Szegedi Tudományegyetem
Mesterséges Intelligencia Kutatócsoport

Evolutionary Tree Reconstruction and its Applications in Protein Classification

PhD értekezés tézisei

Busa-Fekete Róbert

Témavezetők:

Prof. Csirik János, Dr. Kocsor András

Szeged
2008

Bevezetés

Az összefoglaló ismerteti a „Evolutionary Tree Reconstruction and its Applications in Protein Classification" című PhD disszertáció eredményeit. A disszertáció témáját tágabb értelemben a mesterséges intelligencia és a bioinformatika, szorosabb értelemben pedig a gépi tanulás és az evolúciós fák rekonstrukciója képezi.

Az evolúció már több mint egy évszázada a fajok kialakulásának a legelfogadottabb modellje. A törzsfejlődés ezen modellje elsősorban a fajok rokonsági fokát probálja meghatározni. A filogenetika (a szó a görög phülon = törzs és geneszisz = születés szavakból ered) a fajokat, élőlényeket rendszerezi evolúciós rokonsági fokuk alapján. A filogenetikában a legelterjedtebb módszerek a fajok fejlődésének a folyamatát egy úgynevezett filogenetikus fával reprezentálják, amely egy súlyozott fa-gráfnak felel meg, ahol a levelek reprezentálják a vizsgált biológiai objektumokat. Az ilyen típusú fák rekonstrukciója mind biológiai, mind számítástudományi szempontból számos érdekes problémát vet fel.

A különböző fajokból izolált fehérjék szekvenciáinak összehasonlítási lehetősége új típusú vizsgálatok elvégzésére adott alapot a filogenetikában. Ez merőben átformálta a biológia ezen ágát. Míg korábban a filogenetika egyet jelentett a fajok evolúciós fejlődéstanával, addig az új eredmények hatására a kutatások kiterjedtek a fehérjék öröklődésének vizsgálatára. Fehérjék azon csoportját, melyek szekvenciái nagyon hasonlóak egymással, rokon fehérjéknek tekintik, vagy más szóval homológ csoportnak hívjuk. A homológ csoportok általában hasonló funkciókkalrendelkeznek az élő szervezetben. A filogenetika egyik fontos alapfeladatának tekintjük a különböző fajokból izolált, hasonló funkciójú és hasonló szekvenciájú fehérjék vizsgálatát, és ezen fehérjecsoportok evolúciós történetének a meghatározását.

Mivel a disszertáció két fő részre tagolódik, az eredményeket is ennek megfelelően két csoportra fogjuk felosztani.

Az eredmények *első csoportját* filogenetikusfa-építő módszerek bemutatása képezi. A faépítő algoritmusok bemenete sokféle biológiai objektum lehet, úgy mint gén szekvenciák, fehérje szekvenciák vagy mitokondriális DNS szekvenciák egy halmaza. Kimenetük egy fa struktúra, melyben a levelek reprezentálják a vizsgált biológiai objektumokat. Számos faépítő algoritmust dolgoztak ki, amely közül néhány széles körben elterjedt, mint például a Neighbor-Joining [1] és az UPGMA [2]. Ezek a módszerek az úgynevezett távolság-alapú módszerek közé tartoznak, mert a vizsgált szekvenciák előre adott távolságai alapján rekonstruálják az evolúciós történetüket. Ezek a módszerek az evolúciós történetet általában egy úgynevezett súlyozott filogenetikus fa formájában reprezentálják. Az általunk kidolgozott távolságalapú Multi-Stack (MS) algoritmus [3] azt a súlyozott fatopológiát keresi, amely a legjobban képes visszaadni az előre definiált távolságot: azaz a keresett súlyozott fában a fehérjék távolságai –a közöttük lévő út élsúlyainak az összege– a legkevésbé térnek el az előre definiált távolságoktól. Mivel nem minden esetben létezik olyan súlyozott filogenetikus fa, amely által meghatározott távolságok az előre adott távolságokat teljes mértékben visszaadják, ezért arra törekszünk, hogy a kapott fa topologiája a legjobban igazodjon a "távolságviszonyokhoz". Ennek a fának a megtalálása egy NP-teljes problémára vezet [4], ezért csak heurisztikus megoldást lehet rá adni. Az MS módszer először a vizsgált fehérjék egy-egy részhalmazára épít optimális fát, majd ezeket a részfákat iteratívan összekapcsolja. Ezt a bottom-up megközelítést hatékonyan tudtuk alkalmazni több tesztkörnyezetben, és számos tradicionális faépítőnél jobbnak bizonyult.

Mivel a filogenetikusfa-építő algoritmusok sokszor több lehetséges evolúciós történetet is képesek meghatározni vagy a különböző algoritmusok különböző fát rekonstruálnak, ezért sokszor olyan mód-

szerre van szükségünk a filogenetikus analízis utolsó fázisaként, amely több filogenetikus fa által hordozott információt képes egyetlen reprezentatív fába összegyűjteni [5]. Az ilyen célú algoritmusokat konszenzusfa-építőknek nevezzük. Általában minden gyökeres filogenetikus fa egy belső pontja egyértelműen meghatározza a vizsgált biológiai objektumoknak egy részhalmazát (a belső pont alatt található levelek által reprezentált objektumok halmaza). Tehát a filogenetikus fa ekvivalens a hierarchikus halmazrendszerek vagy más szóval a kompatibilis halmazok konstrukciójával. Ezt a megközelítést alkalmazva, kézenfekvő, hogy azokat a kompatibilis részhalmazokat szeretnénk a konszenzusfa belső pontjaiként kiválasztani, amelyek a vizsgált fákban a legtöbbször fordulnak elő. Természetesen az input fákban előforduló részhalmazokon értelmezhetünk tetszőleges valós értékű súlyfüggvényt, amely nem csupán előforduláson alapszik, hanem az input fák más tulajdonságait is figyelembe veszi. Ezt a konszenzusfa-építési problémát oldottuk meg hatékonyan [6], és megmutattuk, hogy egy alkalmas részhalmaz súlyozással a legelterjedtebb konszenzus módszereknél (mint például Majority-Rule, Strict vagy Greedy konszenzus [7]) pontosabb filogenetikus analízist lehet végrehajtani.

A tézisek *második csoportját* a faépítő módszerek egy alkalmazása képezi. A fehérje-osztályozás az egyik legfontosabb feladat a mai biológiában. Egy-egy szervezet génjeinek adatait szekvenciák – gének által kódolt fehérjéket jelképező néhány száz karakter hosszú sorozatok– formájában tárolják. Mára mindennapi rutinná vált, hogy ezeket az adatokat a közelítő mintaillesztés módszerével összehasonlítják a már ismert fehérjék hasonló adataival, majd valamely osztályozási eljárással megkísérlik besorolni őket a már ismert (szerkezeti, funkciós stb.) kategóriák valamelyikébe [8]. A gyakran emlegetett genom-kutatások automatikus adat-annotációs rendszerei lényegében erre a módszerre épülnek.

Munkáinkban a fehérje-osztályozás újszerű módszereit fejlesztettük ki, melyekben filogentikus információt is használtunk. Alapfeltételezésünk az, hogy a szekvencia adathalmazok belső szerkezete filogenetikus fa formájában ábrázolható, és hogy ennek révén az osztályozás hatékonnyá tehető [9; 10]. Módszereinkben az ismert és ismeretlen osztállyal rendelkező szekvenciákra megkonstruálunk egy filogenetikus fát csupán a szekvenciák hasonlósági viszonyai alapján. Majd a megkonstruált fából nyerünk ki olyan információt, amely hasznos az osztályozás szempontjából. Azok a fehérjeosztályozási módszerek, amelyekben filogenetikus információt is felhasználnak a filogenomika tárgykörébe tartoznak [11], ezért az általunk kifejlesztett módszerek is ide sorolhatóak.

