

TDK-dolgozat

Dr. ICLĂNZAN David Andrei

NAGHI Mirtill - Boglárka

SAPIENTIA ERDÉLYI MAGYAR TUDOMÁNYEGYETEM

Marosvásárhelyi Kar
Matematika-informatika tanszék



TDK-dolgozat

A Collatz-probléma jellemzése lokális optimum hálózatokkal

Témavezető

Dr. ICLĂNZAN David Andrei

Docens

Hallgató

NAGHI Mirtill - Boglárka

Szoftverfejlesztés, II. évf.

Kivonat

A hírhedt Collatz-sejtés, más néven a $3n+1$ probléma kimondja, hogy két alapvető aritmetikai művelet megismétlése végül minden pozitív egész számot egyé alakít. Míg megfogalmazása nagyon egyszerű, a sejtés „rendkívül nehéz problémának bizonyult, amely a mai matematika számára teljesen megoldhatatlan” (Lagarias, 2010), Erdős Pál is „teljesen reménytelenek” vélte. Ebben a dolgozatban a sejtés nehézségének újszerű jellemzését adjuk. Először is átfogalmazzuk a sejtést bináris optimalizálási feladatként, ahol gyakorlatban az a cél, hogy megtaláljuk azokat a bitsztringeket, amelyek relatív hossznövekedése maximális a Collatz-folyamat során. Ezt követően elemezzük az optimalizálási probléma szerkezetét és jellemzőit különféle fitness landscape metrikák és a helyi optimumok hálózat alapú modelljeinek segítségével. Az összes vizsgált problémaméretnél azt találtuk, hogy az így létrejövő tájak egyenetlenek és gyenge szerkezetűek. Úgy tűnik, hogy a helyi optimumok nem korrelálnak egymással, és nem is a keresési tér egy kis részére korlátozódnak.

Kulcsszavak: Collatz-sejtés, fitness táj analízis, helyi optimum hálózatok, fitness

1. fejezet

Bevezető

A Collatz-sejtés azt kérdezi, hogy két alapvető aritmetikai műveletet ismételve (ha a szám páros, akkor a sorozat rákövetkező elem ennek fele lesz, különben viszont a háromszorosánál eggyel nagyobb szám) végül minden pozitív egész szám eggyé alakítható-e át. Közelmúltbeli előrehaladása ellenére még mindig nem tudjuk, hogy egy szám kikerülhet-e a végtelen (1-4-2-1) hurkából.

Vegyük például a 10-et: ez páros szám, ezért elosztjuk kettővel, hogy 5-öt kapjunk. Mivel az 5 páratlan, megháromszorozzuk, és hozzáadunk 1-et. Most 16-ot kaptunk, ami páros, ezért felezzük, hogy 8-at kapjunk, majd ezt ismét felezzük megkapva a 4-est, amit szintén felezni kell, így 2-öt kapunk, és még egyszer osztunk, hogy 1-et kapjunk. Mivel az 1 páratlan, megháromszorozzuk, és hozzáadunk 1-et. Most visszatértünk a 4-hez, amelyről már tudjuk hová tart: 4, 2, 1. Az 1 után újból a 4-es következik... Beragadtunk egy hurokba.

Páratlan szám esetén hasonlóképp történik, például 11 esetén: ez páratlan, ezért megháromszorozzuk, és hozzáadunk 1-et. Most 34-et kaptunk, ami páros, tehát felezzük, hogy 17-et kapjunk, ezt megháromszorozzuk, és hozzáadunk 1-et, hogy 52-t kapjunk, ezt felezzük, hogy 26-ot kapjunk, és újra, hogy megkapjuk a 13-t, ezt megháromszorozva és hozzáadva egyet 40-hez érünk, amelyet felezni kell, eljutván a 20-hoz. Ezt felezzük, megkapva a 10-es számot, majd ismételt felezés után az 5-öst. Ezt megháromszorozva és hozzáadva egyet, kapjuk a 16-ot, amely páros és a felezett értékei is azok lesznek, tehát osztva őket kettővel kapjuk, hogy: 8, 4, 2, 1. 1 után megállapítottuk fentebb, hogy ismét a 4-es következik, vagyis ismét beleragadtunk a hurokba.

