Hasonlósági keresés molekulagráfokon: legnagyobb közös részgráf keresése

Kovács Péter

ChemAxon Kft., ELTE IK

kpeter@inf.elte.hu

Budapest, 2018.11.06.

Bevezetés

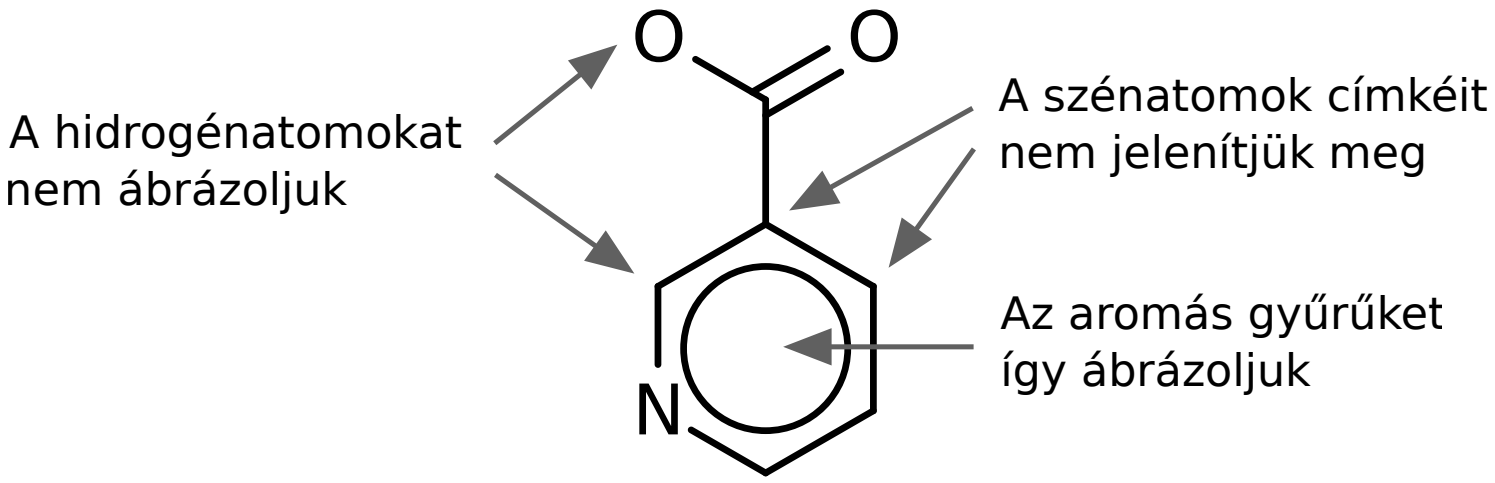
Feladat: két molekulagráf legnagyobb közös részgráfjának meghatározása.

Alkalmazások:

- hasonlósági keresés
- klaszterezés
- 2D/3D illesztés
- reakciók elemzése
- stb.

Forrás: P. Englert and P. Kovács. *Efficient heuristics for maximum common substructure search*. J. Chem. Inf. Model., 55:941–955, 2015. IF: 3.657.

Molekulák reprezentációja

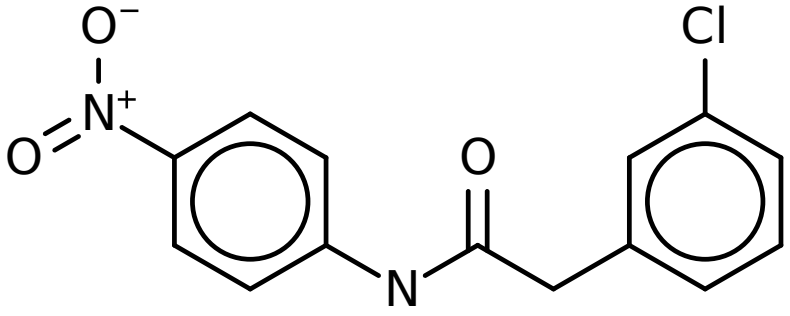


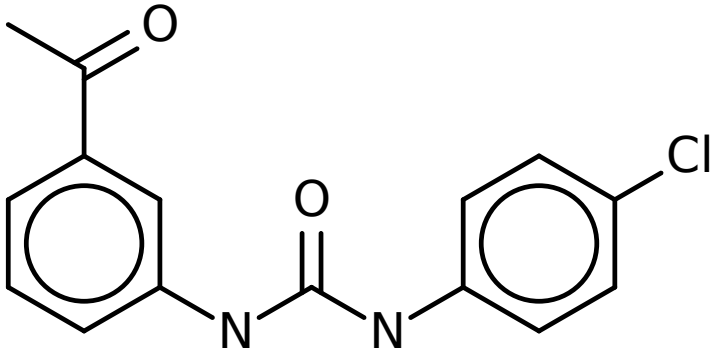
Molekulagráf

- irányítatlan, egyszerű, csúcs- és élcímkézett gráf
- csúcsai az atomokat reprezentálják (pl. C, N, O, F, Cl, Br stb.)
- élei a kémiai kötéseket reprezentálják (egyszeres, kétszeres, aromás stb.)
- a hidrogénatomokat közvetlenül nem reprezentáljuk (általában)

Molekulagráfok hasonlósága

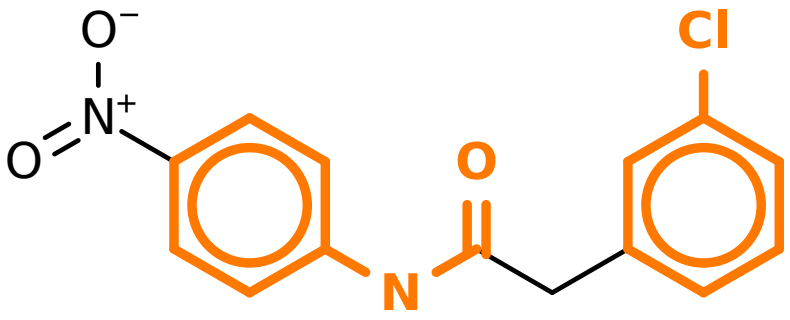
Mennyire hasonló ez a két molekulagráf?

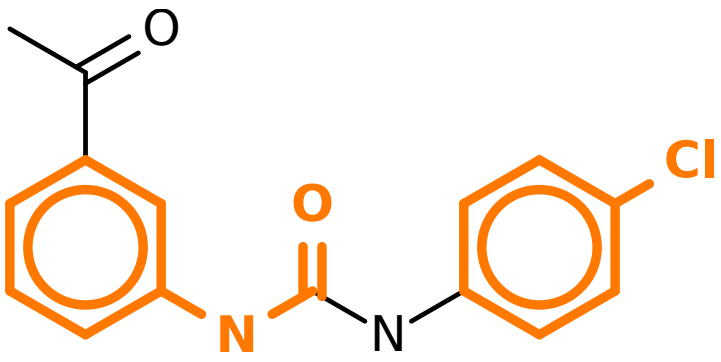




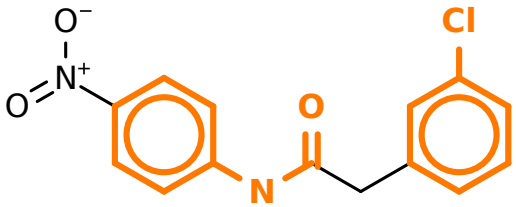
Molekulagráfok hasonlósága

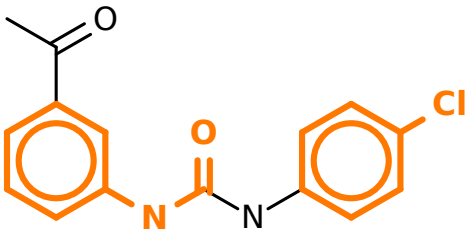
Mennyire hasonló ez a két molekulagráf?





Molekulagráfok hasonlósága

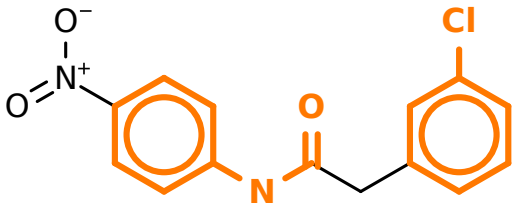


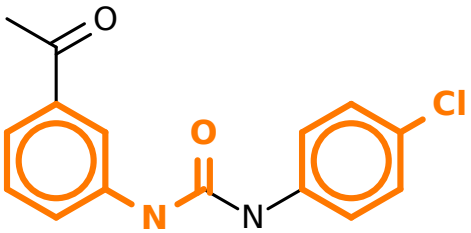


Hasonlóság (Jaccard, Tanimoto):

$$\frac{|A \cap B|}{|A \cup B|} =$$

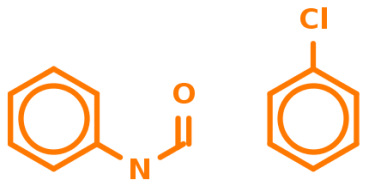
Molekulagráfok hasonlósága

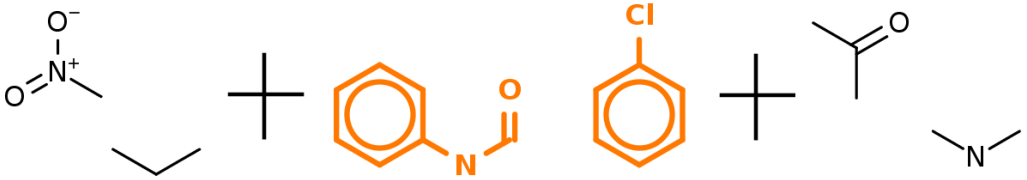




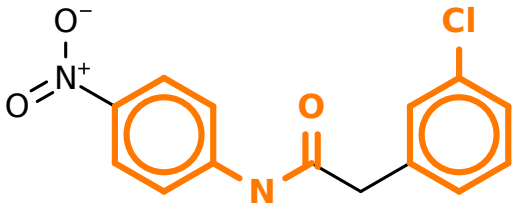
Hasonlóság (Jaccard, Tanimoto):