I. Evolúciós következtetési módszerek

Filogenetikus fa definíciója

Fajok illetve biológiai objektumok (fajok, gének, fehérjék, genomok, stb.) evolúciós történetét általában egy fastruktúrával ábrázolják, ahol minden levél egy biológiai objektumot reprezentál. Azon biológiai objektumok halmazát, amely a filogenetikus analízis tárgyát képezik, a továbbiakban X fogja jelölni. A fa belső pontjai a hipotetikus ősöknek felelnek meg. Az általunk követett terminológiában a filogenetikus fa egy levélcímkézett fa.

1. Definíció. *Egy $\mathcal{T} = (T; \phi)$ rendezett párt* **filogenetikus fának** *nevezünk, ha T egy olyan fa (körmentes összefüggő gráf), melynek a $V(T)$ ponthalmaza legfeljebb egy olyan pontot tartalmaz, amelynek kettő a fokszáma, továbbá $\phi : X \to L(T)$ egy bijektív leképezés az X halmaz és a T fa $L(T)$ levélhalmaza között. A ϕ leképezést a $\mathcal{T}$ címkéző leképezésének nevezzük. Egy filogenetikus fát* **gyökeres vagy gyökereztetett filogenetikus fának** *nevezünk, ha rendelkezik egy $r \in V$ csúccsal, amelynek a fokszáma kettő. Az r pontot nevezzük a filogenetikus fa gyökérpontjának.*

2. Definíció. *Ha a $\mathcal{T}$ gyökeres filogenetikus fa minden nem-gyökér belső pontjának a foka pontosan három, akkor a $\mathcal{T}$ fát* **bináris filogenetikus fának** *nevezzük.*

3. Definíció. *Egy filogenetikus fát* **súlyozott filogenetikus fának** *nevezünk, ha a T fa élein értelmezve van egy nem negatív valós leképezés: $w : E(T) \to \mathbb{R}_{\geq 0}$.*

Egy fában bármely két levélpont között létezik út. Ha megköveteljük, hogy egy út minden élet legfeljebb egyszer tartalmazhat, akkor egyszerű útról beszélünk. Egy fában minden levélpárra pontosan egy olyan egyszerű út van, amely összeköti őket.

1. Következmény. *Egy $x, y \in X$ elempárra jelölje $p(x, y)$ az őket összekötő egyszerű utat az X feletti $\mathcal{T}$ súlyozott filogenetikus fában. A $\mathcal{T}$ fa w súlyfüggvénye meghatároz egy távolságfüggvényt az X halmazon:*

$$d_{\mathcal{T}}(x, y) = \sum_{e \in p(x,y)} w(e).$$

Az így értelmezett távolságot az X halmazon értelmezett fatávolságnak nevezzük.

Ezek ismeretében szeretnénk meghatározni azt az evolúciós történetet illetve filogenetikus fát, amely bizonyos kritérium alapján minél jobban képes visszaadni az evolúciós viszonyokat. A belső pontoknak több értelmezésük lehet, attól függően, hogy a biológiai objektumok, amelyeket vizsgálunk, milyen típusúak. Ha például gének egy rokon csoportjára építünk fát, akkor a belső pontokat úgynevezett evolúciós eseményeknek is értelmezhetjük (gén duplikációnak vagy gén specializációnak).

Meg kell jegyeznünk, hogy ha $RB(n)$ jelöli az összes gyökeres filognetikus fát az $|X| = n$ halmaz felett, akkor $|RB(n)| = (2n-3)!!$ [12]. Tehát az X halmaz méretétől függően a fatér mérete szuperexponenciálisan nő. Emiatt az olyan elemi megközelítés, mint például a kimerítő keresés technikája, már kis elemű adatbázisokon is nehezen alkalmazható.

Távolságalapú megközelítés

Számos megközelítés került kidolgozásra, amely filogenetikus fa konstruálását célozza meg. A filogenetikus módszerek fejlesztését a molekuláris biológia robbanásszerű fejlődése követeli meg. A faépítő módszereket durván három nagy csoportba lehet besorolni annak megfelelően, hogy milyen megközelítést követnek: távolság-, szekvencia- és kvartet-alapú faépítőket különböztetünk meg [1; 13; 14]. Ebben a disszertációban mi egy újszerű távolságalapú módszert mutatunk be.

A távolságalapú faépítésnél csupán a vizsgált biológiai objektumok közötti hasonlóságokból indulunk ki[12]. Ezért a filogenetikai analízis előtt szükségünk van egy távolságfüggvényre, hogy számszerűen definiáljuk az X halmazbeli objektumok páronkénti távolságát. Számos távolságfüggvény terjedt el a szakirodalomban, amely rögzített ábécé feletti karaktersorozatok között definiál távolságértéket[1].

Az evolúció során a szekvenciákban végbemenő változásokat mutációknak hívjuk. Ahhoz, hogy megállapítsuk, hogy két szekvencia milyen szoros kapcsolatban van egymással, először ezeket a változásokat kell azonosítanunk. Erre a legeltejedtebb módszer az általános illesztési modell, amely a bioinformatikában Needleman-Wunsch algoritmusként ismert [15]. Ez a módszer egy páronkénti illesztést végez a szekvenciákra. Ezen illesztés alapján meg tudjuk határozni, hogy két adott szekvencia hány pozicióban egyezik illetve tér el. Számos más alternatív illesztési módszer is kidolgozásra került, mint például a Smith-Waterman és a Gotoh algoritmusok [16; 17].

Az illesztési lépés után a különböző szekvenciák páronkénti evolúciós távolságát egy időfolytonos Markov Lánc (ML) segítségével határozzák meg [18; 19]. Ezek a modellek a páronkénti illesztésnél azonosított változások száma és változatossága alapján modellezik a szekvenciák evolúciós távolságát, és képesek modellezni azt az esetet, amikor egy pozícióban többször fordul elő mutáció. Például először egy $A \to C$ mutáció, majd egy $C \to A$ fordul elő.

Illesztés-mentes szekvencia távolságokat is alkalmaznak a bioinformatika különböző területein, azonban nem terjedtek el olyan széles körben, mint az illesztés-alapúak [20]. Az egyik legegyszerűbb és számítási szempontból a leghatékonyabb illesztés-mentes szekvenciatávolságon például két szekvencia nuklein- vagy aminosav eloszlásának relatív entrópiáját értjük. A korai automatikus fehérjeosztályozási rendszereknél többnyire ezt alkalmazták.

A távolságalapú faépítési algoritmusok bizonyos értelemben rokon módszerek a klaszterező eljárásokkal. A cél itt is az, hogy a kérdéses objektumok felett egy előre definiált hasonlóság/távolság alapján csoportokat, vagy idegen szóval klasztereket alakítsunk ki. Számos klasszikus agglomeratív hierarchikus klaszterező módszert alkalmaztak filogenetikus analízisre is, mint például a Single Linkage (SL), a Complete Linkage (CL) vagy az UPGMA algoritmusokat. Ezek alkalmazása azért kézenfekvő faépítési célokra, mert egy fastruktúraként ábrázolják az általuk meghatározott klaszterezést. A legismertebb távolságalapú faépítő módszer a Neighbor-Joining (NJ) algoritmus, amely metodikájában rokon a korábban említett SL vagy CL klaszterezőkkel. Az NJ egy divizív módszer, azaz felülről-lefelé klaszterezés alapján rekonstruálja a filogenetikus fát. Az NJ sikerét a kidolgozása után több mint 20 évvel is kutatják, és számos előnye és hátránya csak mostanában nyert bizonyítást [21; 22].

