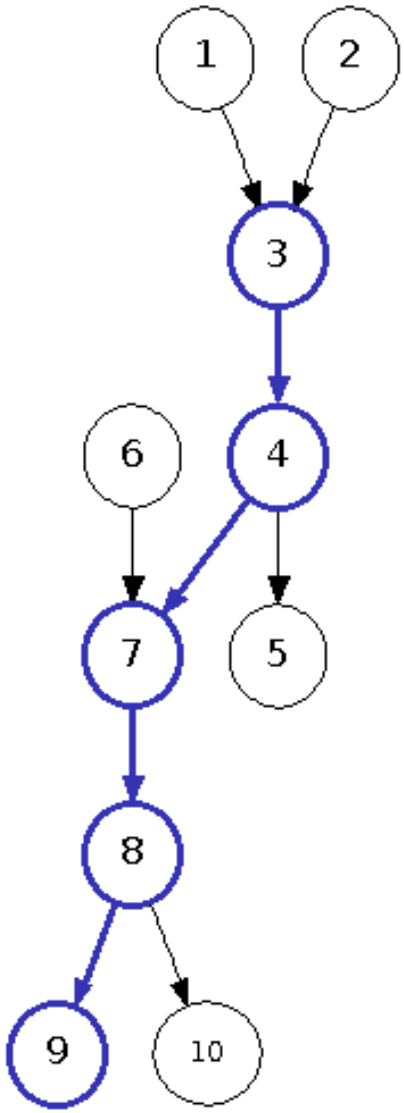
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

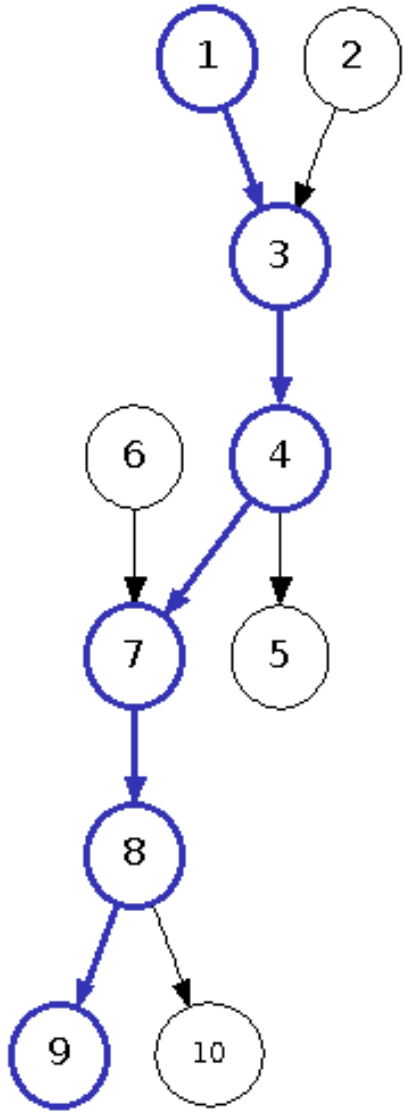
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. return P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

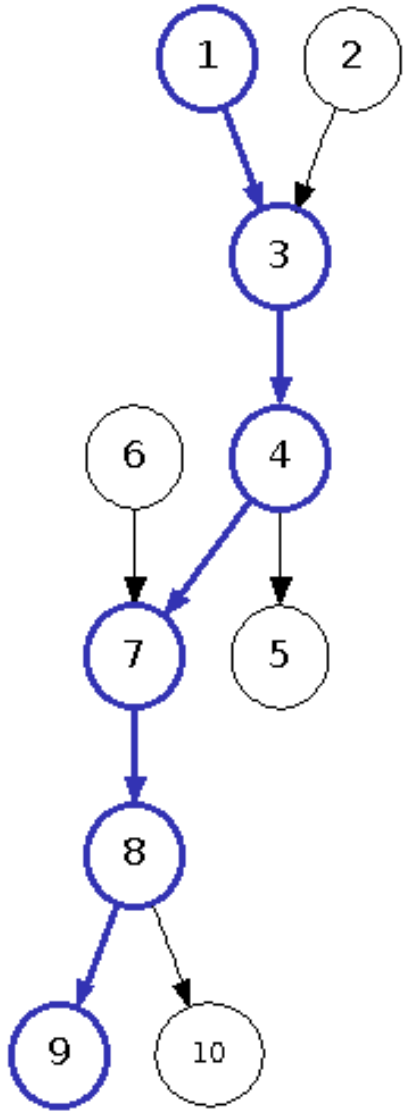
1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. return P