$$f(n) = \begin{cases} n/2, & \text{ha } n \text{ páros} \\ 3 \times n + 1, & \text{ha } n \text{ páratlan} \end{cases} \quad (1.1)$$

A fenti függvény egy n bemenetet vár, és két szabály egyikét alkalmazza rá, attól függően, hogy a bemenet páros vagy páratlan. Ez az f függvény a fent leírt

eljárás szabályait hajtja végre: Például $f(10) = 10/2 = 5$, mivel a 10 páros, és $f(5) = 3 \times 5 + 1 = 16$, mivel az 5 páratlan. A páratlan bemenetekre vonatkozó szabály miatt a Collatz-sejtés $3n + 1$ problémaként is ismert. A számítások során megkapott számokra a későbbiekben Collatz-pályaként hivatkozunk majd.

A kutatás első és legfontosabb célja a fent bemutatott probléma bináris optimalizálási feladatként való átfoglamzása, annak nehézségi jellemzőinek megadása.

Az egyszerűbb mérőszámok mellett, mint például a fitness-távolság-korrelációk[1] és a megoldások sűrűség-eloszlása[2], a lokális optimum hálózatok (angolul *local optima networks*, rövidítve LON-ok)[3, 4, 5] a kombinatorikus tájak szerkezetének gazdag jellemzését adja. A LON-ok diszkrét fitness-tájak durva szemcsés és tömörített modelljei, ahol a csomópontok helyi optimumok, az élek pedig átmeneti vagy elérhetőségi valószínűségek. Ez a hálózati modell egy új, sokrétű metrikakészletet hoz a komplex hálózatelemzés területéről[6], amely megkönnyíti a bőséges jellemzést. Az első cél elérése érdekében ki szeretnénk számítani és elemezni ezeket a mérőszámokat minden helyi optimumra, 16 és 32 bites hosszúság között.

A keresőalgoritmusok hatékonyságát befolyásoló tájjellemzők azonosítása nem csak a probléma keménységének kifejezésére, hanem a tervezésük és a paraméterezésük javítására is releváns. Ezért a második célunk esetén olyan optimalizálási módszerek kifejlesztései szerepelnek, amelyek következetesen megtalálják azokat a bitkarakterláncokat, amelyeknek megfelelő $f(b_x)$ 1-hez közeli vagy akár 1 feletti. Ez túlmutat jelen dolgozat céljain, de a további munka során mindenképpen foglalkoznunk kell a kérdéssel!

Ezek a módszerek lehetővé tehetik a Collatz-probléma hatékonyabb feltárását nagyon nagy (több száz bites hosszúságú) kiindulási értékek esetén, segítve a probléma dinamikájának feltárását a már felvázolt 68 bites határ felett.

A sejtés megszületése óta számtalan vizsgálatot tett eddig a tudományos világ, a következő fejezetben ezeket fogjuk áttekinteni.

2. fejezet

Szakirodalmi áttekintés

A hírhedt Collatz-sejtés[7], más néven $3n + 1$ -probléma, Kakutani-probléma, Ulam-probléma, Hasse-algoritmus és Syracuse-probléma, azt kérdezi, hogy két alapvető aritmetikai műveletet ismételve végül minden pozitív egész szám egygé alakítható-e át. Egyszerű megfogalmazása ellenére ez az 1937-ben javasolt sejtés nagyon nehéznek bizonyult. Fél évszázaddal később Jeffrey Lagarias az [7]-ben megemlíti Erdős Pál gyászos megjegyzését a probléma megoldhatatlanságára vonatkozóan: „A matematika nem biztos, hogy készen áll az ilyen problémákra”, majd csaknem negyedszázaddal később, 2010-ben ő maga kijelentette, hogy a sejtés „egy rendkívül nehéz probléma, amely a mai matematika számára teljesen elérhetetlen”[8].