Az eddig említett távolságalapú faépítők az agglomeratív hierarchikus klaszterező algoritmusok oszályába tartoznak, mivel minden iterációban egy klasztert szétvágnak (top-down) illetve két klasztert összevonnak (bottom-up megközelítés) egy, a klasztereken értelmezett függvény alapján. Továbbá a

[1] A szekvenciatávolságok, amelyek elterjedtek a bioinformatikában általában nem teljesítik a háromszög egyenlőtlenséget. Továbbá sok esetben úgy vannak definiálva ezen függvények, hogy hasonlóságot fejeznek ki. Tehát a nagyobb érték nagyobb hasonlóságot jelent.

faépítő módszerek a kimeneti fának az éleihez rendelnek élsúlyokat is. Mivel a filogenetikus fa pontjai biológiai objektumokat reprezentálnak (a belső pontok hipotetikus objektumokat), ezért az élsúlyok, vagy élhosszak evolúciós távolságokat jelölnek. A távolságalapú módszereket arra a kérdésre nem adnak választ, hogy miért preferálunk egy fát jobban, mint a többit. Egy lehetséges megközelítésnek tűnik, hogy azt a fát tekintjük jobbnak, amelyre az X halmazon értelmezett fatávolságok és az X-en előre definiált d távolságértékek négyzetesen kevésbé térnek el. Vagyis értelmezzük egy $\mathcal{T}$ fára és egy d távolságra az úgynevezett fahibát a következő formulával:

$$e_{\mathcal{T}} = \sum_{x,y \in X} (d(x,y) - d_{\mathcal{T}}(x,y))^2 \tag{1}$$

Vezessük be a $P_{\mathcal{T}}$ él-út szomszédsági mátrixot egy adott n levelű $\mathcal{T}$ filogenetikus fára:

$$P_{\mathcal{T}}(p,e) = \begin{cases} 1 & \text{, ha} \quad e \in p \\ 0 & \text{különben,} \end{cases} \tag{2}$$

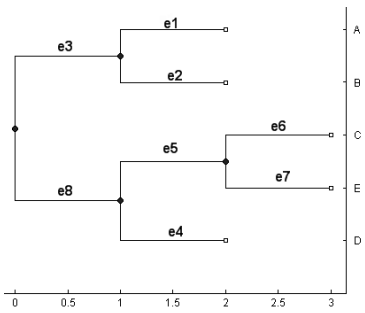
ahol az oszlopok az éleknek felelnek meg és a sorok pedig a fában található $\binom{n}{2}$ darab levélpárok közötti útnak. A 1. ábrán egy példa látható az él-út szomszédsági mátrix konstrukciójára. Ekkor egy optimalizálási feladat megoldásával minden filogenetikus fára és távolságfüggvényre egy minimális fahibát tudunk meghatározni, ahol az x vektor tartalmazza az optimális élsúlyokat és d vektorban pedig a távolságmátrix elemei vannak felsorolva olyan sorrendben, ahogyan az él-út szomszédsági mátrix sorainak megfelelnek:

$$e_{\mathcal{T}} = \min_{\mathbf{x} \in \mathbb{R}_+^{\mathbf{n-1}}} \| \left(P_T \mathbf{x} - \mathbf{d} \right) \| \tag{3}$$

Egy adott fára az $e_{\mathcal{T}}$ minimális fahiba értékét a Legkisebb Négyzetek módszerével tudjuk kiszámolni $O(n^4)$ időben egy n levelű fára [14]. Azóta bizonyítást nyert, hogy ez $O(n^2)$ időben is meghatározható [23]. A szakirodalomban ezt a távolságalapú kritériumot többnyire arra alkalmazták, hogy a faépítés után egy optimális élsúlyozást nyerjenek[24].

Az így definiált minimális fahiba alapján rangsorolni tudjuk az összes X-en értelmezett n levelű súlyozott filogenetikus fát egy rögzített d távolságfüggvényre. Day megmutatatta, hogy a legkisebb fahibával rendelkező n levelű fa megtalálása általában NP-teljes [4]. Emiatt a disszertációban erre a problémára az úgynevezett Multi-Stack módszert alkalmaztuk [3]. Ezt a beszédfelismerésben elterjedt megközelítést adaptáltuk erre a feladatra, mivel ez olyan problémák heurisztikus megoldására szolgál, melyeknek a megoldástere hatalmas.

Az MS módszer először az összes 3-nál kevesebb levelű fát előállítja, mivel tetszőleges távolságértékek mellett bármely 3-nál kevesebb levelű fára létezik élsúlyozás úgy, hogy a fahiba 0 lesz (azaz a 3 egyenletben leírt optimum értéke 0 lesz). Tehát nem tudunk fahiba alapján különbséget tenni a 3-nál kevesebb levelű fák között. Ezután a kezdeti lépés után a módszer iteratívan kapcsolja össze azokat a részfákat, amelyeket már előállított illetve a fahibája már meg van határozva. Az algoritmus a k. iterációban a k és k-nál kevesebb levelű fákat kapcsolja össze egy filogenetikus fákon értelmezett operátor segítségével. Ezt ismételten végrehajtva lépésről-lépésre egyre több elemet tartalmazó filogenetikusfa-kezdeményeket kapunk. Az azonos számú levéllel rendelkező fakezdeményeket korlátozott prioritási sorban tároljuk, melyekben legfeljebb K darab fát tudunk tárolni. Ezáltal hatékonyan tudjuk bejárni a fatér egy releváns részét.



	e1	e2	e3	e4	e5	e6	e7	e8
(A,B)	1	1	0	0	0	0	0	0
(A,C)	1	0	1	0	1	1	0	1
(A,D)	1	0	1	1	0	0	0	1
(A,E)	1	0	1	0	1	0	1	1
(B,C)	0	1	1	0	1	1	0	1
(B,D)	0	1	1	1	0	0	0	1
(B,E)	0	1	1	0	1	0	1	1
(C,D)	0	0	0	1	1	1	0	0
(C,E)	0	0	0	0	0	1	1	0
(D,E)	0	0	0	1	1	0	1	0

1. ábra. Egy filogenetikus fa és a hozzárendelt él-út szomszédsági mátrix, ahol $X = \{A, B, C, D, E\}$.

Konszenzusfa módszerek

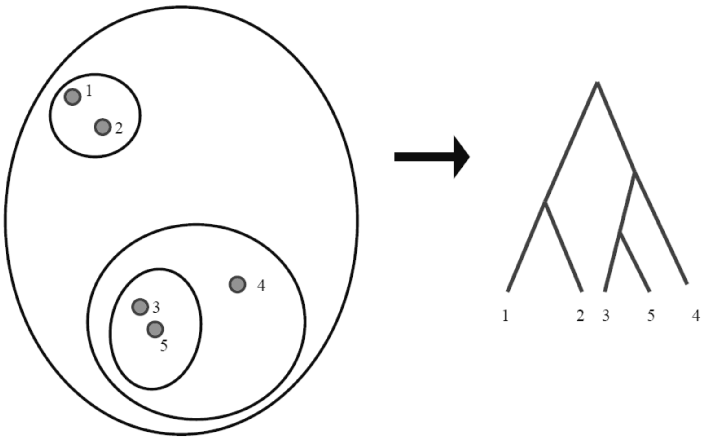
A konszenzusfa módszerek bemutatásánál célszerű kihasználni a gyökeres filogenetikus fák és a hierarchikus halmazrendszerek között található párhuzamot.

