Bioinformatikai Algoritmusok 10. GY

Legrövidebb közös superstring

Lehotay-Kéry Péter
lkp@caesar.elte.hu

Forditotta: Nyíri Tamás
nytuaai@gmail.com
http://people.inf.elte.hu/nytuaai

(Ben Langmead diasora alapján)

Az assembly probléma megfogalmazása

Az átfedések keresése fontos, és vissza fogunk térni rá, de a végső célünk a genom rekreálása (assembly-je).

Hogyan tudjuk ezt a problémát megfogalmazni?

Első próbálkozás: a legrövidebb közös superstring (*SCS*) keresése

Legrövidebb közös superstring

Adott egy S string-ek halmaza. Találjuk meg a legrövidebb string-et aminek minden S-beli string rész-string-je.

A „legrövidebb” megszorítás nélkül egyszerű: csak konkatenáljuk őket:

Példa: *S*: BAA AAB BBA ABA ABB BBB AAA BAB

Konkatenáció: BAAAABBBAABAABBBBBBAAABAB

24

SCS(S): AAABBBABAA

10

AAA
AAB
ABB
BBB
BBA
BAB
ABA
BAA

Legrövidebb közös superstring

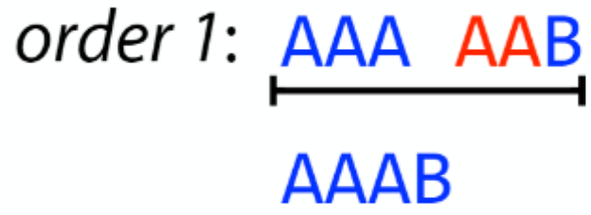
Ötlet: Vegyük egy rendezését az S-beli stringeknek és konstruáljunk egy superstringet!

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAA

Legrövidebb közös superstring

Ötlet: Vegyük egy rendezését az S-beli stringeknek és konstruáljunk egy superstringet!



Legrövidebb közös superstring

Ötlet: Vegyük egy rendezését az S-beli stringeknek és konstruáljunk egy superstringet!

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAABA

Legrövidebb közös superstring

Ötlet: Vegyük egy rendezését az S-beli stringeknek és konstruáljunk egy superstringet!

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAABABB

Legrövidebb közös superstring

Ötlet: Vegyük egy rendezését az S-beli stringeknek és konstruáljunk egy superstringet!

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAABABBAABABBABBB ⟵ superstring 1

Legrövidebb közös superstring

Ötlet: Vegyük egy rendezését az S-beli stringeknek és konstruáljunk egy superstringet!

order 1: AAA AAB ABA ABB BAA BAB BBA BBB

AAABABBAABABBABBB ⟵ superstring 1

order 2: AAA AAB ABA BAB ABB BBB BAA BBA

AAABABBBAABBA ⟵ superstring 2

Keressünk meg minden lehetséget rendezést és válasszuk a legrövidebb superstringet!

Legrövidebb közös superstring

Meg tudjuk oldani?

Képzeljünk el egy módosított átfedési gráfot ahol minden él költsége = -(az átfedés hossza)