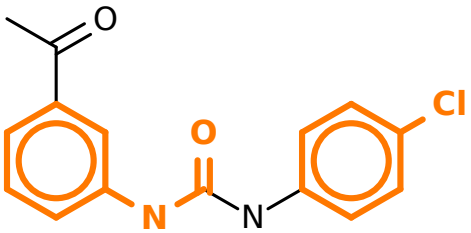
$$\frac{|A \cap B|}{|A \cup B|} =$$





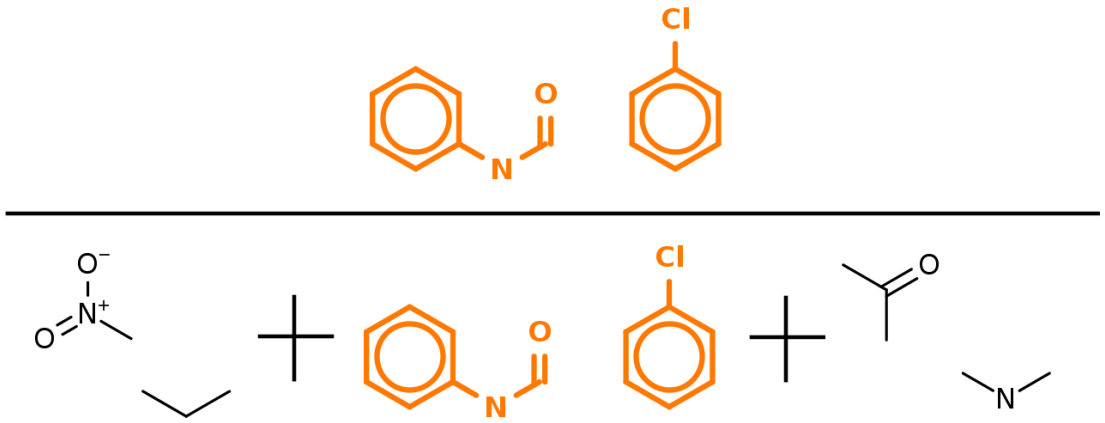
Molekulagráfok hasonlósága





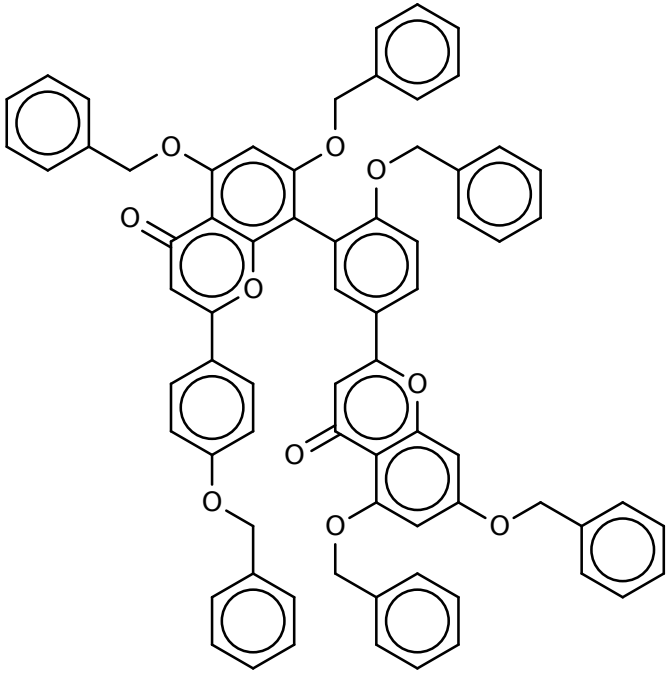
Hasonlóság (Jaccard, Tanimoto):

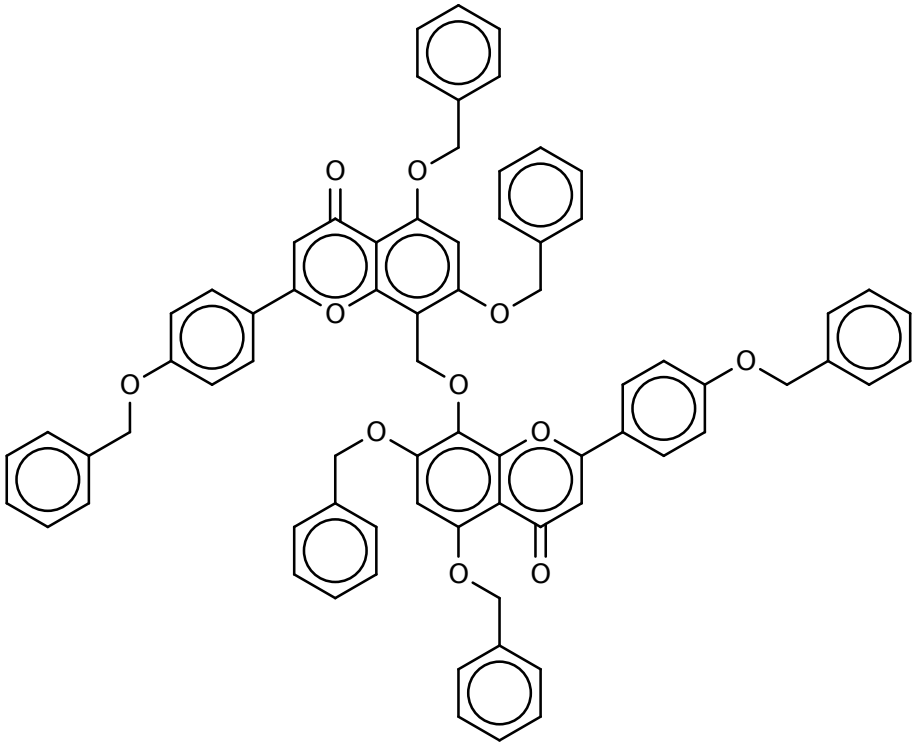
$$\frac{|A \cap B|}{|A \cup B|} =$$



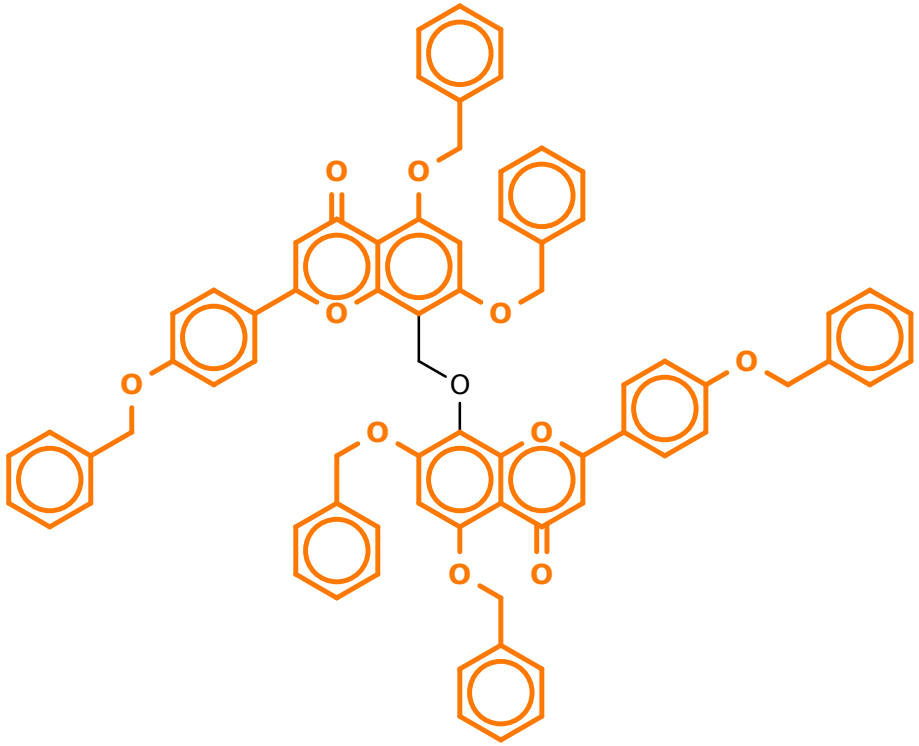
$$= \frac{m_c}{m_1 + m_2 - m_c} = 0{,}615$$