4. Definíció. *Legyen M egy véges halmaz. Az M nem üres részhalmazainak egy halmazát $\mathcal{H}$* hierarchiának *nevezzük ha minden $A, B \in \mathcal{H}$ elemre teljesül, hogy*

$$A \cap B \in \{\emptyset, A, B\} \tag{4}$$

A $\mathcal{H}$ hierarchia elemeit páronként kompatibilisnek nevezzük.

Mivel egy X halmaz feletti $\mathcal{T}$ gyökerezett filogenetikus fa minden belső pontja kijelöli X-nek egy részhalmazát (az alatta található elemek halmazát), ezért természetes módon le tudjuk képezni $\mathcal{T}$ belső pontjait olyan X feletti részhalmazokra, amelyekre igaz, hogy bármely kettő vagy tartalmazási relációban áll, vagy diszjunktak (2. ábra). Ezzel megadtunk egy kölcsönösen egyértelmű megfeleltetést a két struktúra között. A továbbiakban jelöljük $\mathcal{T}^C$-vel a $\mathcal{T}$ gyökerezett filogenetikus fához tartozó hierarchiát. Mivel a konszenzusfa módszerek több filogenetikus fából képeznek egy reprezentatív fát, ezért úgy is megfogalmazhatjuk ezeknek az algoritmusoknak a működését, hogy több hierarchikus



2. ábra. A filogenetikus fák és a hierarchikus halmazrendszerek analógiája.

halmazrendszer elemeiből egy kompatibilis részhalmazt választanak ki, amelyek az output konszenzusfa belső pontjainak felelnek meg.
A legelterjedtebb módszerek közé a következő algoritmusok tartoznak, melyeknek a leírása és néhány tulajdonsága David Bryant cikkében található meg [7]:

1. Strict consensus: csak azokat a részhalmazokat választja ki a konszenzusfa belső pontjának, amelyek minden fában megtalálhatóak

2. Majority-rule consensus: csak azokat a részhalmazokat választja ki a konszenzusfa belső pontjának, amelyek az input fák legalább felében megtalálhatóak

3. Greedy consensus: a részhalmazokat sorba rakja az input fákban való előfordulásai gyakorisága szerint, és akkor választ ki egy részhalmazt a keletkező konszenzusfa belső pontjának, ha az előtte lévőkkel kompatibilis

Ezek a konszenzusfa módszerek a részhalmazok előfordulásainak gyakorisága alapján választják ki azokat a részhalmazokat, amelyek a konszenzusfa belső pontjainak felelnek meg. Azonban megfogalmazhatunk egy általánosabb megközelítést is. Tegyük fel, hogy az input fák $\mathcal{C} = \bigcup_i \mathcal{T}_i^C$ hierarchiáinak az unióján adott egy $w^C : \mathcal{C} \to \mathbb{R}^+$ súlyfüggvény. Célunk azon $\mathcal{C}' \subseteq \mathcal{C}$ részhalmaz meghatározása, melyre $\sum_{c \in \mathcal{C}'} w(c) \to \max$. Ezt a megközelítést Max Clique Consensus (MCC) megközelítésnek nevezzük, és amennyiben az input fák száma nagyobb mint kettő, akkor ez a probléma NP-teljes[25].

Az MCC problémát felírtuk egy egészértékű bináris programozási feladatként. Ezt a problémát a jólismert Branch&Bound algoritmussal [26] hatékonyan meg lehet oldani, és ezáltal az MCC alkalmazhatóvá válik filogenetikus analízisre.

A [6] cikkben számos evolúciós modell alkalmazása mellett megvizsgáltuk ennek a metodológiának a pontosságát. Azt az esetet is megvizsgáltuk, amikor az input fák több különböző faépítő algoritmus által lettek előállítva, továbbá bevezettünk egy Maximum Likelihood (ML) alapú részhalmaz súlyozást.

A tapasztalatok azt mutatták meg, hogy érdemes az ML alapú súlyozást használni az MCC konszenzusfa-építővel, mivel számos teszt helyzetben jobb lett ez a megközelítés. Habár a Strict Consensus módszer néhány esetben pontosabb volt, azonban ekkor nagyon rossz minőségűek voltak a bemeneti filogenetikus fák.

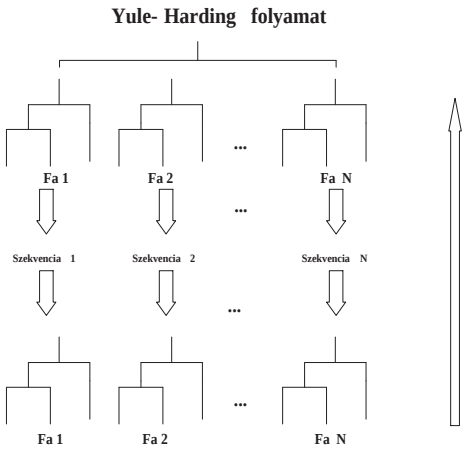
A filogenetikus analízis pontosságának becslése

A fehérjék pontos evolúciós története általában nem ismert. Ezért fontos az olyan módszerek kidolgozása, melyekkel a filogenetikus analízis pontosságát meg lehet becsülni, illetve a módszereket össze lehet hasonlítani. Bemutatunk egy tesztelési protokollt, mely erre alkalmas. A protokoll kidolgozásánál és implementálásánál számos biológiai megfigyelést szem előtt tartottunk, ezért ebben a keretrendszerben a faépítő módszerek előnyeit-hátrányait egyszerűen meg lehet határozni.

A tesztelési folyamat magja három lépésből áll:

1. Állítsunk elő egy N levelű filogenetikus fát, melynek az élhosszait valamely eloszlás szerint generáljuk.

2. Valamely evolúciós folyamat alapján generáljunk N elemű szekvencia halmazt az előző pontban generált véletlen fa elágazási mintázata szerint.



3. ábra. A filogenetikus fák tesztelésének folyamata.

3. Építsünk a vizsgált faépítő használatával egy fát az előbb generált szekvencia halmazra, majd hasonlítsuk össze az eredeti filogenetikus fával valamely fahasonlósági mérték szerint, ezzel kapunk egy becslést a filogenetikus analízis pontosságára.

Az 1. pontban két elterjedt fagenerálási módszert alkalmazhatunk: vagy egyenletes eloszlás alapján választunk ki egy véletlen fát az N levelű fák teréből, vagy az úgynevezett Yule-Harding folyamat szerint, amely nagyobb valószínűséggel generál kiegyensúlyozottabb fát [27; 28]. A második lépésben valamely evolúciós folyamat szerint generálhatunk véletlen szekvenciákat. Az evolúciós modellek széles arzenálja áll rendelkezésre ilyen célra [12]. Az utolsó lépésben felépítünk a vizsgált filogenetikus algoritmussal egy filogenetikus fát. Összehasonlítva a kapott fát az eredeti fával tudunk következtetni a faépítő pontosságára. Több fahasonlósági mérték terjedt el az evolúciós faépítésben. Ha csak a fatopológiák különbségét akarjuk vizsgálni, akkor a Robinson-Foulds távolságot alkalmazhatjuk, illetve súlyozott fákra a Kuhner-Felsenstein távolságot [29; 30]. A tesztelési folyamatot többször végrehajtva egy megbízható mérőszámot kaphatunk a faépítő pontosságáról.