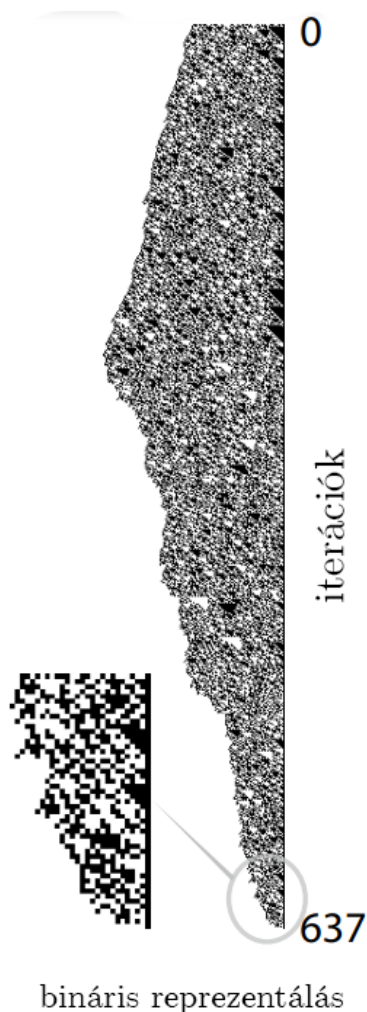
A problémát vizsgáló matematikusok általános véleménye az, hogy a sejtés igaz, mert alátámasztják valószínűségi heurisztikák[7, „A heuristic argument” fejezet.] és a megállási idő[9] érve. Terence Tao 2019-ben logaritmikus sűrűség segítségével kimutatta, hogy szinte minden Collatz-pálya a kiindulási pont bármely adott függvénye alá süllyed, feltéve, hogy ez a függvény a végtelenségig divergál, függetlenül attól, hogy milyen lassan[10]. Erős kísérleti bizonyítékok is vannak a közelmúltban[11], a sejtést minden kiindulási értékre ellenőrizték $2^{68}2,95 \times 10^{20}$ -ig.

A sejtésnek sok más ekvivalens megfogalmazása is létezik, amelyek a Collatz-függvényt enyhén módosított formában veszik figyelembe, mint például paritásorozat[9], mátrix értelmezések[12] vagy címkerendszer[13] stb. Egy másik megfogalmazás, amely releváns a kutatásunkkal kapcsolatosan, egy absztrakt gép, amely a Collatz iterációkat kettes, vagyis bináris számrendszerben számítja ki, bitsorozatot kezelve. A gép a következő három lépést hajtja végre bármely pozitív páratlan számon, amíg csak egy 1 bites marad:

1. 1 bit hozzáfűzése binárisan x -hez legkisebb helyiértékű bitként (jobb oldali hozzáfűzés), ami $2x + 1$ -t eredményez;
2. Az első lépés eredményét adjuk hozzá az eredeti számhoz bináris összeadással, amiből kapjuk, hogy $2x + 1 + x = 3x + 1$;

3. Távolítsuk el az összes nullát jobb oldalról, azaz osszuk folyamatosan 2-vel, amíg az eredmény páratlan lesz.

Az 2.1. ábra a bináris leképezést mutatja, amely a fent leírt algoritmusnak a 2^{68} [11] alatti legnagyobb ismert útvonalára való alkalmazásából adódik. A térképek kezdetben kibővülnek, és a 246. iterációnál elérik a 135 bites hosszúságot, de végül 637 lépés után 1 bitesre zuhannak.



2.1. ábra. Az iterátumok alakulása a $x = 274133054632352106267$ kezdőérték esetén, amely a legnagyobb ismert útvonal 2^{68} [11] alatt. A bal oldalon, kinagyítva, az utolsó 50 iteráció látható.

A bináris reprezentáció új utakat tesz lehetővé a probléma tanulmányozásában. Például ezen ábrázolási mód tulajdonságai egy másik perspektívát kínálnak

a Collatz-sejtés azon elégséges feltételeire vonatkozóan, hogy igaz legyen[14, 15]. Ezenkívül a [16]-ben a szerzők a sejtautomaták elméletéhez és a diszkrét dinamikus rendszerek tanulmányozásához kapcsolódó geometriai megközelítést alkalmaztak bizonyos kiindulási értékosztályok iterációinak általános periodikus és kaotikus mintázatainak feltárására. Később a [17]-ben bemutatott két egyszerű sejtautomatát, amelyek a Collatz iterált térképét utánozzák. Ezek az eredmények azt mutatják, hogy a bináris reprezentáció olyan mintákat és struktúrákat tárhat fel, amelyek további vizsgálatot érdemelnek.

