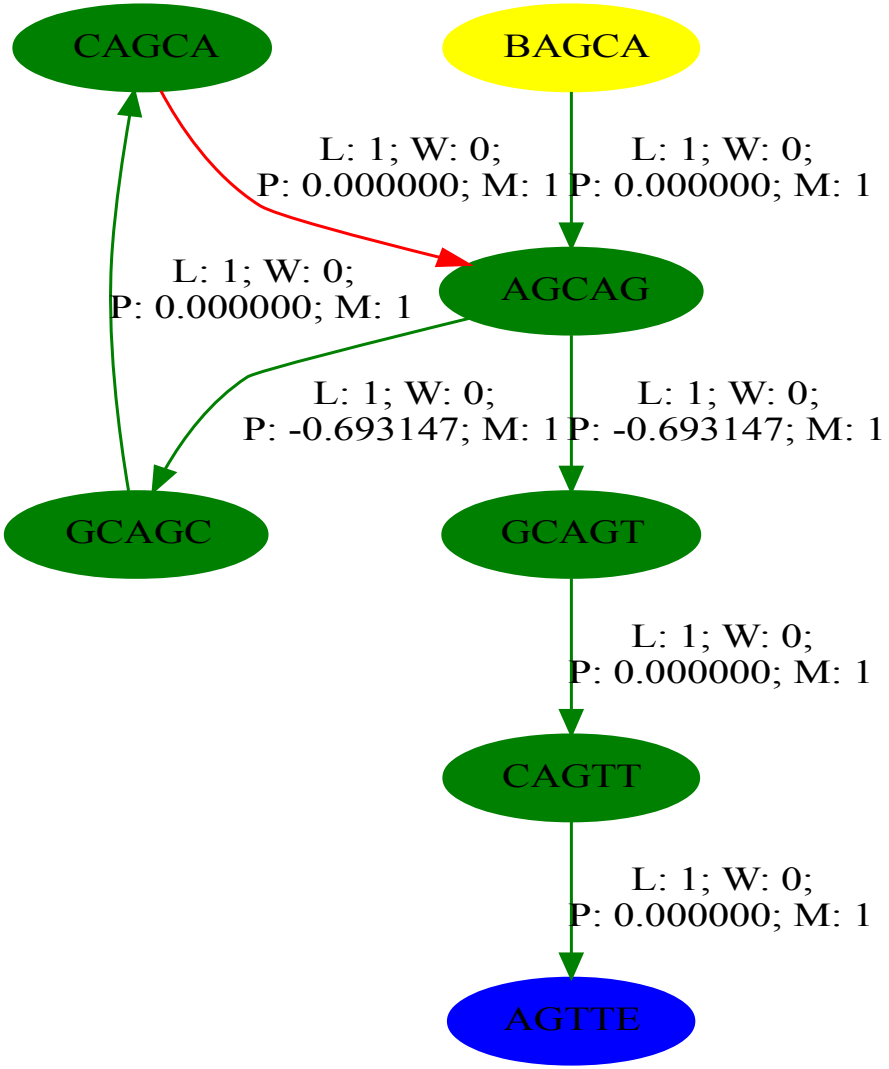
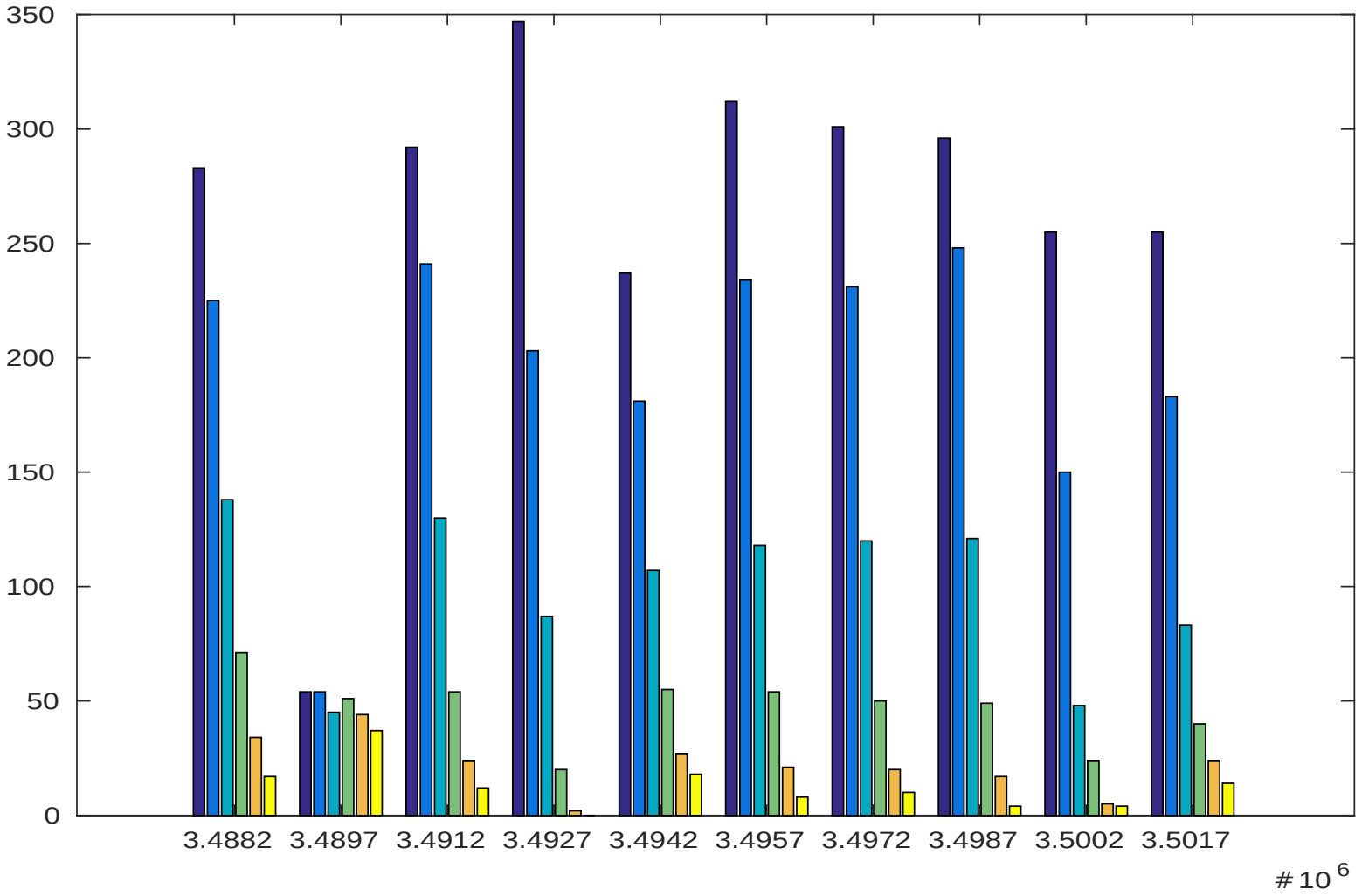
Zpětné hrany