I/1. Tézis

A szerző egy Multi-Stack alapú faépítő módszert dolgozott ki, amely a legkisebb négyzetek kritériumot alkalmazza. Ezáltal egy újszerű faépítő módszert kapunk, amely kompetitív a legelterjedtebb módszerekkel, és pontosabban meg lehet határozni olyan adatbázisok evolúciós történetét, ahol az objektumok hasonlósága alacsonyabb. Ezt a javulást mind illesztés-mentes mind evolúciós távolságok alkalmazásánál ki lehet mutatni. Továbbá a módszer jelentőségét emeli az, hogy a szintén legkisebb négyzetek kritériumot használó Fitch-Margoliash faépítő módszernél[31] számos tesztesetben jobb eredményt ér el a Multi-Stack megközelítés[3].

I/2. Tézis

A szerző visszavezette az MCC problémát egy bináris egészértékű programozási feladatra. Ezáltal tetszőleges részhalmazsúlyozás mellett meg lehet határozni a maximális súlyú kompatibilis részhalmazokat. Továbbá a szerző bevezetett egy Maximum Likelihood alapú részhalmaz súlyozást, mely által

az MCC hatékonyan alkalmazható konszenzusfa építésre összehasonlítva a legismertebb konszenzusfa-építő módszerekkel. Módszereinket a széles kőrben elterjedt PAUP programcsomag[32] által konstruált fákon hasonlítottuk össze[6]. Egy valós életből vett fehérjecsoporton bemutattuk a gyakorlati alkalmazhatóságát is.

I/3. Tézis

A szerző megadott egy tesztelési keretrendszert, amely alkalmas a faépítő eljárások teljes körű összehasonlítására több evolúciós modell alkalmazásával[3; 6]. A tesztelési módszerben egy előre meghatározott evolúciós fa alapján állítunk elő mesterségesen egy szekvenciahalmazt. Majd ezen szekvenciahalmazra a vizsgált faépítő módszerek alkalmazásával állítjuk elő a filogenetikus fát. Ezek után az eredeti és a kapott fa hasonlósága alapján meg tudjuk becsülni a filogenetikus analízisünk pontosságát. A tesztelési módszer szélesebb körű tesztelést tesz lehetővé, mint a hagyományos bootstrap módszer, mivel a bootstrap módszer egy rögzített szekvenciahalmazból vesz újra mintát [12]. Ezzel szemben ebben a keretrendszerben megvizsgálhatjuk a faépítők viselkedését különböző evolúciós modellek alkalmazása mellett.

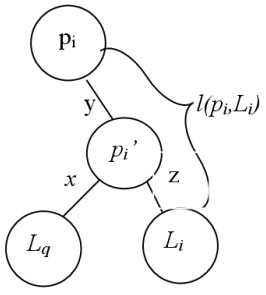
II. Filogenetikus fa-alapú fehérjeosztályozási módszerek

Eisen vezette be a fehérjeosztályozás területén az evolúciós fák használatát, ezt az újszerű megközelítést filogenomikának nevezte el [11]. Mára már több módszert is kidolgoztak ezen a területen. Sajnos az eddig fejlesztett módszerek sok esetben nem képesek megbirkózni nagy méretű adatbázisokkal, mivel általában nem csak páronkénti szekvenciahasonlóságra van szükségük a módszereknek, hanem például génduplikációk felismerésére vagy többszörös szekvencia illesztésre. Ezért sokszor költséges és nehéz ezen módszerek használata valós gyakorlati alkalmazásban.

A disszertációban több módszert mutatunk be, melyek a fehérjeosztályozás problémáját célozzák megoldani. A mi megközelítésünkben csak a fehérje szekvenciák hasonlóságára van szükségünk. Ezeket a hasonlósági viszonyokat egy fastruktúrában tároljuk. Az összes itt bemutatott módszer azon a feltételezésen alapul, hogy egy súlyozott filogenetikus fa segítségével jobban tudjuk reprezentálni a fehérjék hasonlósági viszonyát.

TreeInsert és TreeNN módszerek

A TreeInsert módszer azon a megközelítésen alapul, hogy meg tudjuk konstruálni az ismert osztálycímkékkel rendelkező fehérje adatbázis pontos filogenetikus fáját. Ezt a $\mathcal{T}$ súlyozott filogenetikus fát a fehérjeszekvenciák $d(x, y)$ hasonlóságai alapján építjük fel. Ezek után próbáljuk meg elhelyezni ebbe a fába az ismeretlen q elemet olyan módon, hogy az a legjobban "illeszkedjen" a $\mathcal{T}$ evolúciós fába. Követve a filogenetika konvencióit, a beszúrandó elemet csak levélként szúrjuk be a $\mathcal{T}$ fába. Ennek megfelelően a q elem beszúrása a 4. ábrán látható, ahol L_i jelenti a i. fehérjét reprezentáló levelet, és p_i pedig az L_i beszúrás előtti szülőjét a fastruktúra szerint. A q elem beszúrása után p_i' lesz az új elemet reprezentáló L_q levél és az L_i levél közös őse. Az x, y és z jelölik az ismeretlen élhosszakat.



4. ábra. Az ismeretlen q elem beszúrása levélként az L_i levél mellé.

Az eredeti feltételezésnek megfelelően azt szeretnénk meghatározni, hogy mennyire illeszkedik egy új elem a $\mathcal{T}$ fába. Ezért definiáljuk a q ismeretlen elem L_i levélre vonatkozó beszúrási költségét olyan módon, hogy a beszúrási költség annál kisebb legyen, minél jobban illeszkednek az ismeretlen elemre vonatkozó d hasonlósági mértékek és $d_{\mathcal{T}}$ fatávolságok egymáshoz. Ezért vizsgáljuk meg, hogy a q elem fatávolsága a fa többi levelétől mennyire képes illeszkedni az $d(x, y)$ hasonlóságokhoz:

$$\min_{0\leq x,y} \left(\sum_{j=1}^{n} \left(d\left(L_j, L_q\right) - d_{\mathcal{T}}\left(L_i, L_q\right) \right) \right)^2 \tag{5}$$
$$\text{s.t.} \quad y + z = d_{\mathcal{T}}\left(p_i, L_i\right)$$

Az i. levélre vonatkozó $IC(L_q, L_i)$ beszúrási költségnek legyen az x optimális értéke, ahol az 5 egyenletben leírt kifejezés a minimális értékét veszi fel. A TreeInsert algoritmus minden levél mellé beszúrja a fába az ismeretlen osztálycímkével rendelkező q elemet, majd a minimális beszúrási költség alapján meghatározza, hogy melyik levél mellé lehetett a legjobban az ismeretlen elemet beszúrni. Ilyen módon el lehet helyezni az adatbázis evolúciós történetébe az ismeretlen elemet. Ezek után a TreeInsert módszer az ismeretlen elemhez azon elem osztálycímkéjét rendeli hozzá, amely mellé legjobban illeszkedett, azaz ahol minimális volt a beszúrási költség.

A TreeNN módszer lényegesen egyszerűbb algoritmus. Ennél a megközelítésnél szintén azt feltételezzük, hogy adott egy ismert címkézésű D adatbázis –szekvenciák egy halmaza–, melynél ismerjük az elemek páronkénti szekvenciahasonlóságait. A q ismeretlen elemet ugyanazon szekvenciahasonlóság szerint összehasonlítjuk a D minden egyes elemével. Ezek után egy távolságalapú faépítővel megkonstruálunk egy súlyozott filogenetikus fát a D adatbázisra és a q elemre. Ez a súlyozott filogenetikus fa meghatároz egy fatávolságot a q elem és az ismert címkézésű D adatbázis elemei között. Ezen távolság alapján elvégezhetjük az osztályozást az ismeretlen elemre a Legközelebbi Szomszéd módszerével[33].