Molekulagráfok hasonlósága





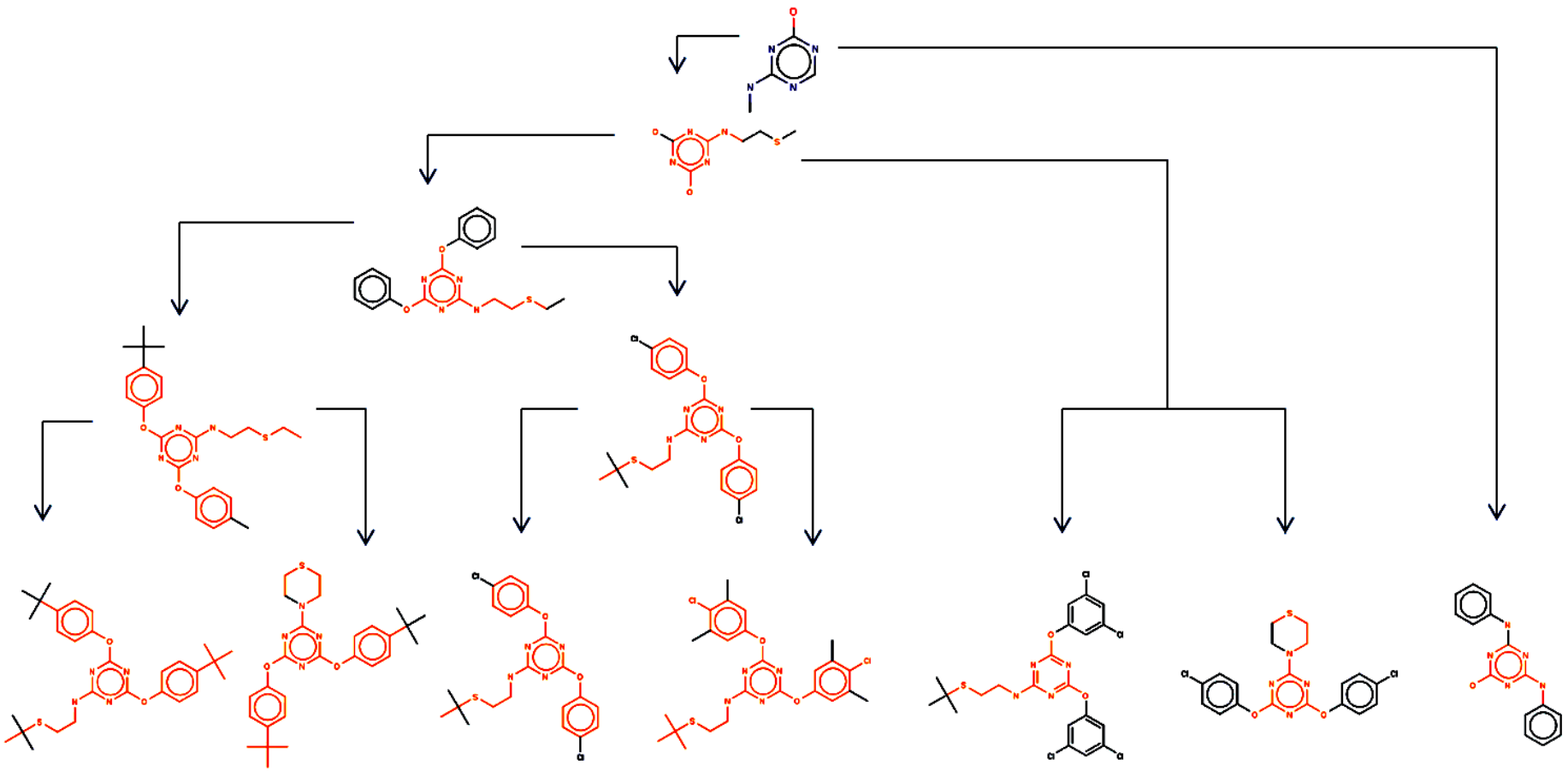
Molekulagráfok hasonlósága



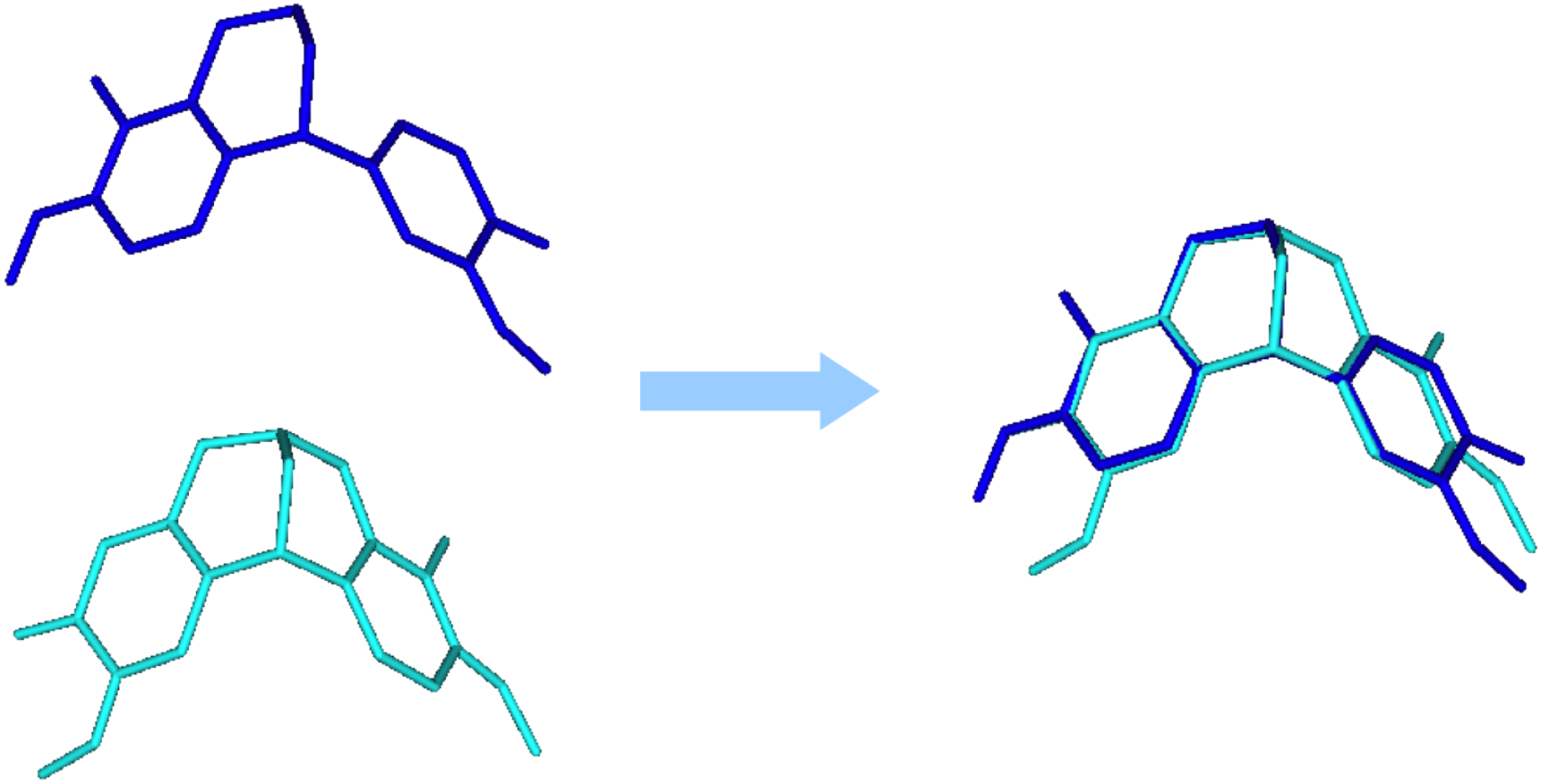
Hasonlóság: 0,958

További alkalmazások

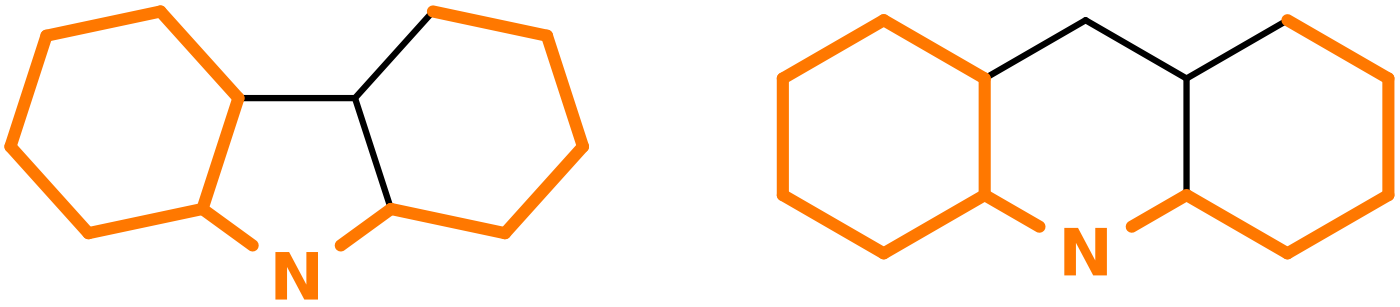
Klaszterezés



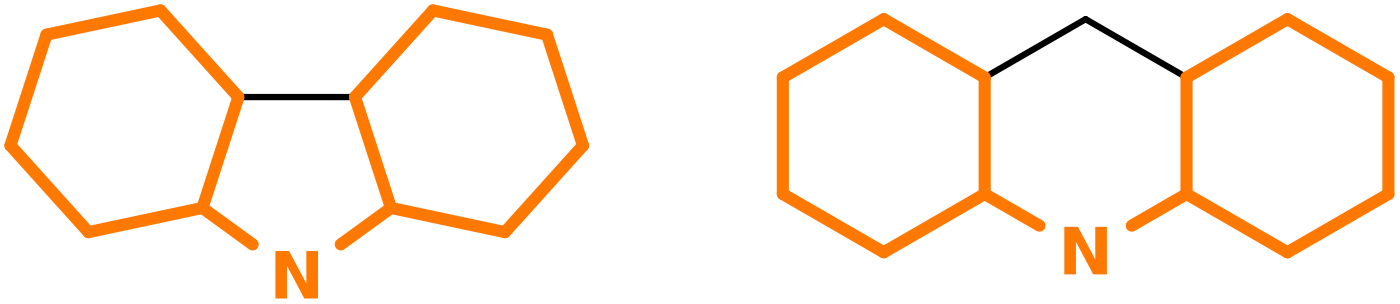
3D illesztés



MCIS – MCES



MCIS = *Maximum Common Induced Subgraph*



MCES = *Maximum Common Edge Subgraph*

MCIS – MCES

Maximum Common Induced Subgraph

MCIS

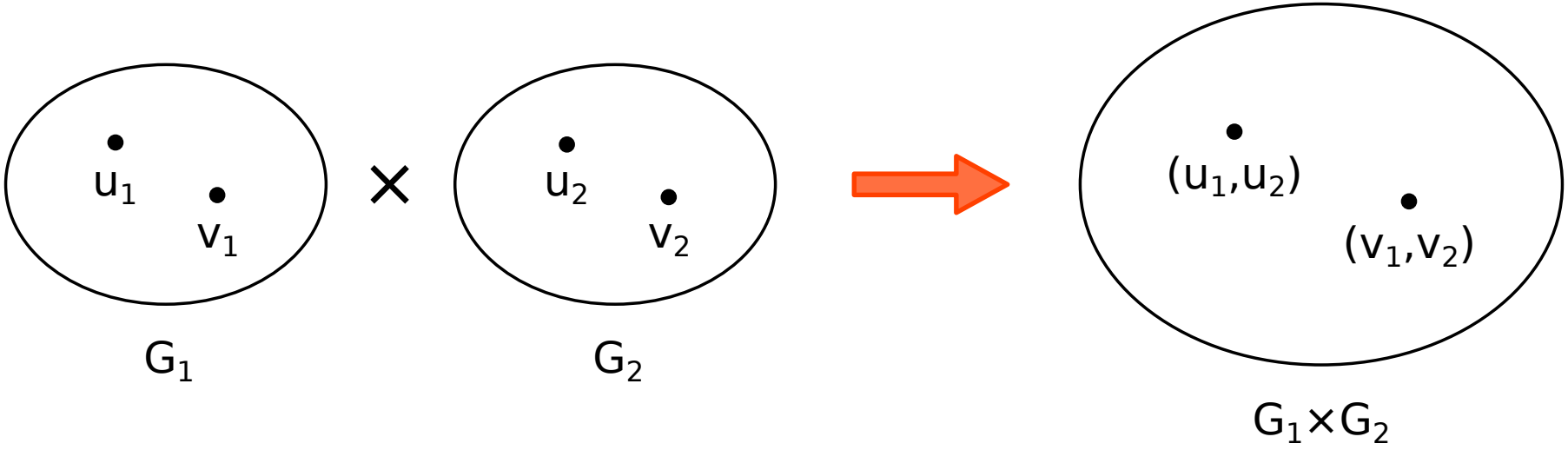
- a közös részgráfnak feszítettnek kell lennie
- *csúcsok* közötti leképezéssel definiálható
- méret = *csúcsok* száma

Maximum Common Edge Subgraph

MCES

- a közös részgráf nem feltétlenül feszített
- *élek* közötti leképezéssel definiálható
- méret = *élek* száma

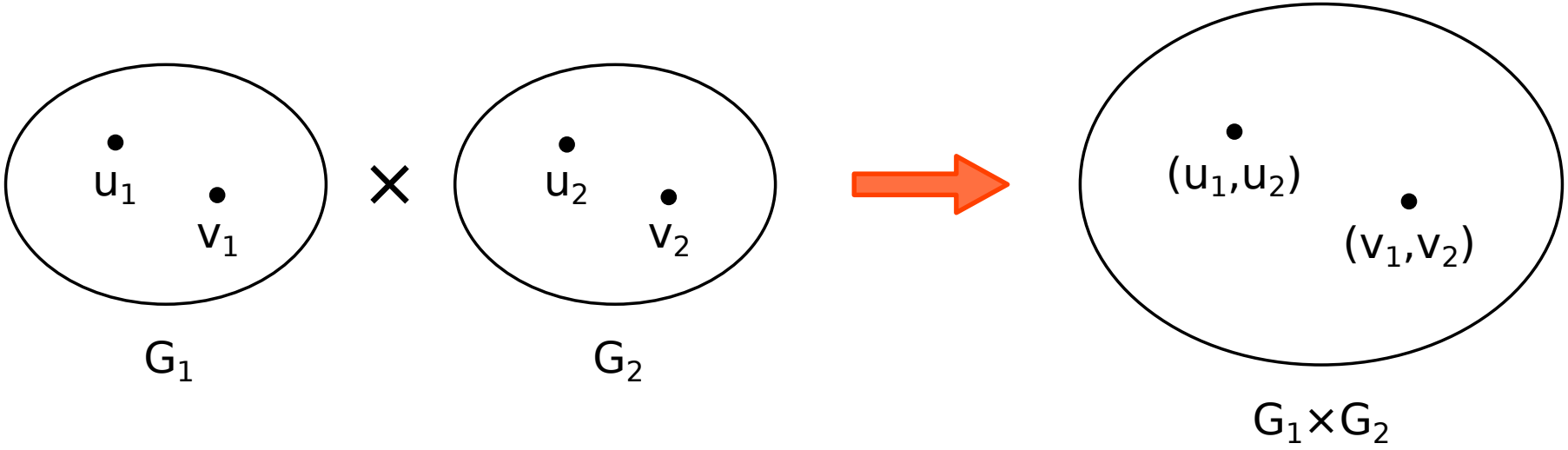
Visszavezetés a maximális klikk feladatra



Direktszorzat-gráf (MCIS feladat)

- csúcsai a G_1 és G_2 azonos címkéjű *csúcsaiból* alkotott párok
- élei ezek kompatibilitását fejezik ki