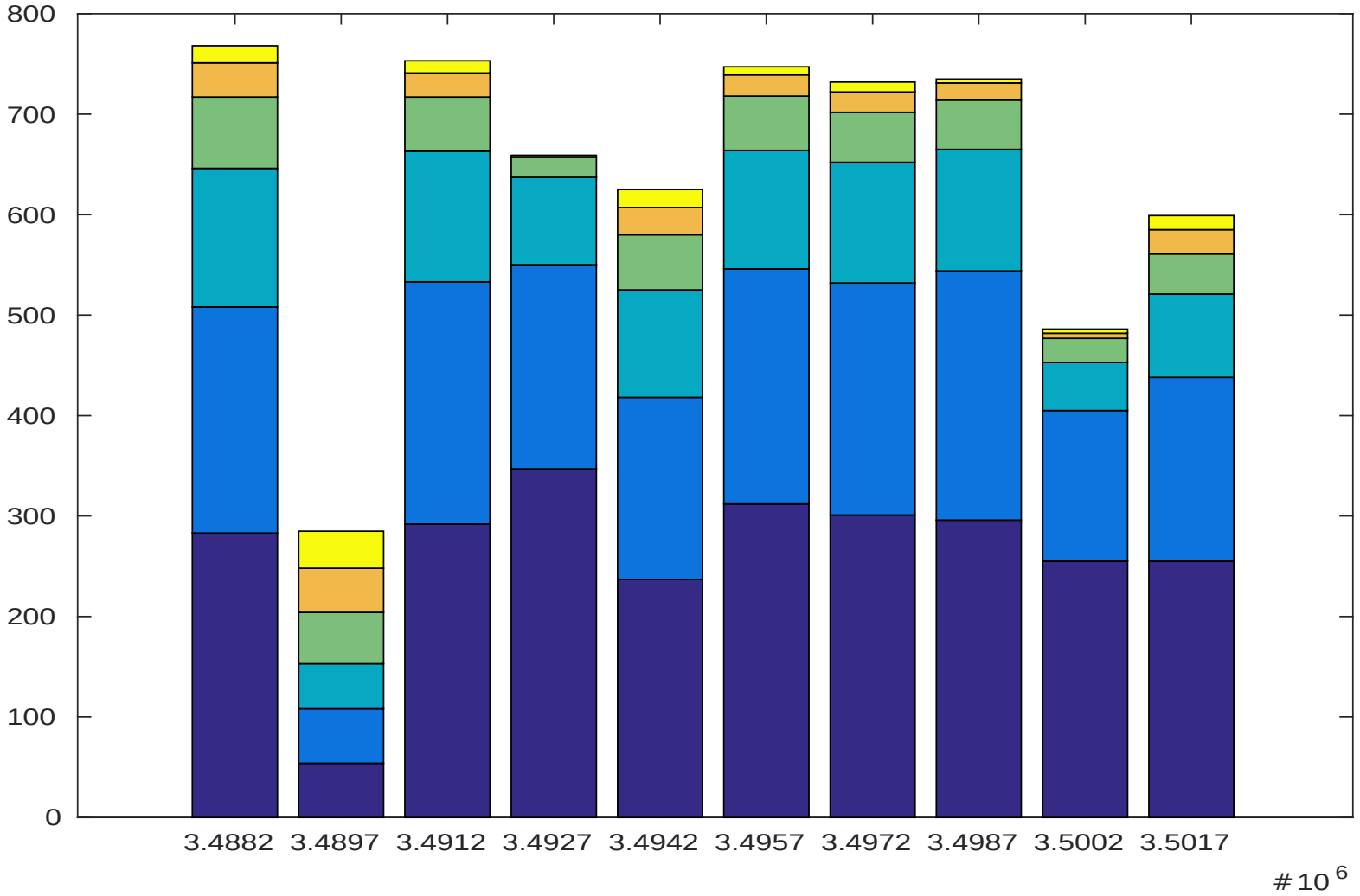
Definice (AGCAGCAGTT)