A TreeInsert módszer először az ismert címkékkel rendelkező, úgynevezett tanuló adatbázisra felépít egy filogenetikus fát, és minden tesztelemre megkeresi a neki legjobban megfelelő helyet a fent leírt módon. Ezzel szemben a TreeNN módszer minden tesztelemre és a tanuló adatbázisra rekonstruál fát. TreeInsert modell osztályozásban jó eredményeket ért el, azonban minden osztályozandó elem esetében el kell végeznünk n darab optimalizálást a minimális beszúrási költség kiszámításához. Több ismert modellkiértékelő metrikát használtunk a módszerek összehasonlításánál, és megmutattuk, hogy a TreeNN és a TreeInsert közel azonos eredményeket értek el[9].

TreeProp-N

Számos területen alkalmaznak olyan módszereket, amelyeknél az adatok hasonlósági viszonyait egy speciális gráfban tárolják (teljes gráf vagy fagráf). Ezen a struktúrán egy úgynevezett propagációt hajtanak végre, amely abból áll, hogy a gráf szomszédos pontjai minden iterációban egy üzenetet küldenek egymásnak. Ezen üzenetek általában valós számoknak felelnek meg. Ilyen propagációs módszerek közé tartoznak a PageRank [34], Message Passing [35], Affinity Propagation [36] algoritmusok. Mi a PageRank egy specialis alakjából indultunk ki, az úgynevezett Personalized PageRank algoritmusból [37; 38]. Ezt a módszert elsősorban információ visszakeresésrere alkalmazták, és a módszer sikerét a Google internetes kereső is jól tükrözi. A fő gondolat emögött az eljárás mögött a következő: ha van egy ismeretlen elemünk -ezt az elemet általában query vagy ismeretlen elemnek hívjuk-, amelynek ismerjük a hasonlósági értékeit a már előre címkézett adatbázis elemeihez, akkor az ismeretlen elem osztályba sorolásánál vegyük figyelembe a címkézett adatbázis hasonlósági viszonyait is.

Hasonlóan a korábbi jelölésekhez, jelölje az ismeretlen osztálycímkével rendelkező elemet q, és

legyen az ismert címkékkel rendelkező adatbázis $D = \{y_1, y_2, \ldots, y_n\}$. Az

$$y(0) = (d(q, y_1), d(q, y_2), \ldots, d(q, y_n))$$

jelölje az eredeti hasonlóságokat tartalmazó vektort a "prior" adatbázis és az ismeretlen elem között. A D adatbázison belüli hasonlósági értékeket jelölje az $S = d(y_i, y_j)$ mátrix. Az $y(t)$ pedig jelölje a t. iteráció utáni hasonlósági értékeket. Ezek után a 6. egyenletben leírt propagációs szabály elvégzésével új hasonlósági értékeket kapunk, amelynél már figyelembe vesszük a hasonlósági viszonyokat, mivel az S matrix tartalmazza ezt:

$$y(t+1) = (1-\alpha)y(0) + \alpha S y(t) \tag{6}$$

A képletből látszik, hogy az ismeretlen q elemhez egy y_i elem iterációnként hasonlóbb lesz, ha több olyan elemet tartalmaz D, amelyek mind q elemhez mind y_i elemhez hasonlóak. Ilyen módon a D adatbázis hasonlósági struktúráját ki tudjuk használni. Ez az iteratív módszer a gyakorlatban lassú, mert sokszor nagyon nagy hálózatokkal kell dolgozni, és minden q ismeretlen elemre propagáltatni kell. Ezért mi egy ritkább struktúrával helyettesítjük a hálózatot, nevezetesen egy gyökértelen bináris filogenetikus fával, melyet a D adatbázisra és az ismeretlen elemre építünk. Mivel n elemre építünk bináris fát, ezért n helyett $2n-1$ pontú lesz a hálózatunk, amelyen propagáltatunk. Azonban minden pontnak legfeljebb 3 a foka. Ebben az esetben $y_i(0)$ az ismeretlen elem fatávolságát jelöli az i. elemtől, az $N(p_i)$ pedig a p_i pont szomszédjait a fában. A propagációs szabály a következőképpen írható fel ebben az esetben:

$$y_i(t+1) = (1-\alpha)y_i(0) + \alpha \sum_{p \in N(p_i)} d_{\mathcal{T}}(p, p_i) y_p(t). \tag{7}$$

Mivel a fa éleinek a hosszai távolság jellegű adatok, ezért a propagálás előtt egy monoton csökkenő leképezéssel hasonlóság jellegű adatokká kell átalakítani. A korábban leírt Personlized PageRank módszerrel teljesen analóg ez a megközelítés, azonban míg ott az adatbázis hasonlósági struktúrája egy teljes súlyozott gráffal volt leírva, addig itt az elemek hasonlósága egy súlyozott binaris filogenetikus faként van reprezentálva.

Fehérjeosztályozásban ezt a módszert egyszerűen lehet alkalmazni. Jelölje a 7. propagációs formula határpontját y^*, az az ahova a $y_1, y_2, \ldots, y_T, \ldots$ konvergens pontsorozat tart[2]. Az osztályozást a y^* maximális értéke alapján hajtjuk végre. Amennyiben az iterációszámot nullának választjuk, akkor ez a módszer ekvivalens a TreeNN-nel, mivel az y_0 vektor a fatávolságokat tartalmazza.

TreeProp-E

A TreeProp-N algoritmusnak kifejlesztettünk egy másik változatát is. Az TreeProp-E módszernél az éleken végezzük el a propagációt hasonló szabályok szerint, mint a TreeProp-N-nél. Egy n levelű $\mathcal{T}$ súlyozott filogenetikus fának az élsúlyait jelölje az $y(0)$ vektor, melynek az i. komponensét $y_i(0)$ jelöli. Egy fa két élét akkor mondjuk szomszédosnak, ha van közös pontjuk. Jelölje az e élnek a szomszédos éleit $N(e)$. Ezek után a propagációs szabályt átfogalmazhatjuk a fa éleire:

[2] A kovergencia a Perron-Frobenius tétel következménye.

$$y_i(t+1) = (1-\alpha)y_i(0) + \frac{\alpha}{|N(e_i)|}\sum_{e_j \in N(e_i)} y_{e_j}(t). \quad (8)$$

Ennek a módszernek is van egy szemléletes motivációja, mint a Personlized PageRank algoritmusnak. Egy súlyozott filogenetikus fában az élek hasonlóságokat reprezentálnak, és egy él abban az esetben megnyúlik –nagyobb értéket kap– ha több olyan szomszédos éle van, amely szintén nagy hasonlóságokat reprezentál.

Osztályozási feladatra ezt a módszert úgy alkalmazhatjuk, hogy az ismeretlen címkével rendelkező q elemre és a D adatbázisra felépítünk egy filogenetikus fát, és erre végrehajtjuk a 8 egyenletben leírt faalapú propagációt [3]. A propagálás után kiszámoljuk a kapott fatávolságokat a q elemhez. Ezen távolsági értékek alapján el tudjuk végezni az osztályozást, például egy Legközelebbi Szomszéd osztályozó segítségével. Kísérletienben az TreeProp-E sok esetben hatékonyabb fehérje osztályozó módszernek bizonyult a hagyományos gépi tanulási eljárásoknál, mint például Artificial Neural Network [39], Support Vector Machine [40].