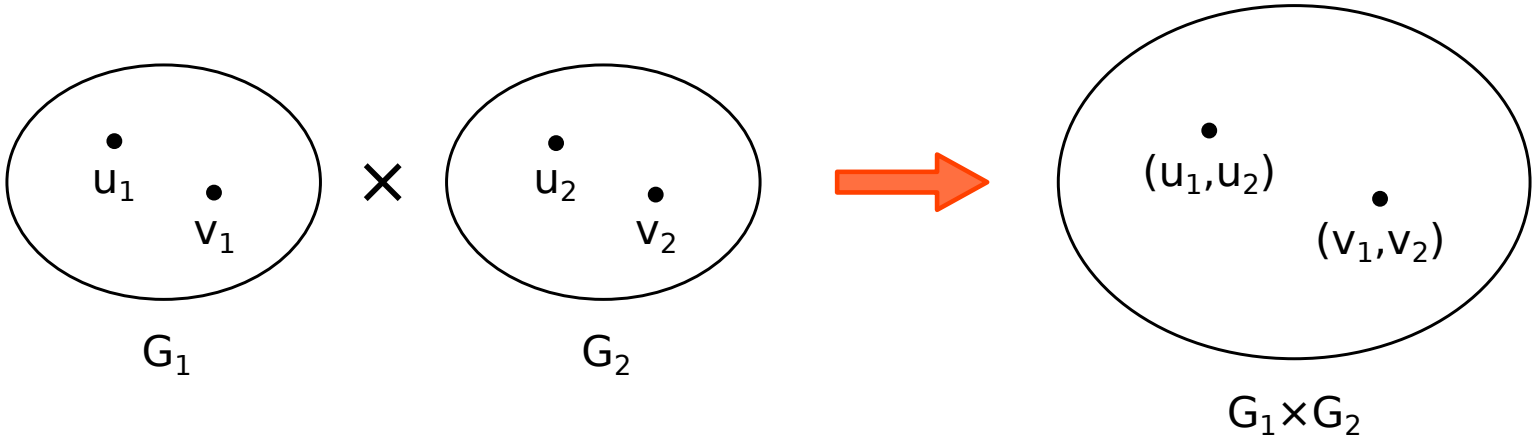
Visszavezetés a maximális klikk feladatra



Direktszorzat-gráf (MCIS feladat)

- csúcsai a G_1 és G_2 azonos címkéjű *csúcsaiból* alkotott párok
- élei ezek kompatibilitását fejezik ki
- (u_1, u_2) és (v_1, v_2) kompatibilis $\iff u_1 \neq v_1$, $u_2 \neq v_2$, valamint
  - u_1 nem szomszédos v_1-gyel és u_2 nem szomszédos v_2-vel
  - vagy u_1-v_1 és u_2-v_2 is szomszédosak és a közöttük lévő élek címkéje azonos G_1-ben és G_2-ben

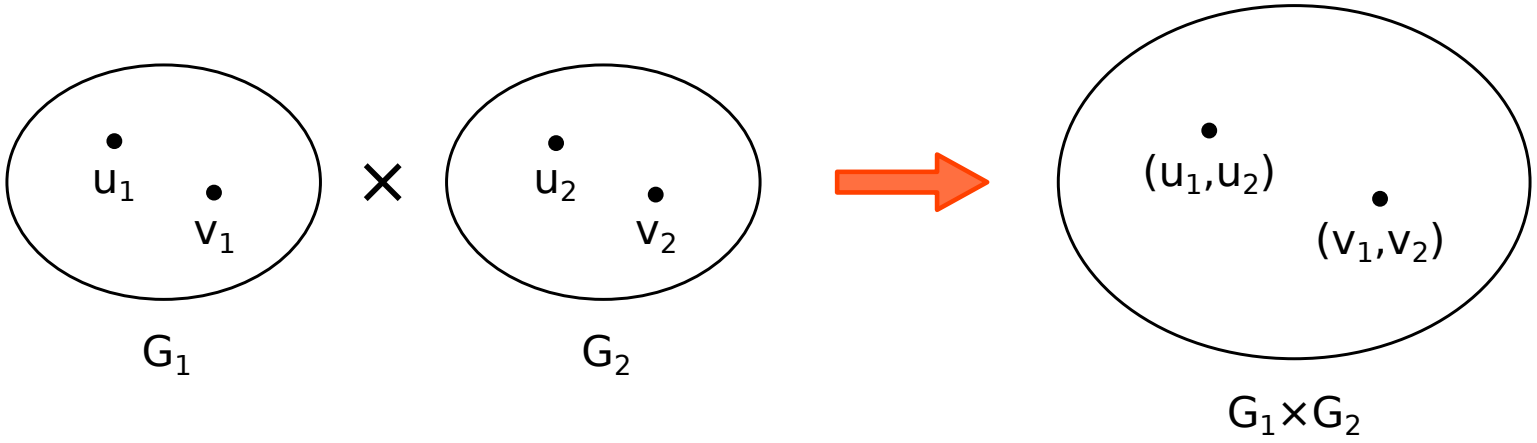
Visszavezetés a maximális klikk feladatra



Direktszorzat-gráf (MCIS feladat)

- (u_1, u_2) és (v_1, v_2) kompatibilis = egy feszített közös részgráfot meghatározó $V(G_1) \to V(G_2)$ leképezés egyszerre tartalmazhatja őket

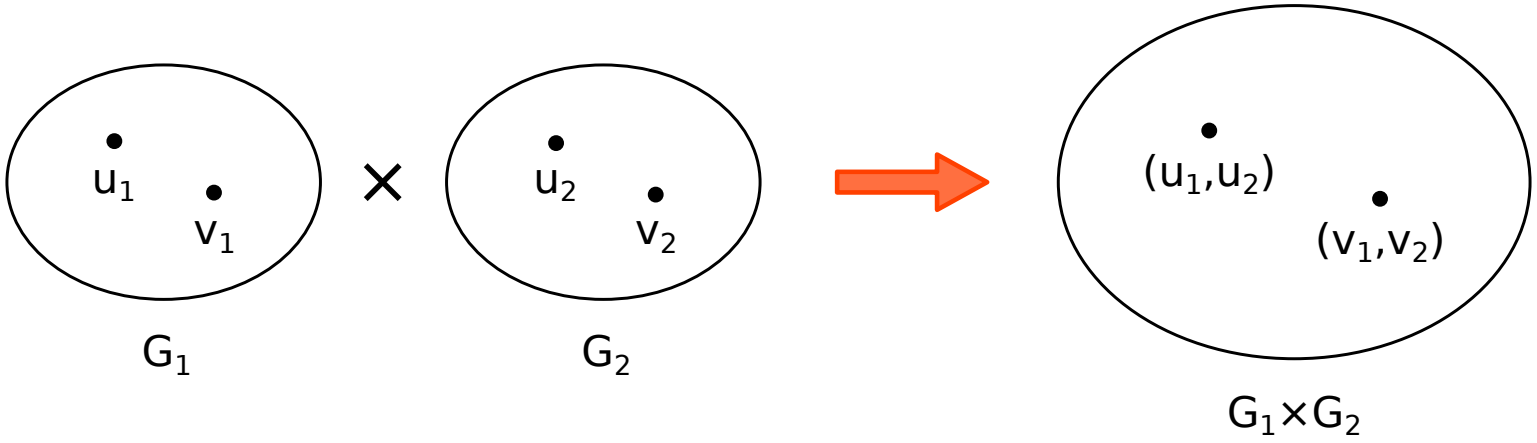
Visszavezetés a maximális klikk feladatra



Direktszorzat-gráf (MCIS feladat)

- (u_1, u_2) és (v_1, v_2) kompatibilis = egy feszített közös részgráfot meghatározó $V(G_1) \to V(G_2)$ leképezés egyszerre tartalmazhatja őket
- **következmény:** $G_1 \times G_2$ klikkjei és a feszített közös részgráfokat meghatározó leképezések kölcsönösen egyértelműen megfeleltethetők egymásnak (és a méretük is azonos)

Visszavezetés a maximális klikk feladatra



Direktszorzat-gráf (MCIS feladat)

- (u_1, u_2) és (v_1, v_2) kompatibilis = egy feszített közös részgráfot meghatározó $V(G_1) \to V(G_2)$ leképezés egyszerre tartalmazhatja őket
- **következmény:** $G_1 \times G_2$ klikkjei és a feszített közös részgráfokat meghatározó leképezések kölcsönösen egyértelműen megfeleltethetők egymásnak (és a méretük is azonos)
- tehát az MCIS feladat visszavezethető a $G_1 \times G_2$ gráfban *maximális klikk* keresésére

Visszavezetés a maximális klikk feladatra

Direktszorzat-gráf (MCES feladat)

- G_1 és G_2 helyett tekintsük az élgráfjaikat: $L(G_1)$ és $L(G_2)$
- $L(G_1) \times L(G_2)$ az élpárok kompatibilitását reprezentálja

Visszavezetés a maximális klikk feladatra

Direktszorzat-gráf (MCES feladat)

- G_1 és G_2 helyett tekintsük az élgráfjaikat: $L(G_1)$ és $L(G_2)$
- $L(G_1) \times L(G_2)$ az élpárok kompatibilitását reprezentálja
- G_1 és G_2 egy izolált csúcsot nem tartalmazó közös részgráfja (nem felt. feszített) megfeleltethető az $L(G_1)$ és $L(G_2)$ egy közös feszített részgráfjának
- fordítva is igaz, feltéve hogy nem történt ΔY csere (l. később)

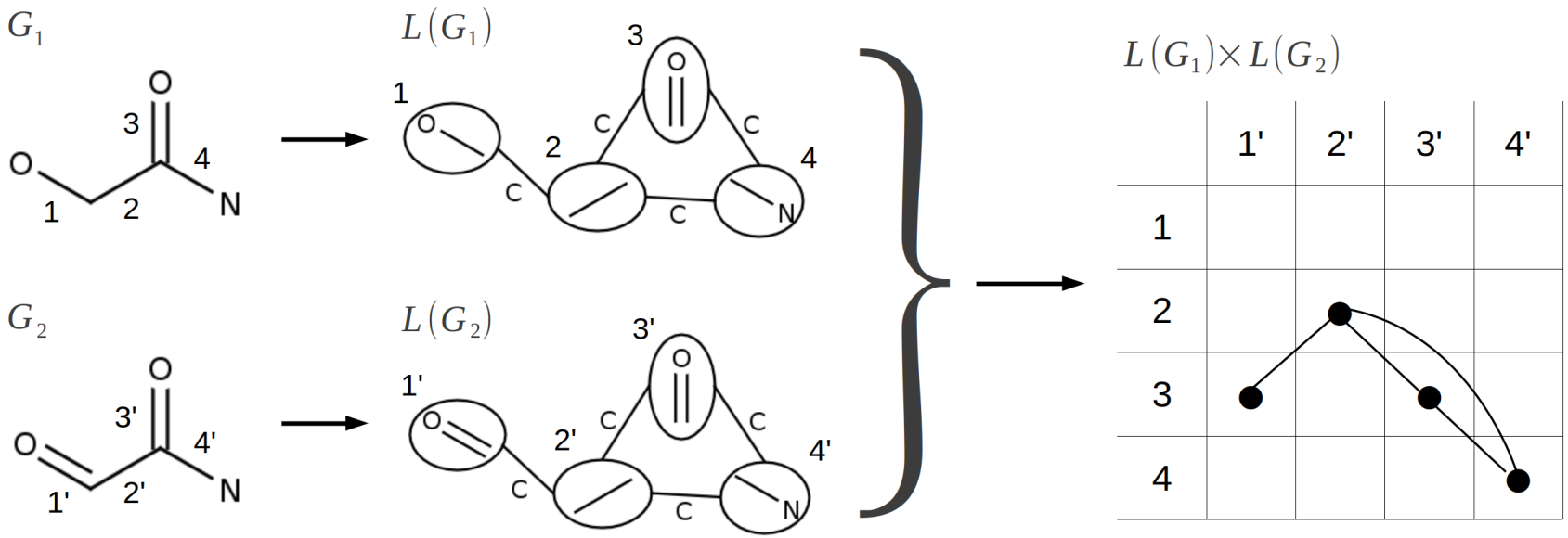
Visszavezetés a maximális klikk feladatra

