

## **k-hierarchiák alkalmazása a távolság alapú filogenetikus farekonstrukcióban**

*Hegedűs István*

*V. programtervező matematikus*

*Témavezetők: Bagyinka Csaba (MTA-SZBK),  
Busa-Fekete Róbert (MTA-SZTE RGAI)*

A molekuláris evolúció vizsgálatában az úgynevezett filogenetikus fák kiemelkedő jelentőségűek, segítségükkel az evolúciós modellnek megfelelően leírhatóak a rokonsági kapcsolatok a különböző biológiai objektumok (pl. fehérjék) között. A fák által reprezentált viszonyok alapján következtetni tudunk például fehérjék tulajdonságaira és térszerkezetére, ami segítséget nyújthat különböző kémiai és gyógyszeripari kísérletek tervezésénél is. A dolgozatban a jelenleg elterjedt farekonstrukciós módszerek közül mi a leginkább alkalmazott, távolság alapú eljárásokat vizsgáljuk, ahol ismertnek tekintjük a biológiai objektumok távolságait, valamint javasolunk egy újabb faépítő algoritmust.

A dolgozat első része a filogenetikus faépítés elméleti hátterébe nyújt bevezetést filogenetikus fa fogalmának bevezetése, az evolúciós modellek és a tesztelésnél használt algoritmusok bemutatása). A második részben pedig az általunk javasolt távolság alapú farekonstrukciós algoritmust mutatjuk be. Az algoritmusunk két fő lépésből áll. Az első lépésben a vizsgált biológiai objektumokból részhalmazokat állítunk elő, kezdve az egy elemű részhalmazokkal. Majd ezeket bővítjük iteratívan egy-egy elemmel. A megfelelő bővítő elem kiválasztásánál a keletkező részhalmazok minőségét a k-hierarchiák segítségével megbecsüljük, és a minimális k-hoz tartozó halmaznak megfelelően bővítjük. Ezt követően minden egyes részhalmazhoz rendelünk egy valós számot, amelyet a részhalmazon belüli elemek és a részhalmazon kívüli elemek távolságaiból származtatunk. Ezt az úgynevezett szeparációs indexet használjuk fel a fa építésekor. A második lépésben pedig az előbbieken megkapott részhalmazok alapján konstruálunk egy úgynevezett kompatibilitási gráfot, amelyben a pontok reprezentálják a klasztereket, a pontokhoz rendelt súly pedig a szeparációs index értékével egyezik meg. Két pont között akkor van él, ha a gráfpontokhoz tartozó részhalmazok kompatibilisek egymással. Ezen gráf minden egyes független csúcshalmaza meghatároz egy hierarchikus halmazrendszert, amely megfeleltethető egy gyökeres filogenetikus fának. A minimális összsúlyú független csúcshalmazhoz tartozó fát tekintjük hipotézisünk szerint a legpontosabb filogenetikus fának.

Az itt javasolt módszert egyrészt mesterségesen előállított, másrészt valós szekvenciák halmazán teszteltük. Eredményeinket három, elterjedt és széles körben használt algoritmussal (Neighbor Joining, Single Linkage, fastME) hasonlítottuk össze. A tesztek alapján az itt javasolt algoritmus hatékonysága és pontossága nemcsak felveszi a versenyt a meglévő módszerekkel, hanem sok esetben pontosabb eredményt is szolgáltat.