ROC analízis

A ROC analízis a legelterjedtebb kiértékelési technika a fehérjeosztályozási modellekre, melyet a bioinformatikában[41] illetve a jelfeldolgozásban [42] igen széles körben alkalmaznak. A fehérje osztályozásban általában többosztályos tanulási feladatokkal találkozunk. Ezekeket a feladatokat több-az-egy-elleni tanulási feladatként, azaz bináris tanulási feladatokként is értelmezhetjük, ahol a kitüntett osztályt – más néven célosztályt – pozitív osztálynak nevezzük, míg a többi osztály elemeit negatív elemeknek tekintjük. Az ilyen bináris gépi tanulási problémáknál egy tanuló halmazon betanított binaris osztályozó minden tesztelemhez hozzárendel egy pozitív osztályra vonatkozó valószínűséget. Ezen valószínűségek alapján rangsorolni tudjuk a tesztelemeket. Ezek után minden $t \in [0,1]$ valós szám természetes módon kettéosztja a teszthalmazunkat olyan módon, hogy a t alatti valószínűségekkel rendelkező elemeket negatívnak tekintjük, míg a többi elemet pozitívnak. Azaz minden t megfelel egy döntési küszöbnek. Ekkor ha a tesztelemek valós osztálycímkéjét is figyelembe vesszük, akkor négy csoportba tudjuk sorolni a tesztelemeket t-től függően:

1. A bináris osztályozónk pozitívnak jósolta a tesztelemet, és a valós címkéje is pozitív. Ezen elemek a valós pozitív elemeknek felelnek meg. A teszthalmazban jelölje ezen elemek számát $TP(t)$.

2. A bináris osztályozónk pozitívnak jósolta a tesztelemet, de a valós címkéje negatív. Ezen elemek a hibás pozitív elemeknek felelnek meg. A teszthalmazban jelölje ezen elemek számát $FP(t)$.

3. A bináris osztályozónk negatívnak jósolta a tesztelemet, de a valós címkéje pozitív. Ezen elemek a hibás negatív elemeknek felelnek meg. A teszthalmazban jelölje ezen elemek számát $FN(t)$.

4. A bináris osztályozónk negatívnak jósolta a tesztelemet, és a valós címkéje is negatív. Ezen elemek a valós negatív elemeknek felelnek meg. A teszthalmazban jelölje ezen elemek számát $TN(t)$.

[3] A kísérletek azt mutatták, hogy általában 20 iteráció után nem változott $y_i(t)$ értéke.

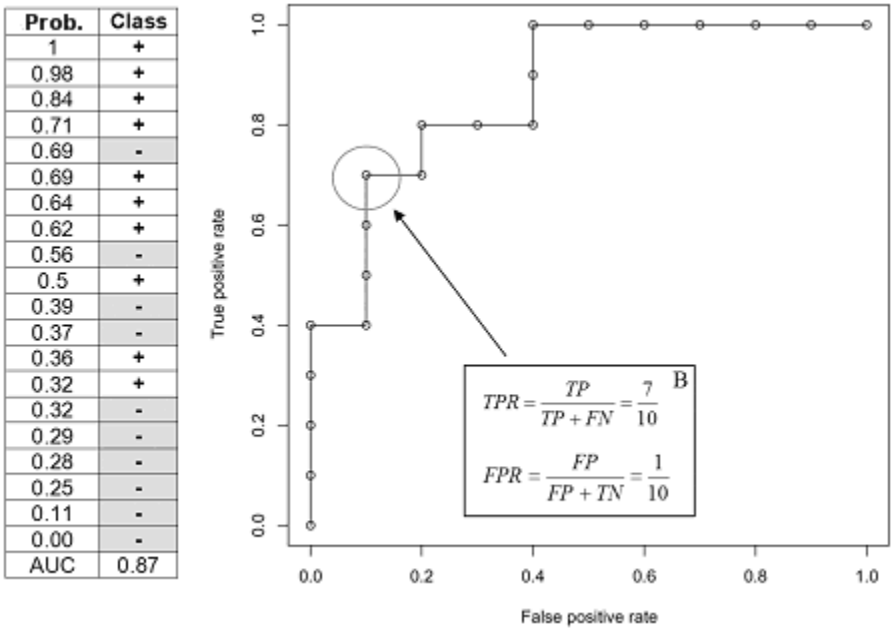
A fent megadott értékek alapján a teszthalmazon definiáltunk egy hibás pozitív hányadost (FPR), illetve egy valós pozitív hányadost (TPR), amely értékek a t paramétertől függőek:

$$TPR(t) = \frac{TP(t)}{TP(t) + FN(t)}, \qquad FPR(t) = \frac{FP(t)}{FP(t) + TN(t)} \tag{9}$$

Ezek után a ROC görbe definíciója természetesen adódik.

5. Definíció. *Adott egy egységoldalú négyzet $N = [0,1]^2$, továbbá adottak minden $t \in [0,1]$ értékre a $TPR(t)$ és $FPR(t)$ értékek. A ROC görbe a $(TPR(t), FPR(t))$ pontok halmaza.*

A ROC görbe alatti terület értelmezhető úgy, hogy mi annak a valószínűsége, hogy egy tesztelem pozitív címkét kap a modellünk szerint és valóban pozitív az osztálycímkéje[42]. Ezért a tanulómódszerek kiértékelésének egy megbízhatóbb mértékét szolgáltatja a ROC görbe alatti terület (Area Under Curve, AUC) annál, mintha például a modellünk pontosságát használnánk modellkiértékelési metrikaként. A 6. ábra egy egyszerű példát mutat be a ROC analízis alkalmazására.



5. ábra. A ROC görbe számítása rangsorolt elemeken. A bal oldalon látható a bináris osztályozó által előállított pozitív osztályra vett osztályvalószínűségek és a tényleges osztályok. Ezek alapján kirajzolható a ROC görbe, amely a jobb oldalon látható.

A szekvenciaosztályozásban is elterjedt a ROC analízis, azonban itt sokszor a szekvenciahasonlóságok kiértékelésére alkalmazzuk olyan módon, hogy az osztályozni kívánt, ismeretlen szekvenciát összehasonlítjuk az ismert címkékkel rendelkező adatbázis minden egyes elemével a vizsgált szekvencia hasonlóság alkalmazásával. Ezek után a kapott hasonlósági értékek alapján a tanuló adatbázist rangsorolni tudjuk, hasonlósági érték szerint csökkenő sorrendbe. Majd a ROC analízist lehet alkalmazni a kapott rangsorra. Ebben az esetben a rangsor eleje a fontos a kiértékelés szempontjából, mert természetes módon azt várjuk el, hogy sok pozitív elem kerüljön a rangsor elejére. Ezért a rangsort sokszor csak az első n darab negatív elemig szoktuk figyelembe venni, és a rangsor végét

kihagyjuk a kiértékelésből [41]. Az ilyen módon alkalmazott ROC analízist n értékétől függően ROC_n kiértékelésnek hívjuk.

A fehérje adatbázisok alapvető problémája az, hogy az ismert (szerkezeti vagy funkciós) osztályok nagyon heterogének az adatok száma, illetve egymás közti hasonlóságuk szempontjából. Ezért nehéz megfelelően megválasztani n értékét. Ennek a választására adunk egy egyszerű módszert [43]. A ROC analízis módszerét úgy alkalmazzuk, hogy egy sokkal megbízhatóbb mérőszámot kapunk a fehérje-csoportokra vonatkozóan. Ezáltal a ROC analízis kevésbé lesz érzékeny a tanulóhalmaz méretére. Az adatbázisok fejlesztésénél hasznos információt szolgáltathat az itt bemutatott kiértékelő módszer.

II/1. Tézis

A szerző a tézisében megadja a TreeInsert és a TreeNN módszert, melyek újszerű faalapú fehérje-osztályozási eljárások. A korábbi filogenomikai módszerekkel szemben, az itt bemutatott módszerek csak a szekvenciahasonlóságokat használják fel, emiatt egyszerűen alkalmazhatóak széles körben. Több fehérjeosztályozási problémán összehasonlította a faalapú módszereket ROC analízis alkalmazásával, és jelentős javulást ért [9]. Az eredmények rámutatnak, hogy érdemes filogenetikus információt alkalmazni fehérje osztályozásban.