Direktszorzat-gráf (MCES feladat)

- G_1 és G_2 helyett tekintsük az élgráfjaikat: $L(G_1)$ és $L(G_2)$
- $L(G_1) \times L(G_2)$ az élpárok kompatibilitását reprezentálja
- G_1 és G_2 egy izolált csúcsot nem tartalmazó közös részgráfja (nem felt. feszített) megfeleltethető az $L(G_1)$ és $L(G_2)$ egy közös feszített részgráfjának
- fordítva is igaz, feltéve hogy nem történt ΔY csere (l. később)
- **következmény:** az MCES feladat visszavezethető az $L(G_1)$ és $L(G_2)$ gráfokon értelmezett MCIS feladatra, tehát az $L(G_1) \times L(G_2)$ gráfban *maximális klikk* keresésére

Visszavezetés a maximális klikk feladatra

Direktszorzat-gráf előállítása (MCES)



Visszavezetés a maximális klikk feladatra

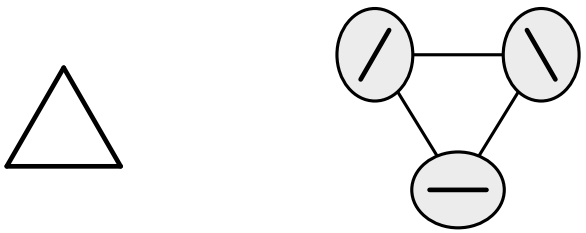
ΔY csere

- ha G_1 és G_2 izomorfak, akkor nyilván az élgráfjaik is
- és megfordítva ?

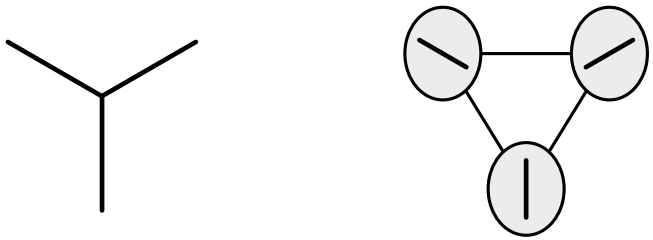
Visszavezetés a maximális klikk feladatra

ΔY csere

- ha G_1 és G_2 izomorfak, akkor nyilván az élgráfjaik is
- és megfordítva?
- **Tétel (Whitney, 1932):** Két összefüggő gráf akkor és csak akkor izomorf, ha az élgráfjaik izomorfak, egyetlen kivételtől eltekintve: K_3 és $K_{1,3}$, amelyek nem izomorfak, de az élgráfjaik izomorfak.
- szerencsére molekulagráfokban a K_3 (háromszög vagy Δ) részgráf elég ritka



K_3 (Δ gráf) és az élgráfja



$K_{1,3}$ (Y gráf) és az élgráfja

A feladat bonyolultsága

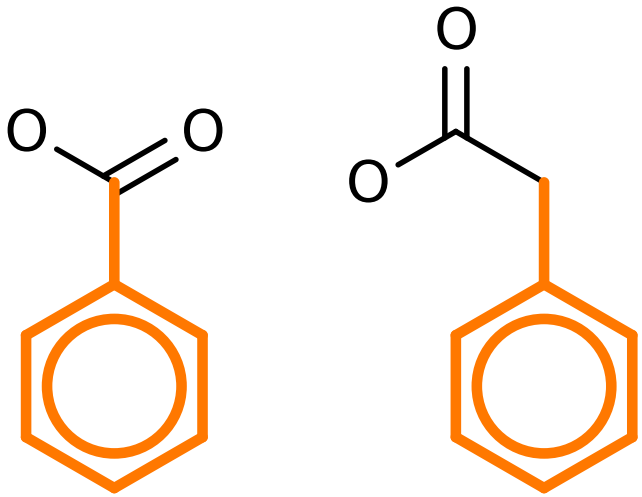
- az MCIS és MCES feladatok NP-nehezek
- a megfelelő döntési problémák NP-teljesek: létezik-e k méretű közös (feszített) részgráf

A feladat bonyolultsága

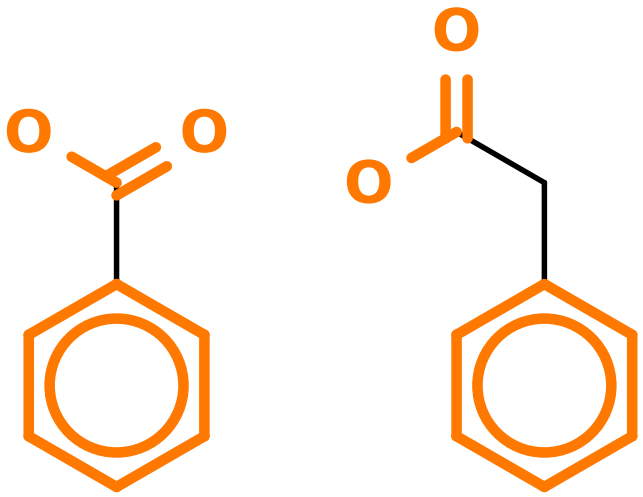
- az MCIS és MCES feladatok NP-nehezek
- a megfelelő döntési problémák NP-teljesek: létezik-e k méretű közös (feszített) részgráf

- MCES feladat speciális esetei:
  - részgráf-izomorfia (NP-teljes)
  - maximális klikk (NP-teljes)
  - Hamilton-kör (NP-teljes, még max. 3 fokú síkgráfokra is)
- a max. klikk probléma az MCIS feladatnak is spec. esete

Összefüggőség



Összefüggő MCES

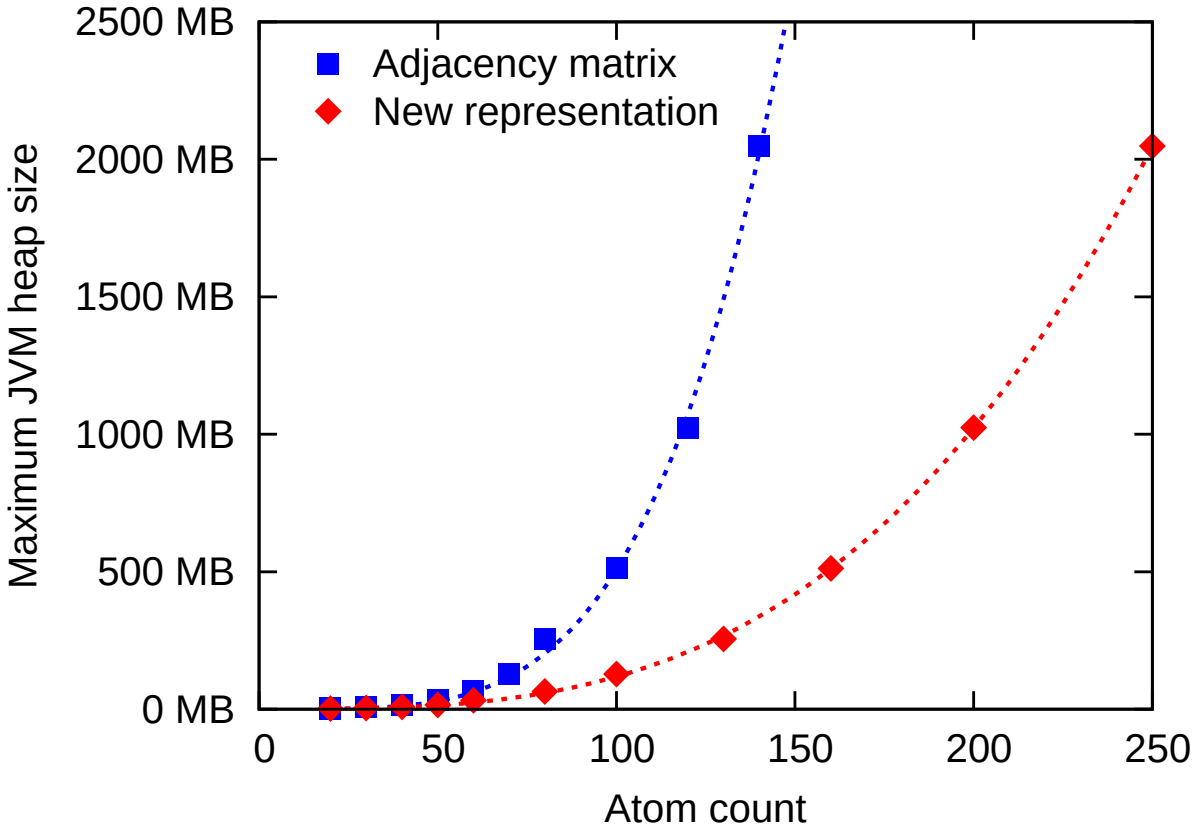


Nem összefüggő MCES

Algoritmus

- MCES feladatot visszavezetjük a max. klikk feladatra ($L(G_1) \times L(G_2)$ gráfban)
- max. klikk keresésére egy hatékony heurisztikus algoritmust alkalmazunk
- kidolgoztunk további heurisztikákat, amelyekkel a módszert pontosabbá és gyorsabbá tettük
- ezekben kihasználjuk a molekulagráfok specialitásait (csúcs- és élcímkék, alacsony fokszám stb.)