A bináris reprezentáció lehetővé teszi a probléma bináris optimalizálási perspektívából való megközelítését is, mely javaslatot legjobb tudomásunk szerint még nem vizsgáltak a szakirodalomban. Itt bináris optimalizálási módszerek segítségével szeretnénk megtalálni azokat a bitsorokat (kezdőértékeket), amelyek olyan iterációkat eredményeznek, amelyek hossza jelentősen meghaladja a kezdőérték hosszát. Legyen $M(b_x)$ a bináris leképezés, amely a fent vázolt 3-lépéses bináris Collatz-folyamat alkalmazásából származik, b_x -tól kezdve, és $M_i(b_x)$ jelölje a térkép i -edik iterációját. Ekkor a maximalizálni kívánt célfüggvény egyszerűen a térkép leghosszabb iterációjának hossza és a kezdő hossz közötti arány, mínusz 1, hogy a kiindulási pont/alap 0 legyen. Formálisan egy n iterációt tartalmazó bináris térkép:

$$f(b_x) = \frac{\max(|M_1(b_x)|, |M_1(b_x)|, \dots, |M_n(b_x)|)}{|M_0(b_x)|} - 1 \quad (2.1)$$

A Collatz-sejtést alapvetően két módon lehet megoldani: ellenpéldát találni egy véges ciklus vagy egy végtelenbe tartó pálya formájában. Mivel bebizonyosodott, hogy nincs 186 milliárdnál rövidebb ciklus (pontosabban 186.265.759.595), kivéve a $4 \rightarrow 2 \rightarrow 1 \rightarrow 4 \dots$ ciklust, így tanulmányunkban elhanyagolható a sejtés ezen részét elemezni. Dolgozatunkban a megoldás második részére koncentrálunk, azaz, ha létezik olyan x , amelyre $f(x_{bin})$ divergál.

Az 2.1. ábrán bemutatott bináris térkép esetén $f(b_x) = 135/68 - 1 = 0,98529411764$, ami azt jelenti, hogy a folyamat körülbelül 98,53%-kal növelte a kiindulási hosszt (a kezdeti hossz csaknem megkétszereződött).

Míg egy divergáló ($f(b_x) = \infty$) kezdőérték megtalálása megoldaná a sejtést, egy ilyen felfedezés esélye rendkívül alacsony. Lagarias és Weiss[18] véletlenszerű sétákra vonatkozó nagy eltérés elméletén alapuló előrejelzések szerint még a $f(b_x) > 1$ értékkel rendelkező bemeneteket is nagyon nehéz megtalálni. Ehelyett az új megközelítéssel azt szeretnénk feltárni és megérteni, hogy a sikeres keresési algoritmusok milyen struktúrákat és mintákat hasznosíthatnak, és hogyan kapcsolódnak ezek a Collatz-problémához.

A fitness tájakat 1932-ben javasolták[19] a biológiai evolúció megértésének elvont fogalmaként, majd később az evolúciós algoritmusok viselkedésének magyarázatára használták. Az elmúlt tíz évben a fitness tájajelemezés rohamosan fejlődött, egy nagyrészt elméleti ötletből mára az optimalizálás általános, újabban pedig a gépi

tanulásban is alkalmazott gyakorlati eszközzé fejlődött. Ezzel a kibővített hatókörrel új típusú tájak jelentek meg, mint például a többcélú tájak (*Multiobjective Fitness Landscapes*), a szabálysértési tájak (*Violation Landscapes*), a dinamikus és kapcsolt tájak (*Dynamic and Coupled Fitness Landscapes*) és a hibatájak (*Error Landscapes*). Katherine Mary Mala 2021-ben megjelent cikke[20] egy 2013-as fitnesztájakkal kapcsolatos felmérésének folytatása[21], és további 11 tájelemzési technikát sorol fel. A cikk egy felmérést is tartalmaz a tájelemzés alkalmazhatóságáról az összetett problémák megértésére és az algoritmusok viselkedésének magyarázatára, valamint ezen algoritmusok teljesítményének előrejelzésére és az automatizált algoritmus-konfigurációra és -kiválasztásra.