Удлинение путей

EXTEND$(\mathbf{G}, P, W, \theta, C)$

1. **for each** $p \in P$:
2. **repeat** :
3. $q \leftarrow p$
4. $u \leftarrow LAST(p)$
5. $H \leftarrow \{(u, v) | (u, v) \in E(\mathbf{G})\}$
6. **for each** $h_i \in H$:
7. $w_i \leftarrow \sum_{e \in p} W(e, h_i, D_p(e, h_i))$
8. **if** $\exists! i : w_i > \theta \wedge w_i > C \cdot w_j,$ $\forall j \in \{1, \ldots, |H|\}/\{i\}$:
9. $p \leftarrow (p, h_i)$
10. **untill** $q = p$
11. **return** P

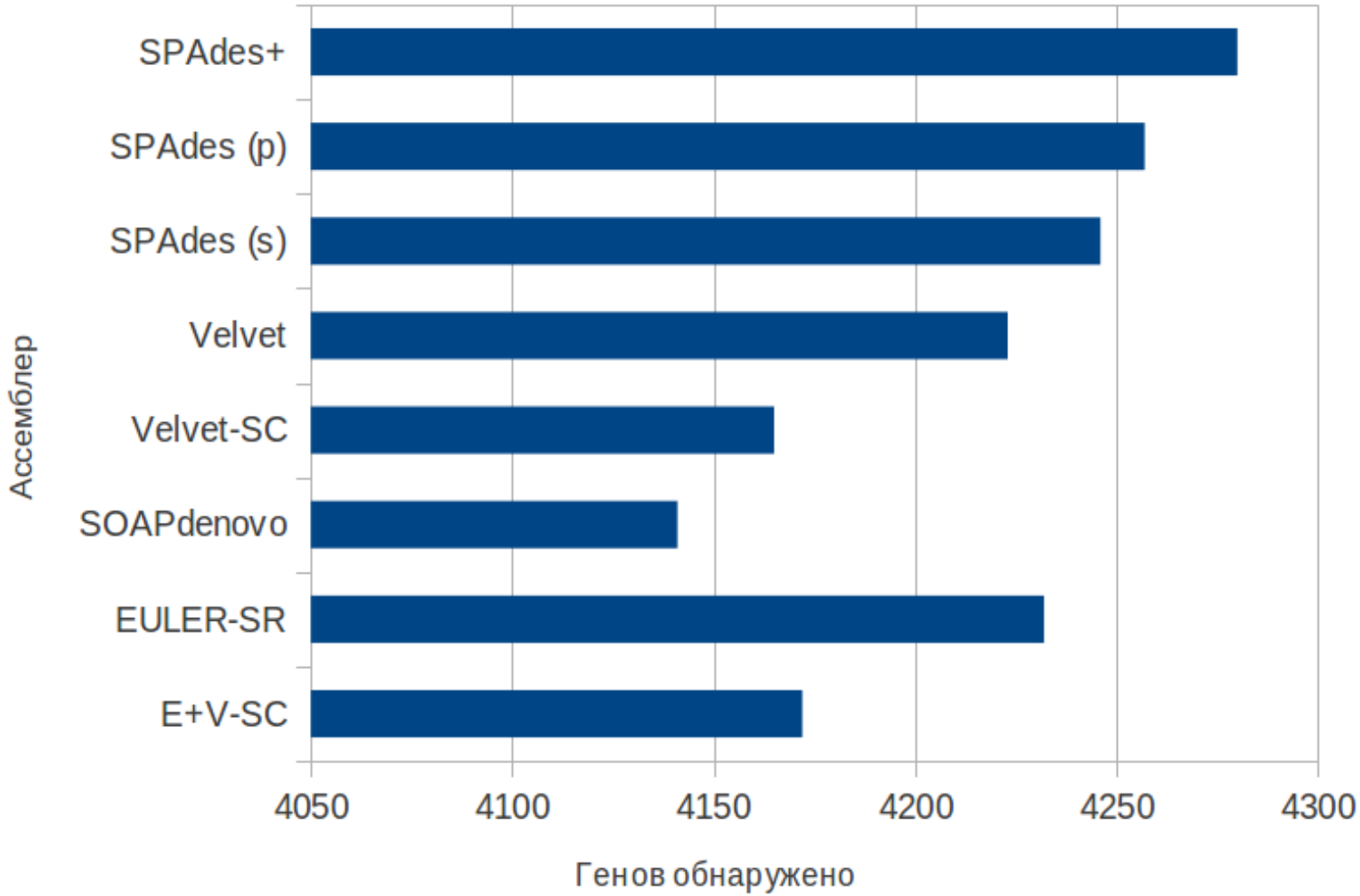


Реализация

- Модуль интегрирован в геномный ассемблер SPAdes
- Реализация включает в себя ряд дополнительных эвристик, учитывающих особенности строения графов и ошибки в начальных данных
- Архитектура проекта позволяет легко добавлять и изменять существующие
  - Процедуры роста путей
  - Стратегии выбора продолжений на каждом шаге
  - Алгоритмы подсчета парной информации