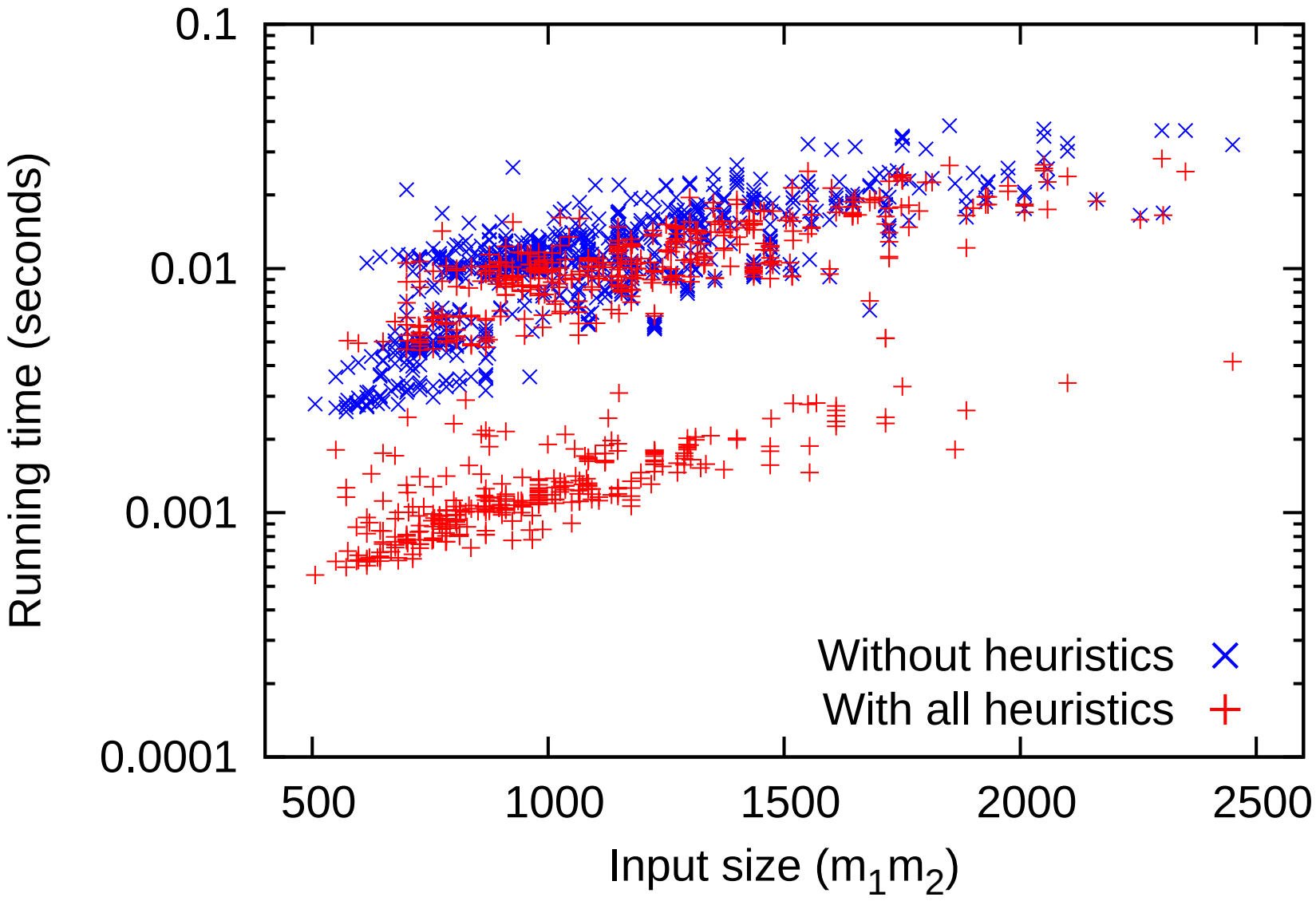
Reprezentáció



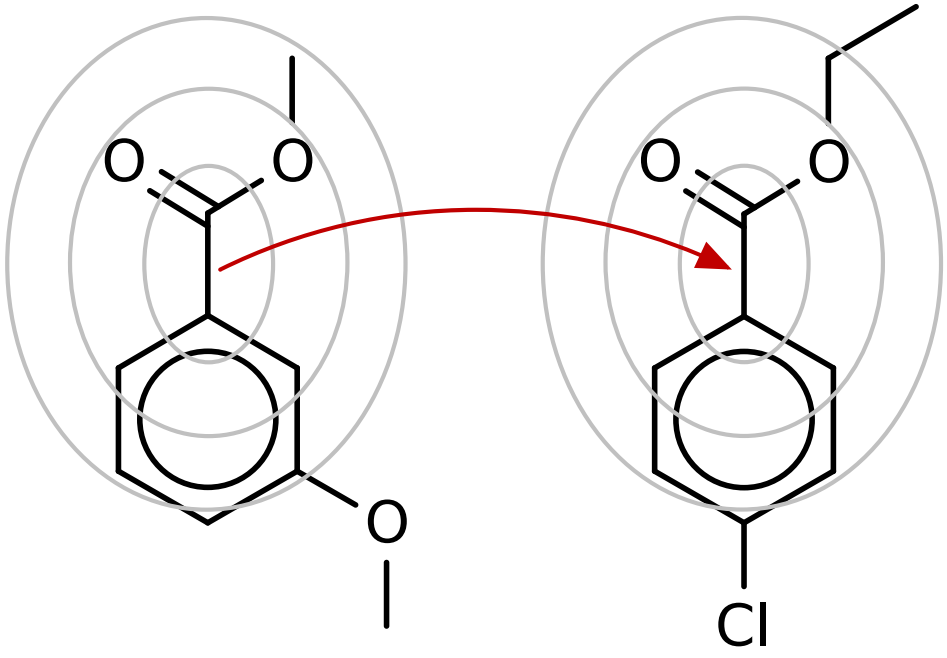
Csúcsmátrixos ábrázolás: $3{,}02\ n^{4{,}11}$

Komplementer gráf tárolása éllistával: $71{,}42\ n^{3{,}11}$

Korai terminálás



ECFP heurisztika

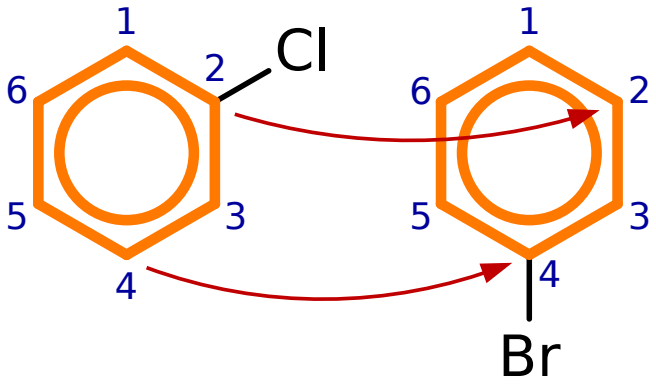


- az ún. extended connectivity fingerprint (ECFP) generálási algoritmusán alapuló heurisztika
- a gráfok csúcsainak környezeteit hash-kódokkal reprezentáljuk
- előnyben részesítjük olyan csúcsok és élek egymáshoz rendelését, amelyek nagy izomorf környezettel rendelkeznek

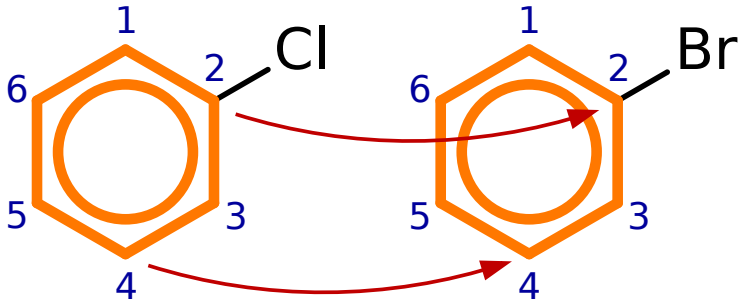
Leképezés optimalizálása

- a közös részgráf méretén kívül gyakran fontos az inputgráfok csúcsait és éleit egymáshoz rendelő leképezés is (pl. reakciók elemzése, molekulák illesztése)
- leképezés szempontjából jelentős csúcsokat azonosítjuk
- a leképezések összehasonlításához egy heurisztikus értékelő függvényt használunk