A 2013-as cikkhez képest röpké 10 év alatt tizenegy új technikát fejlesztettek ki fitnesztájak elemzésére, pontosabban:

- Helyi optimum hálózatok (*Local optima networks, LON*), Ochoa és társai javasolták 2008-ban[22]. A LON a globális struktúrát képviselő keresési tér gráfalapú absztrakciója, ahol a minden csomópontja helyi optimumot jelent és a csomópontok közötti élek az optimumok medencéinek szomszédosságát képviselik (az egyik helyi optimumból a másikra való áttérés lehetősége).
- Feltáró tájelemzés (*Exploratory landscape analysis, ELA*), Mersmann és társai javasolták 2011-ben[23]. Véletlenszerű megoldásokból vett kis minta alapján az alacsony szintű jellemzők hat osztályát határozzák meg: (1) konvexitás, (2) y-eloszlás, (3) szintkészlet, (4) meta-modell, (5) helyi keresés és (6) görbület. A jellemzők olyan attribútumok becslései, mint például a célfüggvény lineáriságának valószínűsége, a függvényértékek eloszlásának ferdesége, az illesztett metamodellek pontossága, a lokális kereséssel azonosított lokális optimumok száma, a becsült numerikus gradiens stb.
- Hosszskála-eloszlás (*Length scale distribution*), 2012-ben javasolta R. Morgan és M. Gallagher[24]. A keresési területen egy véletlenszerű Levy-séta során kapott megoldások mintája alapján a hosszskála (az alkalmasság abszolút különbsége a távolság a térben) a mintában szereplő minden egyes megoldáspárra számítandó. A hosszskála-eloszlást a hosszskálák valószínűségi sűrűségfüggvényeként határozzuk meg, és magsűrűség-becsléssel becsüljük meg.
- Kodinamikus tájmérések (*Codynamic landscape measures*), 2014-ben javasolta H. Richter[25]. Adott egy halmaz a keresési térben lévő pontokból és két összekapcsolt, kodinamikus tájkép: az objektív táj (a probléma fitnesztája) és a szubjektív táj (ahogyan érzékeli a koevolúció a problémát) tájmérései a két táj különbségeinek számszerűsítésére szolgálnak.
- Az elválaszthatóság mértéke (*Degree of separability*), Caraffini és társai 2014-ben javasolta[26].

- Korlátozott tájmutatók (*Constrained landscape metrics*), 2015-ben javasolta Malan és társai[27]. Adott egy progresszív véletlenszerű sétán alapuló megoldássorozat, a hozzá tartozó alkalmassággal és a kényszermegsértés mértékével minden megoldáshoz. A következőket becsüljük: (1) a megfelelő megoldások aránya a keresési térben (FsR), (2), a megfelelő területek közötti szakadás mértéke, számszerűsítve a megfelelő határátlépések arányával ($RFB_{\times}$), (3) az fitness és a szabálysértés közötti összefüggés (FVC), és (4) a fitness szempontjából magas, korlátok megsértése szempontjából alacsony típusú megoldások aránya, két mérőszám formájában: a megoldások aránya a felső 50%-os percentilisben és 20%-os percentilisben.
- Helyi tájjellemzők halmaza (*Bag of local landscape features*), 2016-ban javasolta S. Shirakawa és T. Nagao[28].
- Maximális entrópiás episztázis (*Maximum entropic epistasis, MEE*), 2017-ben javasolta Sun és társai[29]. Minden egyes x_i , x_j döntési változópár esetén a közvetlen interakciók (IM_d) interakciós mátrixát úgy építjük fel, hogy kiszámítjuk az x_j és a célfüggvény becsült parciális deriváltja x_i esetén közötti maximális információs együtthatót. Az IM_d -t ezután egy interakciós gráf felépítésére használjuk, amely az erősen kapcsolódó összetevőket leképezi a közvetett kölcsönhatások azonosítása érdekében.
- Népeségfejlődési mérőszámok (*Population evolvability metrics*), 2018-ban javasolta Wang és társai[30]. Adott a megoldások populációja és a szomszédok halmaza (az algoritmus egy iterációjából), két mérőszámot definiálunk: (1) epp annak a valószínűsége, hogy egy populáció továbbfejlődik, és a becslés a szomszédok arányának kiszámításával történik, amelyek fittebbek, mint a jelenlegi populáció legjobb megoldása, és (2) az eap a populáció evolúciós képessége, amely a szomszédok abszolút fitness javulásával növekszik és a populáció fitness diverzitásával csökken. Az eredmény az evp , amely az előző kettő szorzata $[0, +\infty]$ tartományon.
- Helyi többcélú tájjellemzők (*Local multiobjective landscape features*), Liefvooghe és társai javasolták 2019-ben[31]. Véletlenszerű séták és adaptív séták során kapott megoldások soroza alapján számítja ki a séta jellemzőit, valamint a megoldások szomszédosságából, azzal a céllal, hogy javítsa a szomszédok dominanciát és hipervolumenitását.
- Veszteség-gradiens felhők (*Loss-gradient clouds*), 2020-ban javasolta Bosman és társai[32]. A veszteségértékek és a gradiensértékek mintáját számos véletlenszerű, progresszív gradiens-séta alapján kapjuk. A minta stacionárius pontjait 3 kategóriába soroljuk (lokális minimumok, lokális maximumok vagy nyeregponatok) a Hess-mátrix sajátértékeiből származó lokális görbület alapján