II/2. Tézis

A tézisben kidolgozásra került két filogenetikusfa-alapú propagációs módszer, a TreeProp-N és a TreeProp-E. Ezek a módszerek a TreeNN algoritmus kiterjesztéseinek tekinthetőek olyan módon, hogy a filognetikus fa struktúráját felhasználva a szekvencia hasonlóságokat felüldefiniálják. Ezen propagációs módszerek fehérjeosztályozásban további javulást eredményeztek a korábbi módszerekhez képest. [10].

II/3. Tézis

A szerző definiált egy ROC analízisen alapuló kiértékelési módszert, mellyel egy megbízhatóbb mérőszám kapható a szekvenciahasonlóság minőségére abban az esetben, ha kiegyensúlyozatlan az osztályok eloszlása az adatbázisban [43]. Ezáltal sokkal megbízhatóbb modellkiértékelést lehet végrehajtani a ROC analízis segítségével. Az itt bevezetett módszert a szerző nagy méretű fehérjeadatázison tesztelte.

Konklúzió

A gének és fehérjék nyelvének a megértéséhez elsősorban azt kell megfejtenünk, hogy hogyan alakultak ki az evolúció soran ezen biológia objektumok szekvenciái. Ennek a folyamatnak a modellezése elengedethetetlen feladat. Emiatt szükséges újabbnál újabb filogenetikus módszerek fejlesztése, melyek az evolúció folyamatát minél jobban képesek feltárni.
A disszertációban szereplő Multi-Stack faépítő módszer egy robosztus távolságalapú heurisztikus algoritmus, amely számos tesztkörnyezetben jól szerepelt. Ezzel bebizonyította gyakorlati jelentőségét. A Max Clique Consensus problémára adtunk egy bináris egészértékű lineáris programozási feladatot. Ezáltal egy jól alkalmazható filogenetikus konszenzusfa-építő módszerhez jutottunk, amely remélhetőleg széles körben használt eszközzé fog válni.

A disszertációban bemutatunk számos fehérjeklasszifikációs módszert, amelyek a filogenetikus faépítő algoritmusok egyfajta újszerű alkalmazását mutatják be. Ezek a módszerek a gépi tanulás és a filogenetika határterületét képezik. Számos biológiai osztályozási problémát a korábbi megközelítéseknél hatékonyabban oldottunk meg, ezzel is bizonyítva azt a tényt, hogy fontos és szükséges újabbnál-újabb filogenomikai és filogenetikai módszerek kidolgozása.

Hivatkozások

[1] Saitou N. and Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, 1987.

[2] Rohlf F. J. Classification of aedes by numerical taxonomic methods (diptera: Culicidae). *Ann Entomol Soc Am*, 56:798–804, 1963.

[3] Busa-Fekete R., Kocsor A., and Bagyinka Cs. A multi-stack based phylogenetic tree building method. *Lencture Notes in Bioinformatics*, 4463:49–60, 2007.

[4] Day W.H.E. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49:461–467, 1986.

[5] Adams E.N. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology*, 21:390–397, 1972.

[6] Busa-Fekete R., Bánhalmi A., Kocsor A., and Bagyinka Cs. A binary integer programming relaxation for the max clique consensus. *European Journal of Biophysics*, 2008.

[7] Bryant D. A classification of consensus methods for phylogenetics. *Bioconsensus,Discrete Mathematics and Theoretical Computer Science*, 61:163–184, 2001.

[8] Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, October 1990.

[9] Busa-Fekete R., Kocsor A., and Pongor S. Tree-based algorithms for protein classification. *Computational Intelligence in Bioinformatics, Studies in Computational Intelligence*, 7:0–0, 2008.

[10] Kocsor A., Busa-Fekete R., and Pongor S. Protein classification based on propagation on unrooted binary trees. *Protein and Peptide Letters*, page in press, 2008.

[11] Eisen J.A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, 8:163–7, 1998.

[12] Felsenstein J. *Inferring Phylogeneties*. Sinauer, 2004.

[13] Felsenstein J. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

[14] Fitch W. M. and Margoliash E. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

[15] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970.

[16] Smith T. F. and Waterman M. S. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981.

[17] Gotoh O. An improved algorithm for matching biological sequences. *J Mol Biol*, 162(3):705–708, December 1982.

[18] Jukes T. H. and Cantor C. R. Evolution of protein molecules. *Mammalian Protein Metabolism, Academic Press, New York*, edited by H. N. MUNRO:21–132, 1969.

[19] Kimura M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution.*, 16:111–120, 1980.

[20] Vinga S. and Almeida J. Alignment-free sequence comparison-a review. *Bioinformatics*, 19(4):513–523, March 2003.

[21] Gascuel O. and Steel M. Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11):1997–2000, November 2006.

[22] Bryant D. On the uniqueness of the selection criterion in neighbor-joining. *Journal of Classification*, (22):3–15, 2005.

[23] Bryant D. and Waddell P. Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Journal of Biochemical and Biophysical Methods*, 15(10):1346–1359, 1998.

[24] Grunewald S., Forslund K., Dress A., and Moulton V. Qnet: An agglomerative method for the construction of phylogenetic networks from weighted quartets. *Molecular Biology and Evolution*, 24(2):532–538, February 2007.

[25] Bryant D. *Hunting for trees, building trees and comparing trees: theory and method in phylogenetic analysis*. PhD thesis, Dept. Mathematics, University of Canterbury, 1997.

[26] Land A.H. and Doig A.G. An automatic method for solving discrete programming problems. *Econometrica*, 28:497–520, 1960.

[27] Yule G. A mathematical theory of evolution. *Based on the conclusions of Dr. J. C. Willis. Philos. Trans. Roy. Soc. London Ser. B, Biological Sciences*, 213:21–87, 1925.

[28] McMorris F. R. and Powers R. C. Consensus weak hierarchies. *Bulletin of Mathematical Biology*, 53:679–684, 1991.

[29] Kuhner M. and Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates [published erratum appears in mol biol evol 1995 may;12(3):525]. *Mol Biol Evol*, 11(3):459–468, 1994.

[30] Robinson D. F. and Foulds L. R. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.

[31] Felsenstein J. Phylip program package. *http://evolution.genetics.washington.edu/phylip.html*, 2007.

[32] Swofford D. Paup program package. *http://paup.csit.fsu.edu/index.html*, 2007.

[33] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000.

[34] Brin S. and Page L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.

[35] Lauritzen S. L. and Spiegelhalter D. J. Local computations with probabilities on graphical structures and their application to expert systems.

[36] Brendan J J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, January 2007.

[37] Page L., Brin S., Motwani R., and Winograd T. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[38] Zhou D., Weston J., Gretton A., Bousquet O., and Schölkopf B. Ranking on data manifolds.

[39] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[40] Vapnik V. N. *Statistical Learning Theory*. John Wiley and Son, 1998.

[41] Gribskov M. and Robinson N. Use of receiver operating characteristic (roc) analysis to evaluate sequence matching, 1996.

[42] Egan J.P. *Signal Detection theory and ROC Analysis*. New York: Academic Press, 1975.

[43] Busa-Fekete R., Kertész-Farkas Attila, Kocsor A., and Pongor S. Balanced roc analysis (baroc) protocol for the evaluation of protein similarities. *Journal of Biochemical and Biophysical Methods*, page doi:10.1016/j.jbbm.2007.06.003, 2007.