Hozzárendelés javítása

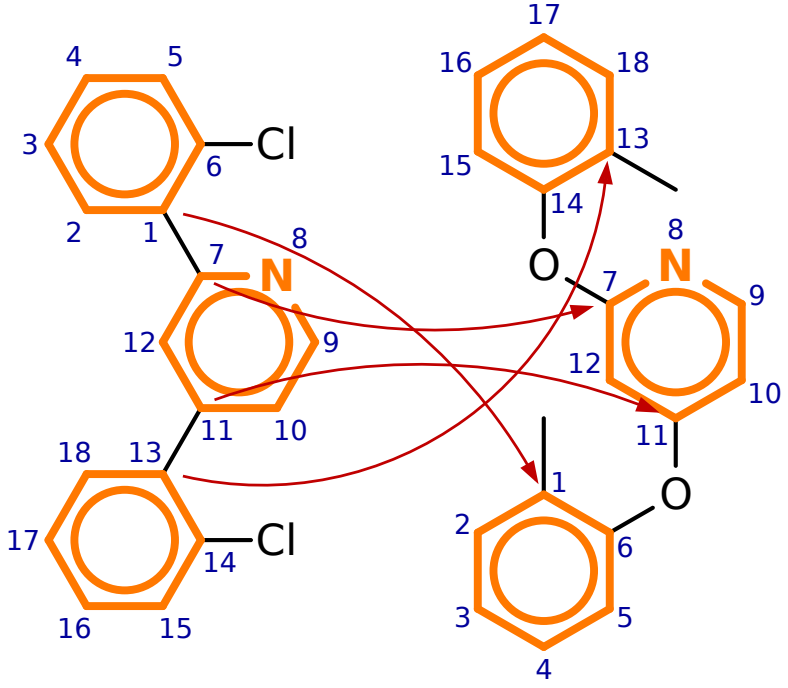


Szuboptimális leképezés

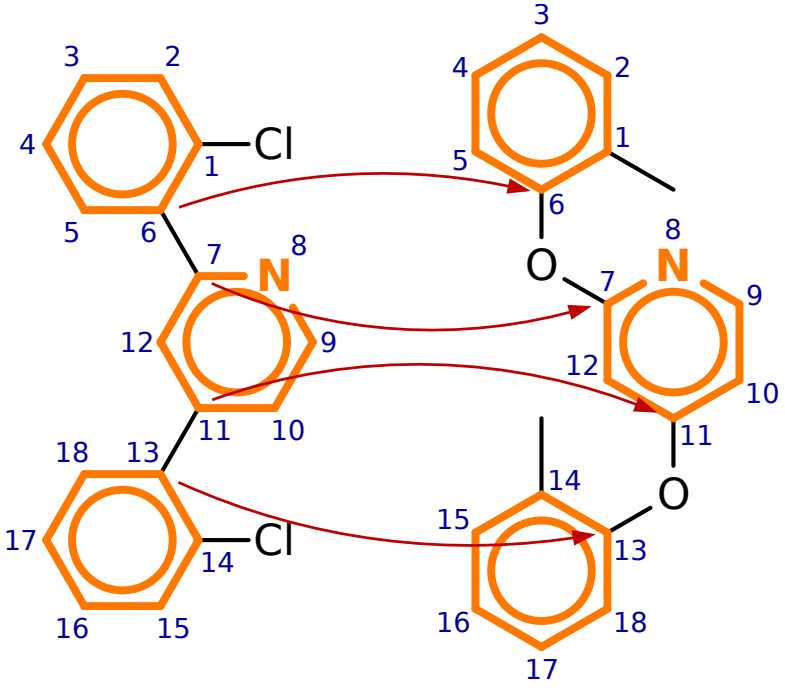


Elvárt leképezés

Hozzárendelés javítása



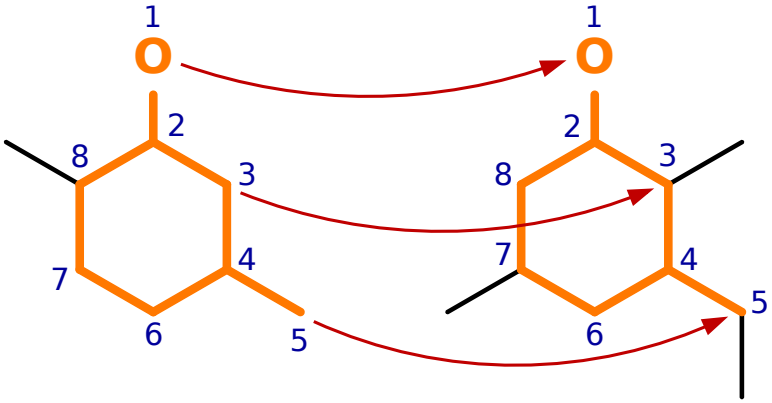
Szuboptimális leképezés



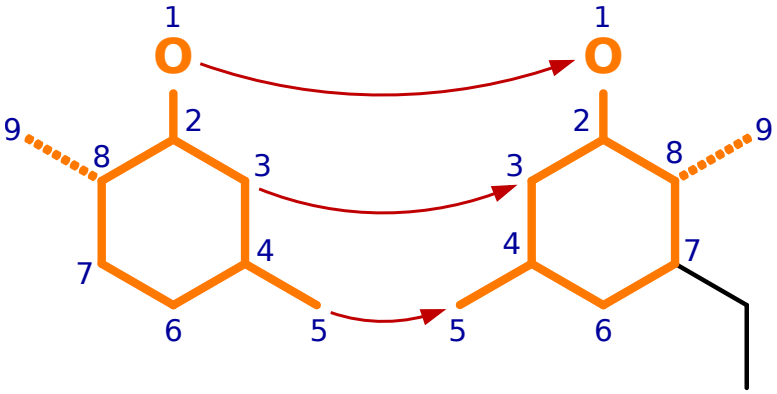
Elvárt leképezés

Hozzárendelés javítása

Ez a heurisztika gyakran segít abban is, hogy nagyobb közös részgráfot találjunk

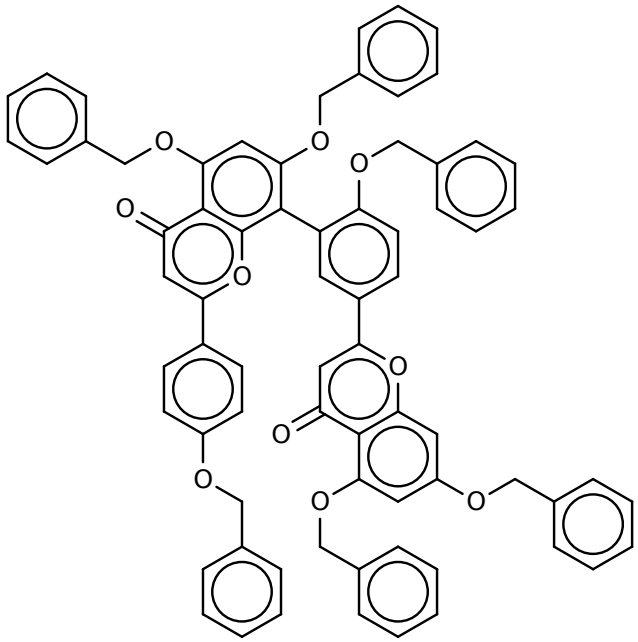


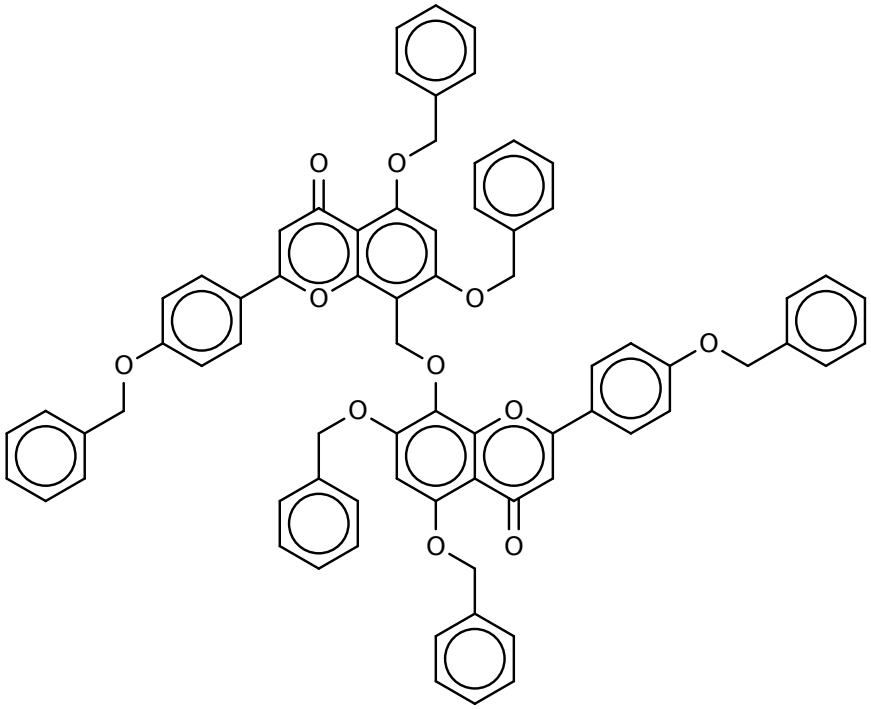
Szuboptimális eredmény



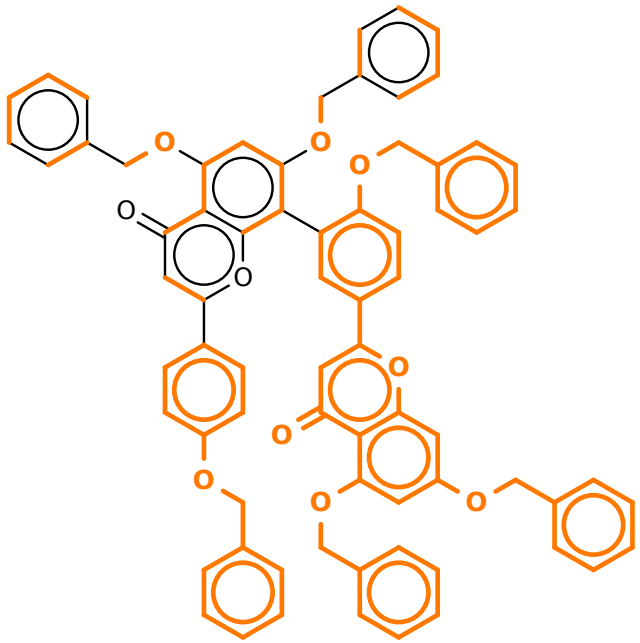
Optimális eredmény

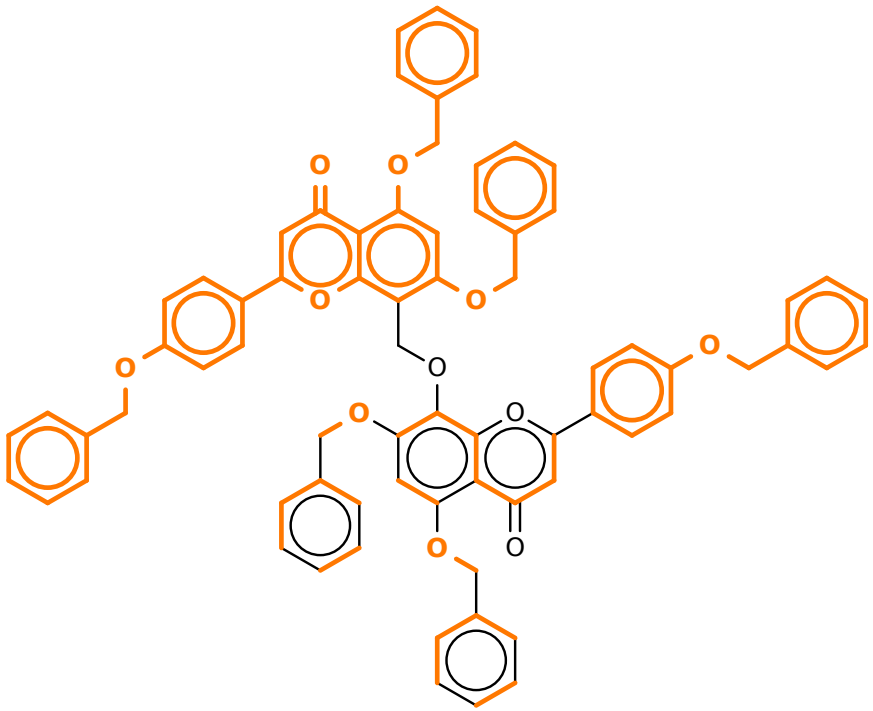
Eredmények – 1. példa





Eredmények – 1. példa

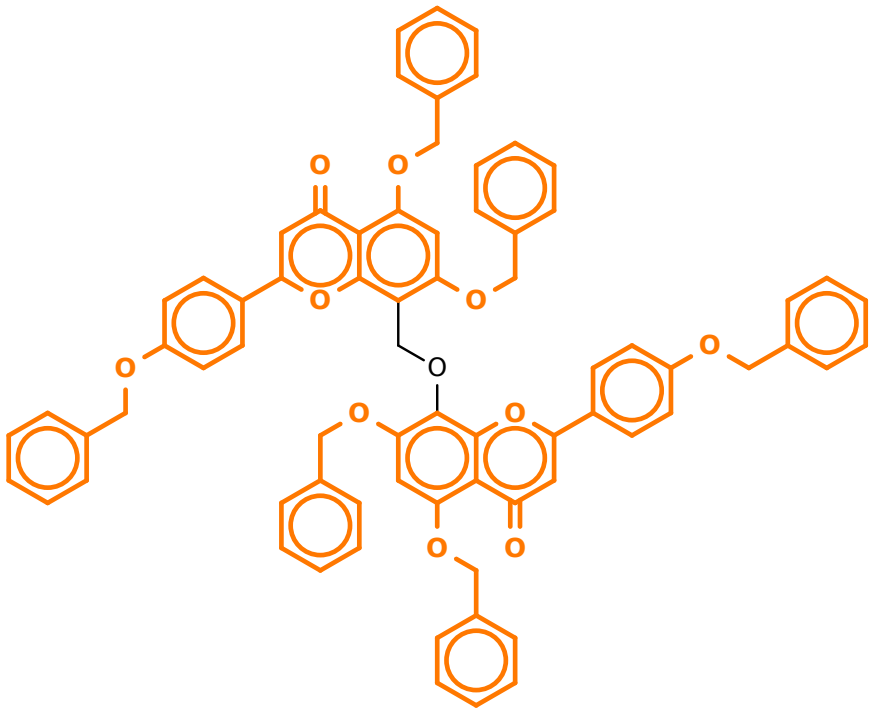




Heurisztikák nélkül:

Atomok száma: **80** Kötések száma: **78** Komponensek száma: **10** Hasonlóság: **0,71**

Eredmények – 1. példa



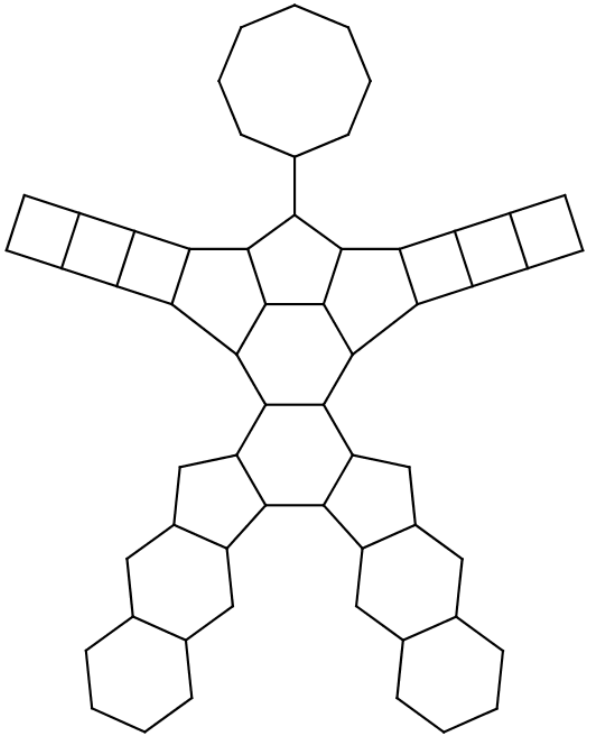
Heurisztikák nélkül:

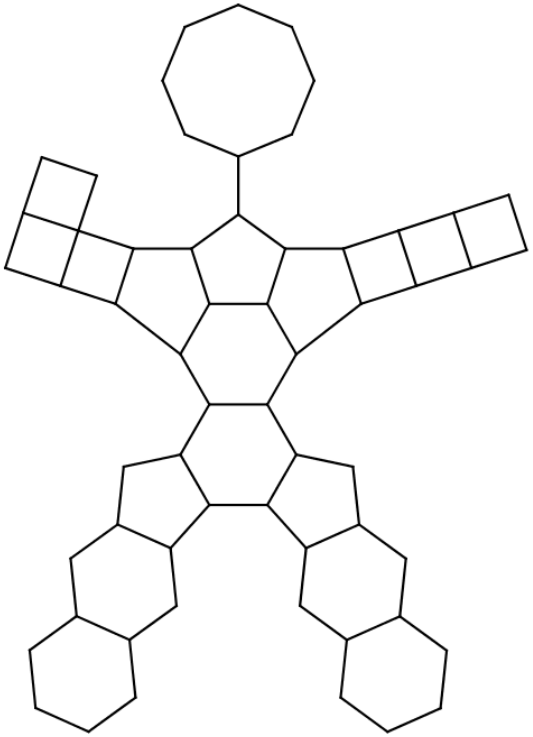
Atomok száma: **80** Kötések száma: **78** Komponensek száma: **10** Hasonlóság: **0,71**

Heurisztikákkal:

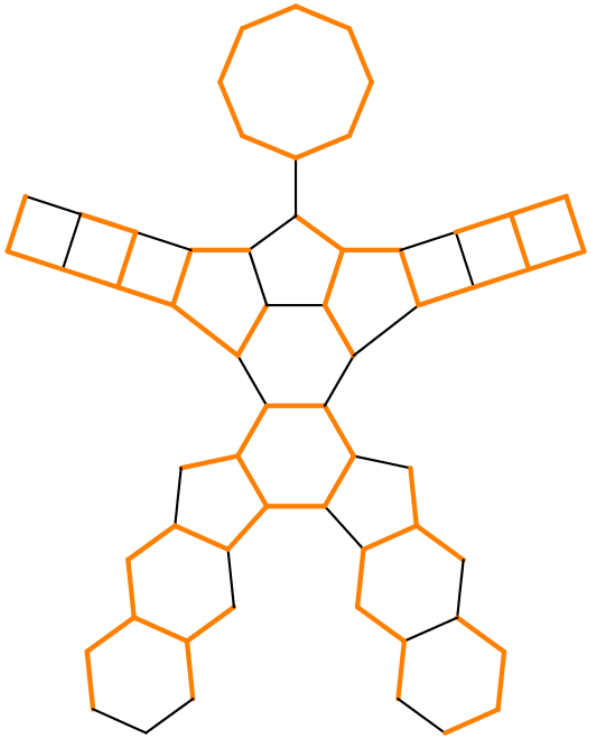
Atomok száma: **82** Kötések száma: **92** Komponensek száma: **2** Hasonlóság: **0,96**

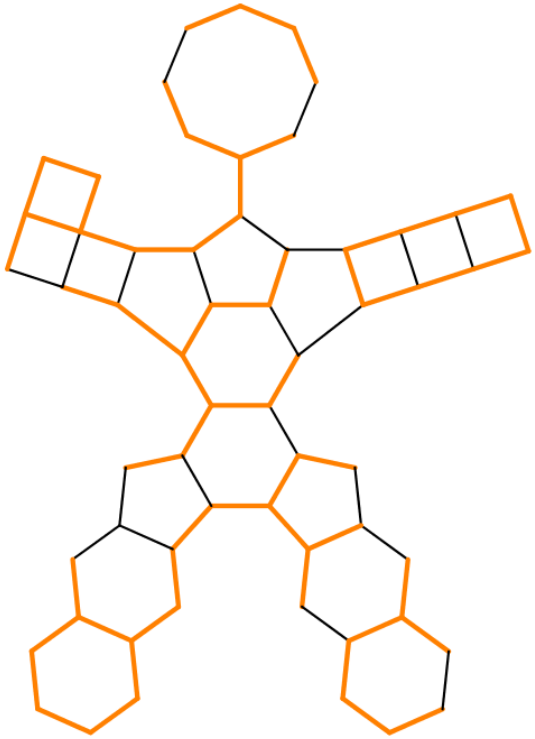
Eredmények – 2. példa





Eredmények – 2. példa

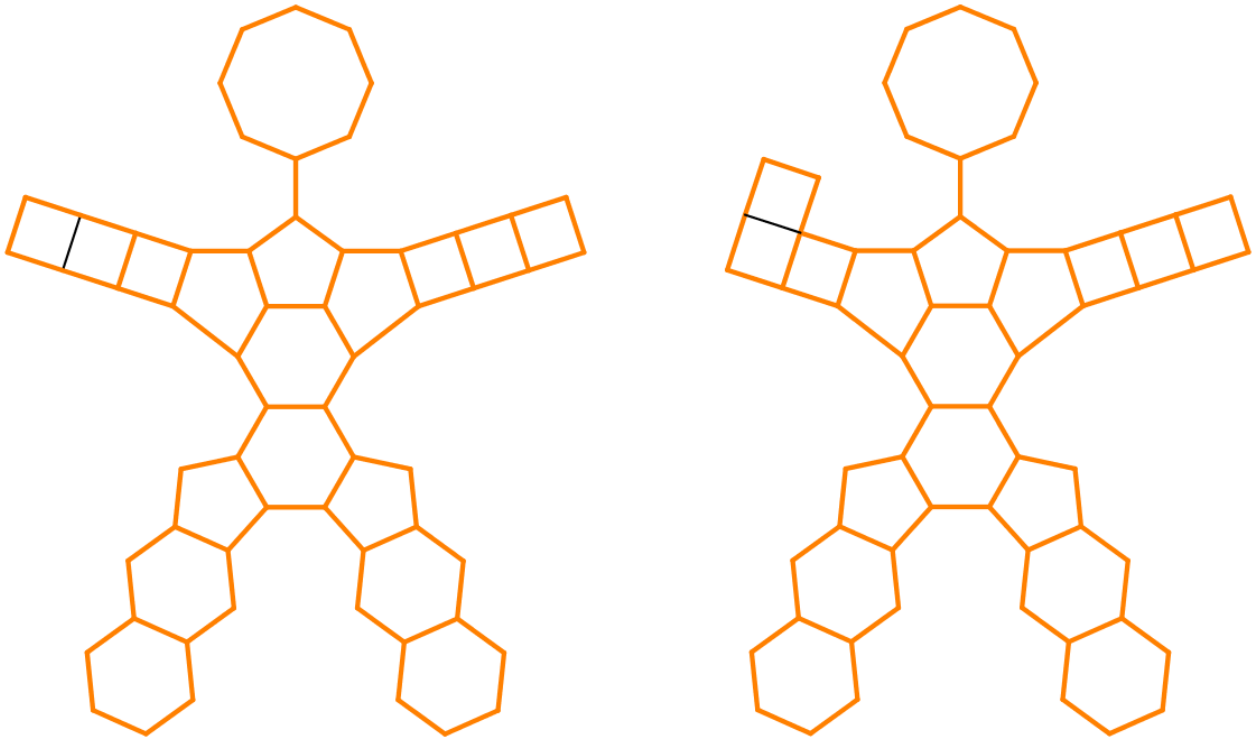




Heurisztikák nélkül:

Atomok száma: **58** Kötések száma: **55** Komponensek száma: **6** Hasonlóság: **0,57**

Eredmények – 2. példa



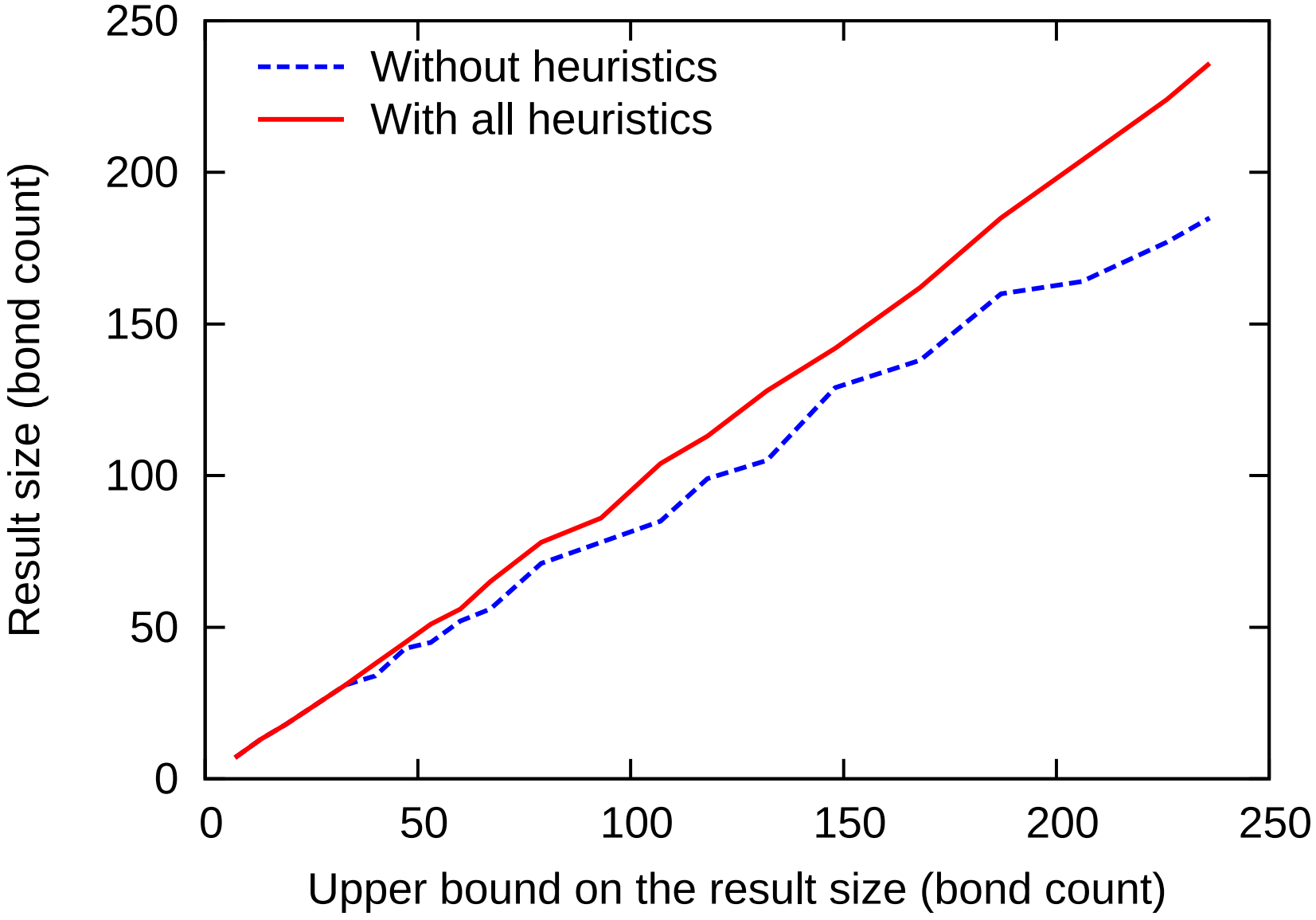
Heurisztikák nélkül:

Atomok száma: **58** Kötések száma: **55** Komponensek száma: **6** Hasonlóság: **0,57**

Heurisztikákkal:

Atomok száma: **59** Kötések száma: **75** Komponensek száma: **1** Hasonlóság: **0,97**

Eredmények


Without heuristics
With all heuristics
Result size (bond count)
Upper bound on the result size (bond count)
0
50
100
150
200
250
0
50
100
150
200
250