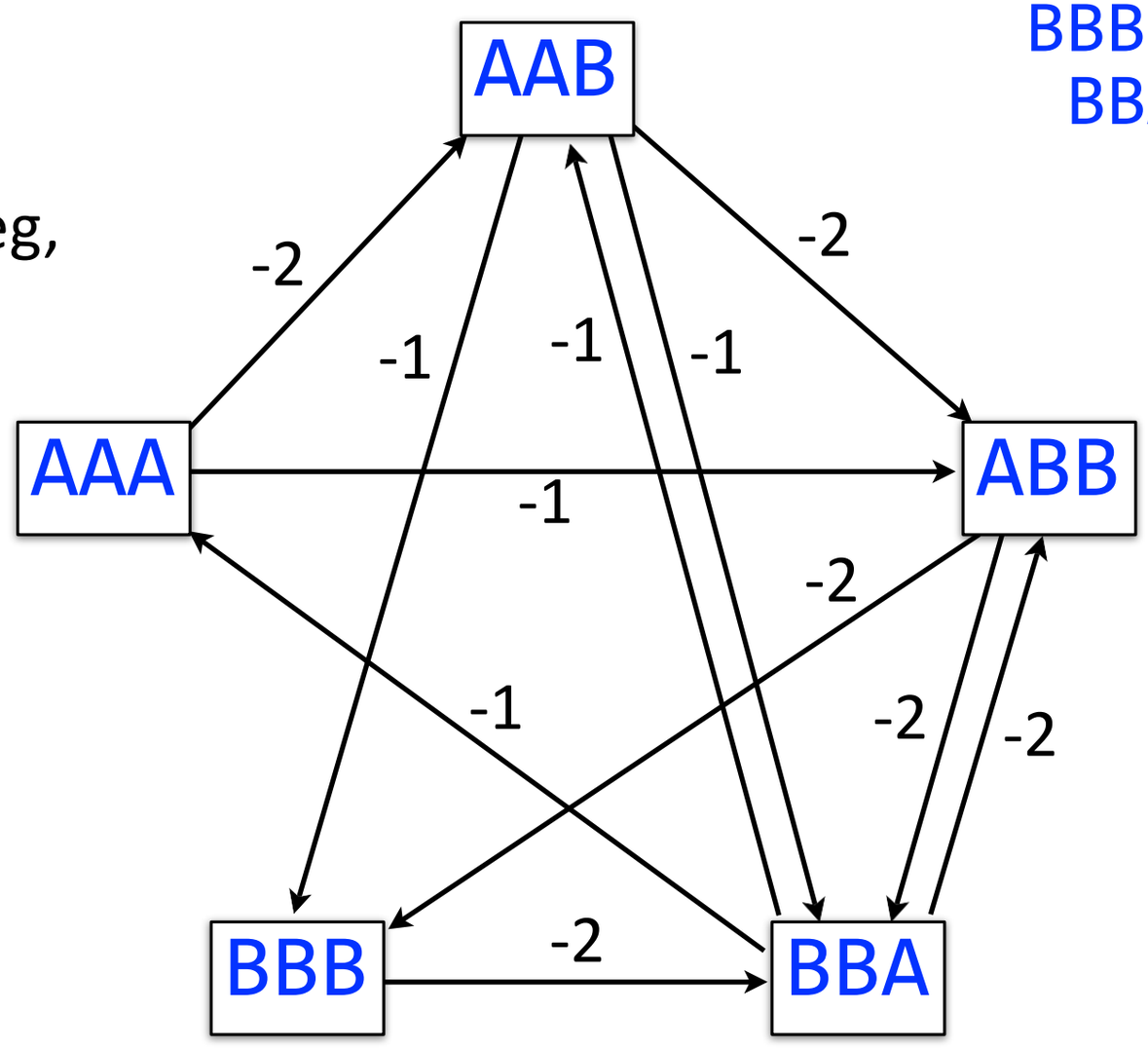
Az SCS megfelel annak az útnak amely mindegyik csúcsot egyszer látogatja meg, minimalizálva az út teljes költségét

Ez az Utazó Ügynök Probléma (*TSP*), ami NP-teljes!

S: AAA AAB ABB BBB BBA

SCS(S): AAABBBA

AAA
AAB
ABB
BBB
BBA



Legrövidebb közös superstring

Most hagyjuk figyelmen kívül az élsúlyokat és keressünk olyan utat ami mindegyik csúcsot pontosan egyszer látogatja meg.

Ezt nevezik Hamilton útnak és ez is NP-teljes.

Ebből láthatjuk hogy az SCS maga is NP-teljes.

S: AAA AAB ABB BBB BBA

SCS(S): AAABBBA

AAA
AAB
ABB
BBB
BBA