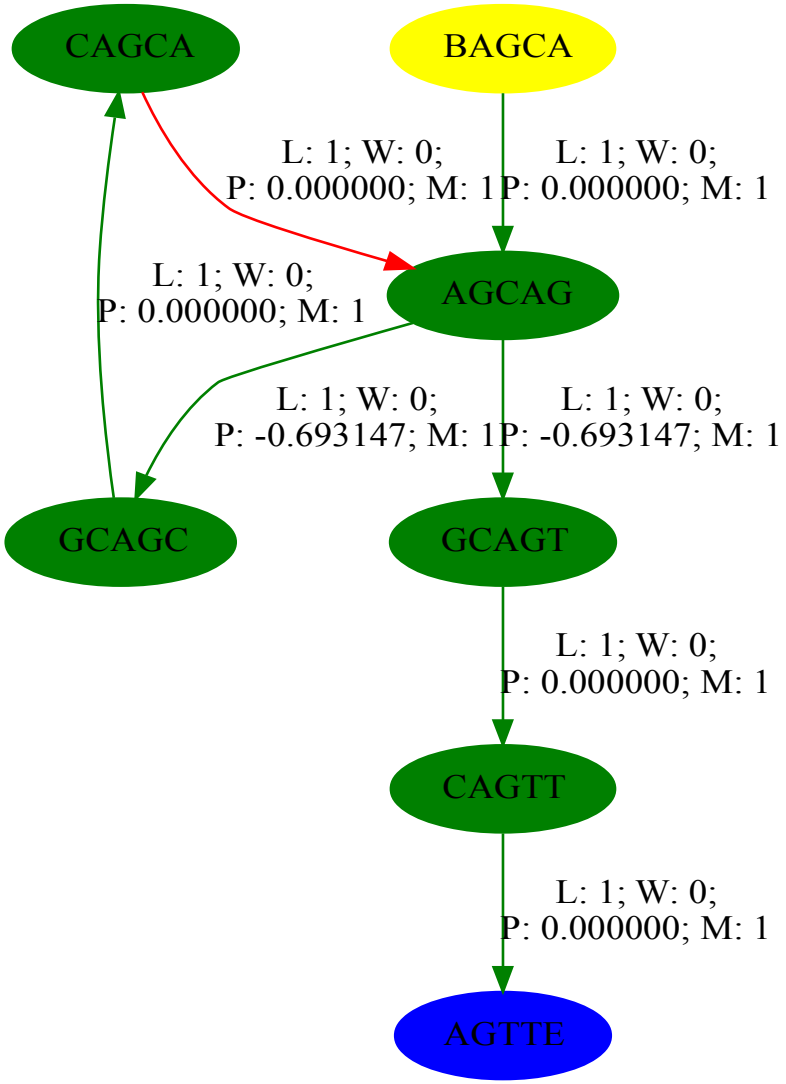
Závislost počtu zpětných hran na K (k = 5..10)