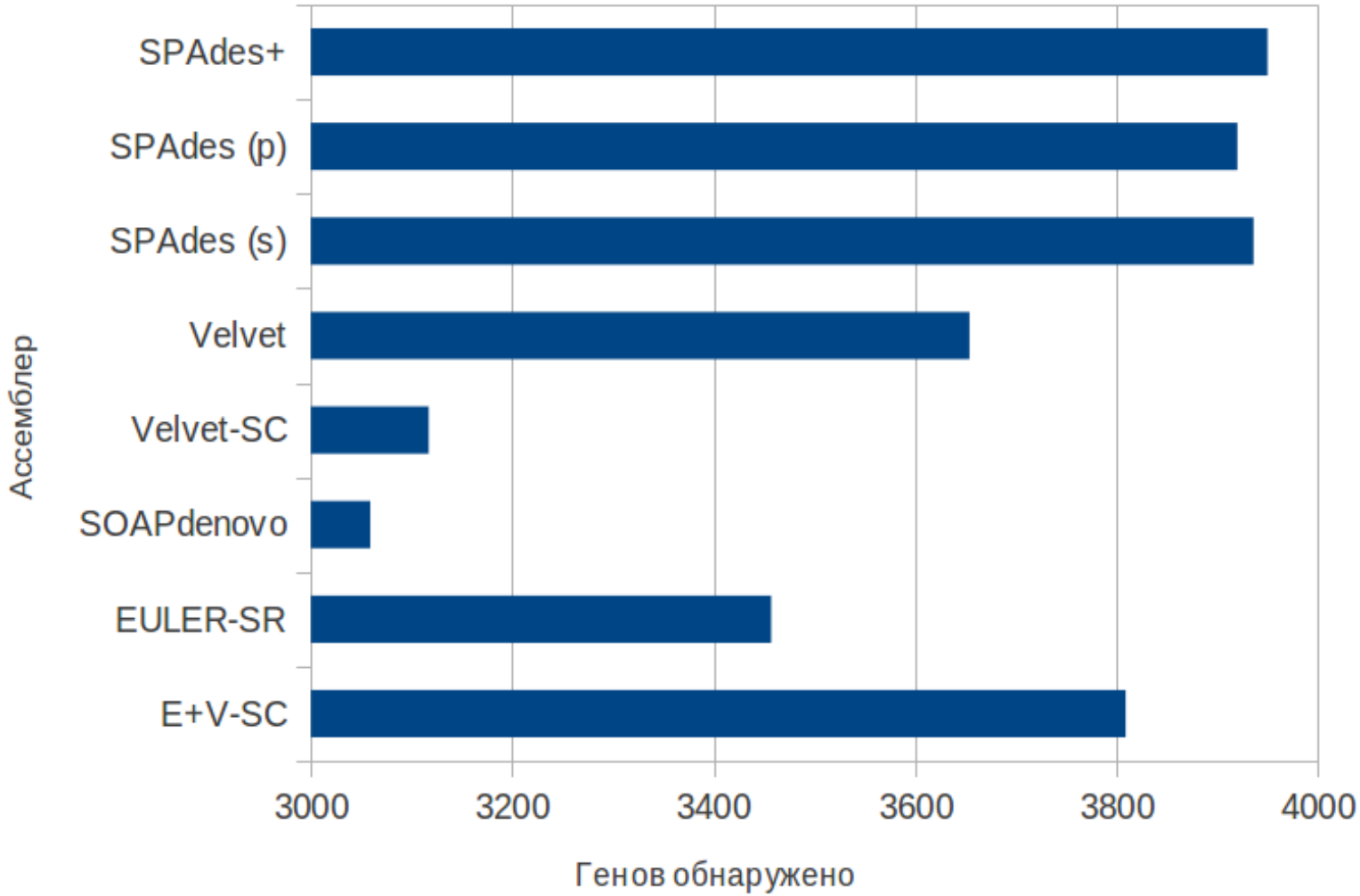
Результаты

- Сравнение по количеству обнаруженных генов
- Геном бактерии *Escherichia coli* (несколько клеток)

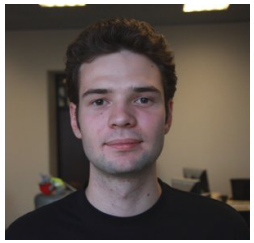


Результаты

- Сравнение по количеству обнаруженных генов
- Геном бактерии *Escherichia coli* (одна клетка)

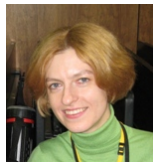


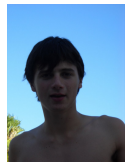
Благодарности



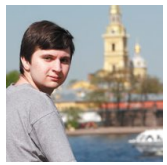
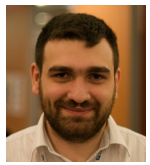
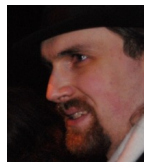








Спасибо за внимание!