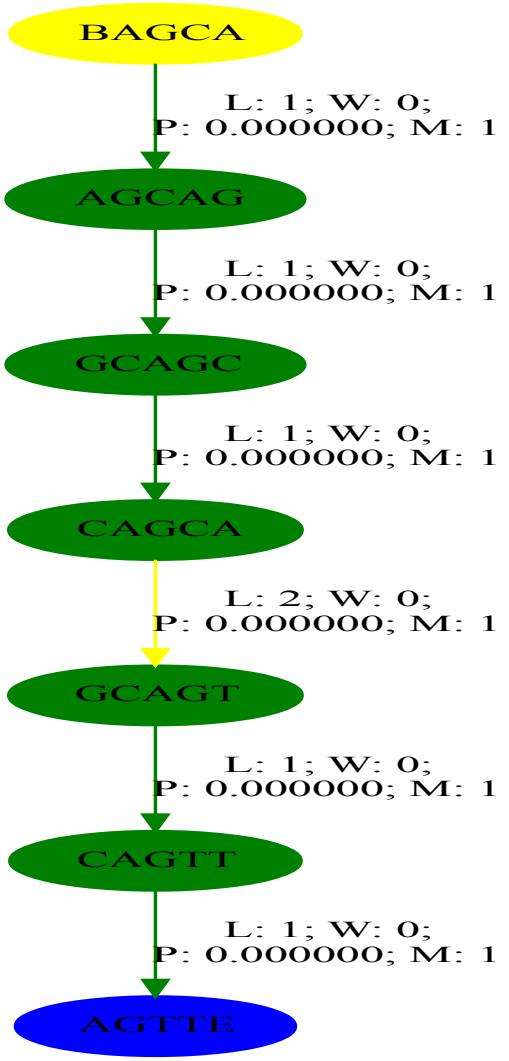


Závislost počtu zpětných hran na K (k = 5..10)

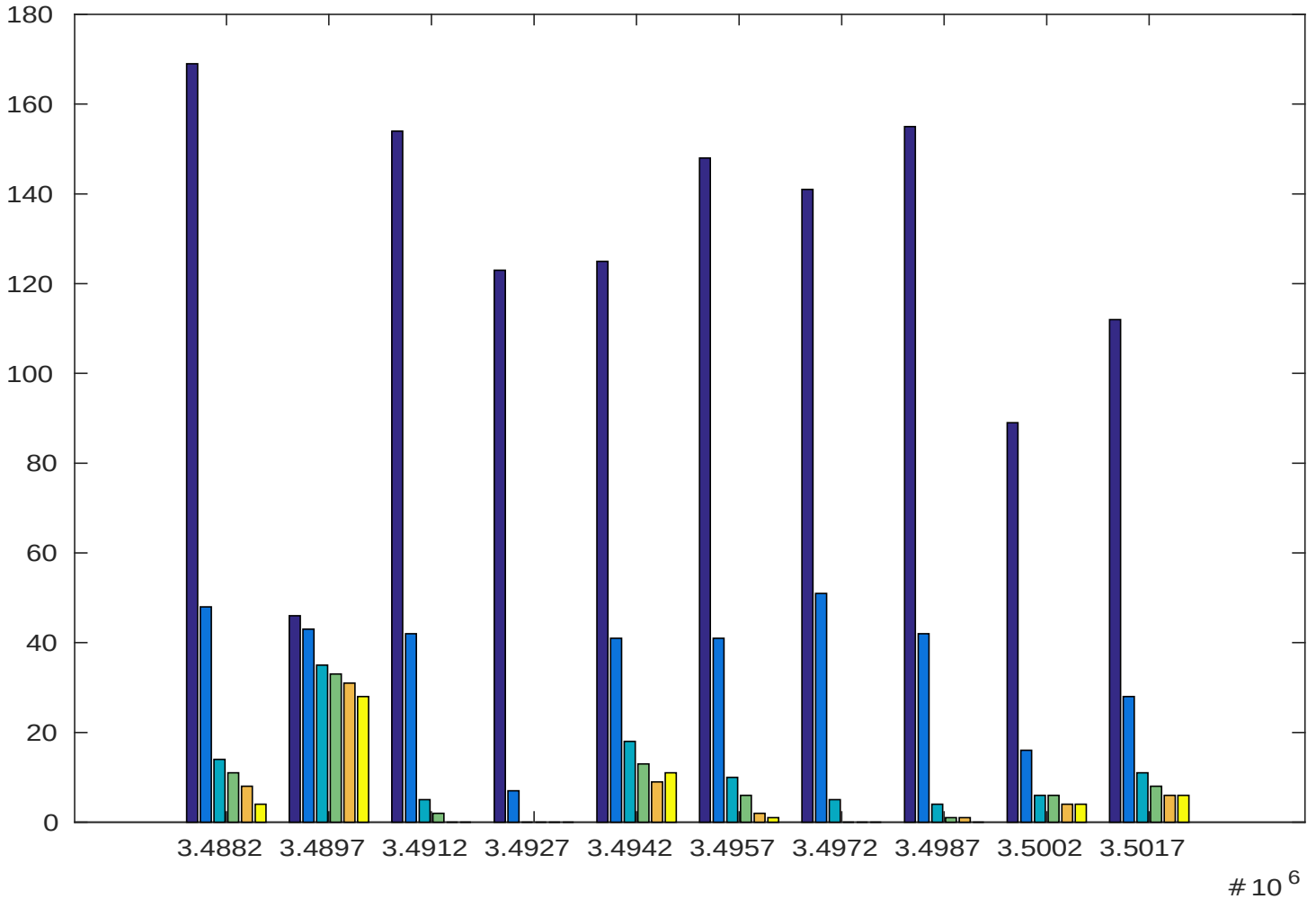


Heuristika (AGCAGCAGTT)

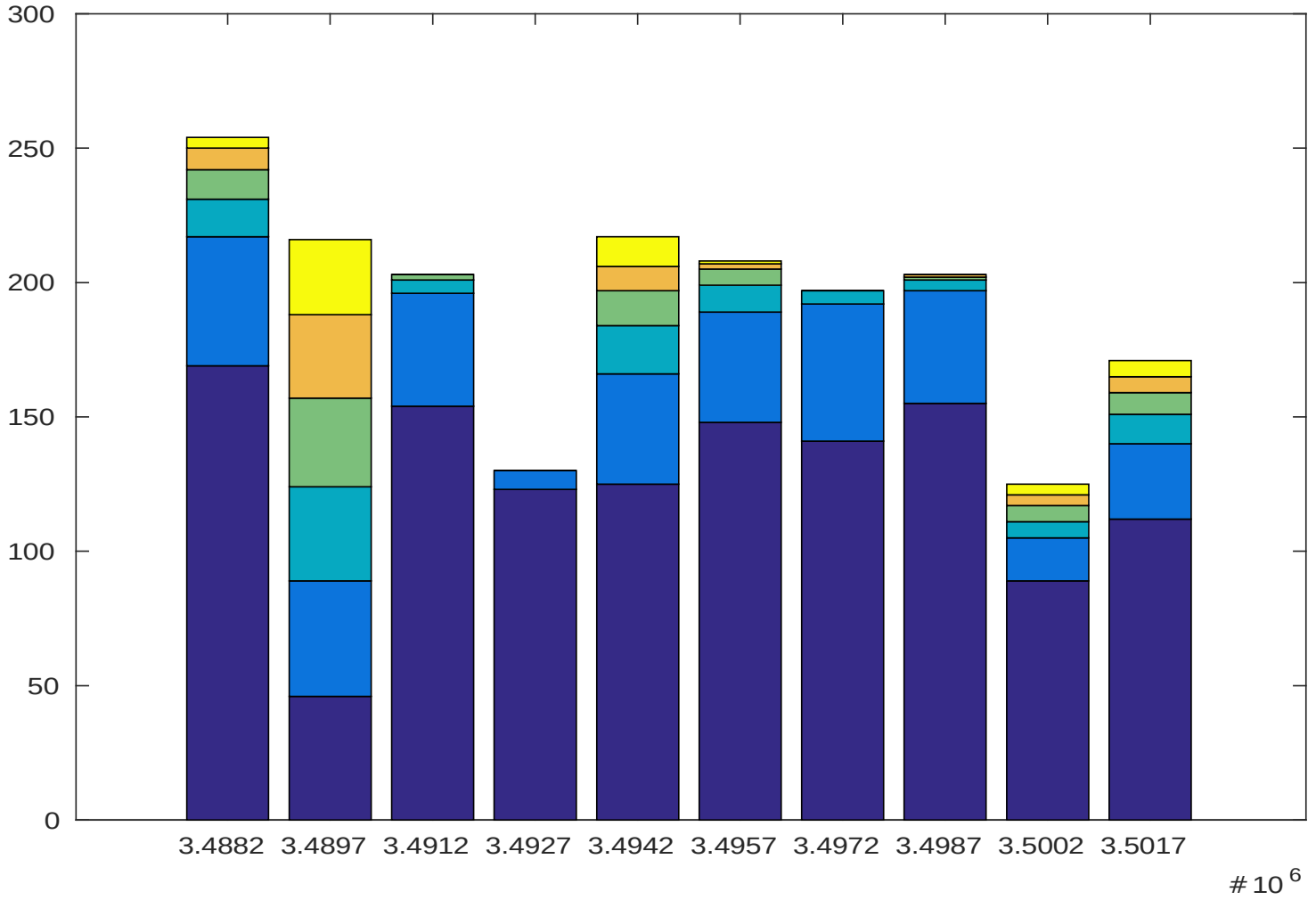




Závislost počtu zpětných hran na K (k = 5..10), heuristika



Závislost počtu zpětných hran na K (k = 5..10), heuristika



Graph Assembly

Referenční sekvence

ACGCTCCGAC

ACGCTCCGAC

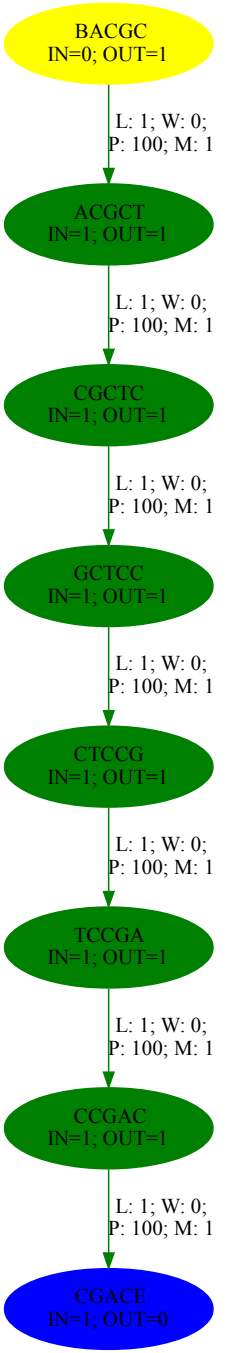
ACGCTCCGAC

ACGCTCCGAC

ACGCTCCGAC

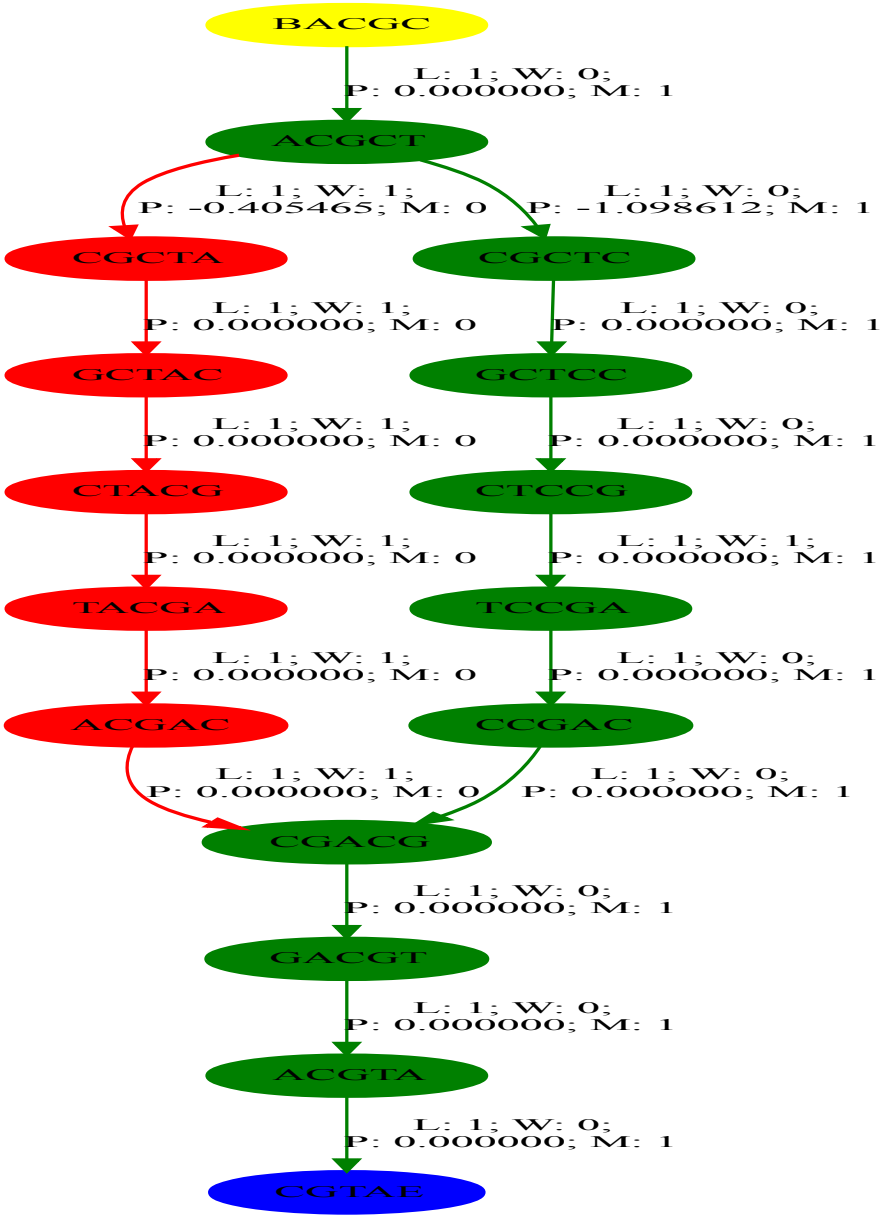
ACGCTCCGAC

Vrcholy značící počátek a konec sekvence



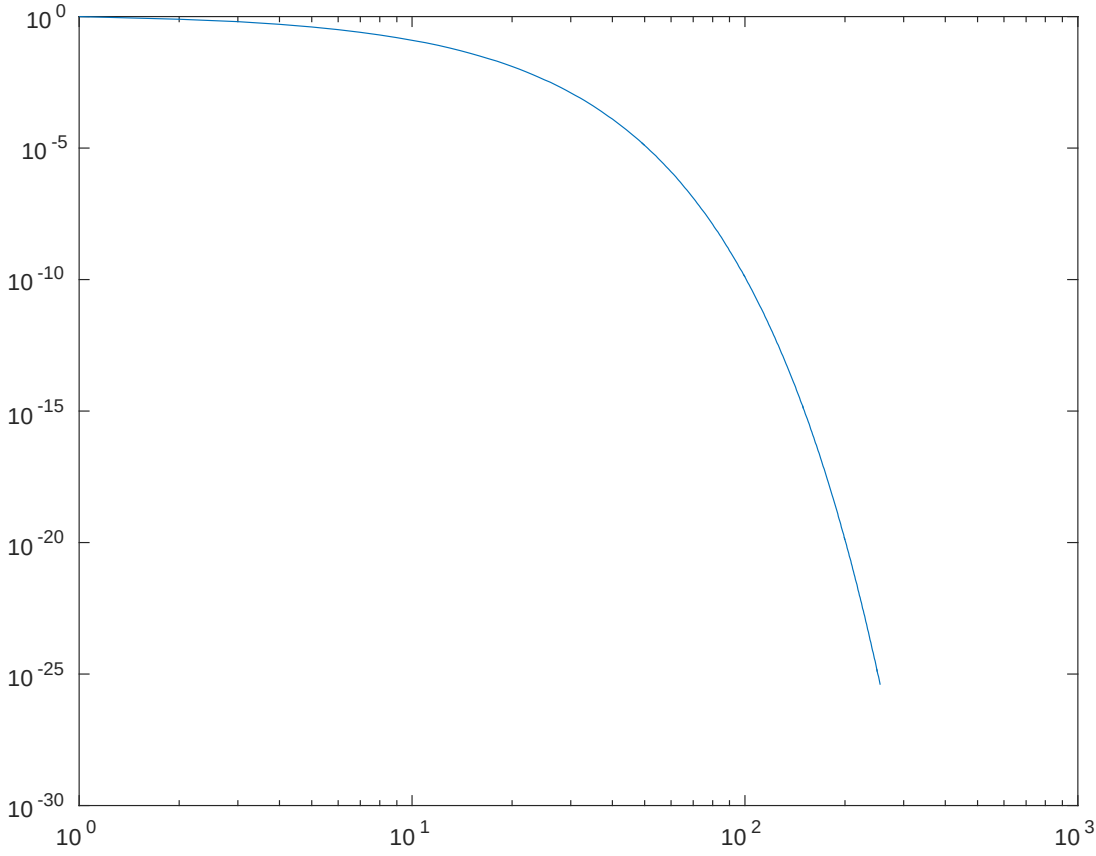
Přidávání čtení (readů)

- W = počet čtení procházejících danou hranou
- L = délka hrany
- M = Maximální povolený počet průchodů (pouze pro hrany referenční sekvence)
- P = pravděpodobnost přechodu



Čtení

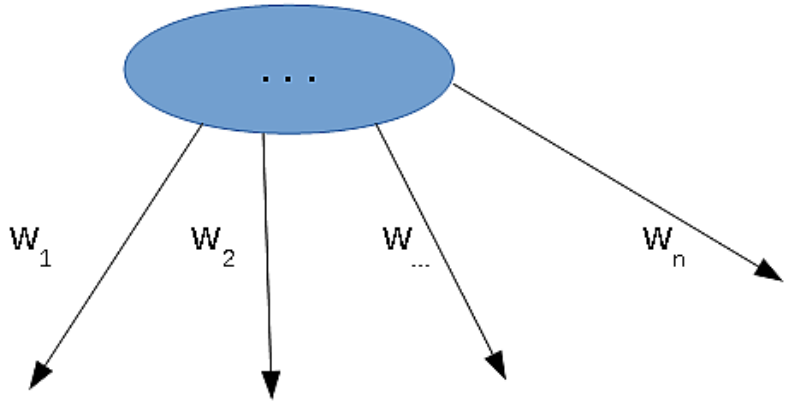
- Načítána z formátu SAM
- Načtena všechna povinná pole (11)
- Zatím se využívá jen POS a SEQ
- Jaké hodnoty QUAL (MAPQ) jsou špatné?
- Co je „reverse complemeted“?



Optimalizace grafu

- Odstranění hran, kterými neprochází dostatečný počet čtení
  - Pouze pro hrany přidané některým ze čtení
  - (asi jsem zkoušel regiony s malým počtem čtení)
- Kontrakce sekvenců vrchoů se vstupním a výstupním stupňem = 1
  - Pouze pro účely zobrazování

Pravděpodobnost přechodu



- $P_i = \ln((w_i + 1) / (w_1 + w_2 + . . . + w_n + n))$
- n – výstupní stupeň vrcolu
- w_i – počet čtení procházejících i-tou hranou

Hledání nejlepších tahů

- Co nejvyšší součet P_i
  - Dobře se ořezává (čím delší tím menší součet)
- Délka odpovídající délce aktivního regionu
  - Nedává moc dobré výsledky (možná špatné SWA)
  - Vyzkouším povlit mírně pohyblivou délku
- Vybírá se N nalezených nejlepších
- Prohledávání do hloubky
- Omezení počtu průchodů hranami
  - Omezení je dáno tím, kolikrát danou hranou prošla referenční sekvence