kapjuk. A séta közbeni stagnáló szekvenciákat is kiszámíthatjuk a hiba simításának eltéréséből. Két mennyiséget mérünk: (1) a stagnálás megfigyelésének átlagos számát és (2) a stagnáló sorozat átlagos hosszát.

A LON-ok és az ELA-k jól megalapozottak, és széles körben használatosak tájélemzést magában foglaló tanulmányokban, ahol a LON-okat többnyire a diszkrét optimalizálásban, az ELA-t pedig a numerikus optimalizálásban alkalmazzák. A tájélemzéshez javasolt új technikák közül számos nagyon specifikus optimalizálási kontextusok esetén alkalmazható, mint például koevolúció, korlátozott optimalizálás, multiobjektív optimalizálás és neurális hálózatok tanítása.

Erdős Pál a Collatz-sejtésről azt nyilatkozta: „A matematika még nem áll készen az ilyen problémákra”, majd nyereséget ígért a helyes megoldónak. Vajon mi készítette arra, hogy ilyen ajánlatot tegyen? Miért olyan nehéz megoldani a Collatz-sejtést?

A Collatz-sejtés megoldásához egy mintát kell találni n és $n+1$ prímfaktorizációja között. Az alapkérdés az, hogy milyen módon befolyásolja az n prím tényezőkre bontása $n+1$ prím tényezőkre való bontását?

Nagyon nehéznek tűnik erre a kérdésre válaszokat találni. Ez részt azért van így, mert n prímtényezőinek kis változása (mondjuk prímszámmal való szorzás) hatalmas változást okozhat $n+1$ prím faktoriázációjában (talán teljesen más prímtényezőket hozhat be). Ezért csábító az 1 hozzáadását a prímtényezők véletlenszerű megkeverésének tekinteni.

A legszembevetőbb a Collatz-sejtésben az, hogy úgy tűnik, mély kijelentést tesz az n és az $n+1$ prímtényezői finom kapcsolatáról. Egy Collatz-iteráció három lépésből áll; amelyek közül kettő „kis jelentőséggel bír” a prímfaktorizáció szempontjából (a 3-mal való szorzás és a 2-vel való osztás kis mértékben befolyásolja a prímfaktorizációt), a harmadik pedig az 1 hozzáadása. A matematikusok ilyen okok miatt gyanítják, hogy a Collatz-sejtés megoldása új távlatokat nyit meg, és új, fontos technikák kifejlesztését segítheti elő a számelmélet terén.

3. fejezet

Kutatás módszertana

Eric Roosendaal, majd később mások, többek között Tomás Oliveira e Silva, David Bařina közzétette az eddig talált pálya rekordokat (*path records*)¹, azaz az eddig legnagyobb növekedést elért számokat. Táblázatuk felsorolja a 2^{68} alatti összes ilyen rekordot. Az eredmények tárgyalása szekcióban bizonyos ilyen rekoordokat felhasználunk ábrák során.

Kutatásunk a következő ütemtervet követte:

- Első lépésben kerestük azokat a 16 és 32 bit közötti hosszúságú b_x bitsorokat, amelyekre a 2.1. egyenletet alkalmazva az eredmény átlép egy magas küszöböt (0.4-0.5), ezáltal garantálva, hogy ezek a pontok lokális optimumok.
- A kapott pontok fitnesztávolsági összefüggéseinek[1] és a megoldássűrűség eloszlásának[2] kiszámítása és elemzése, ezáltal megkapva az optimalizálási probléma nehézségének kezdeti mérőszámát.
- A pontok közötti Hamming-távolság kiszámítása, hogy 16 helyi optimum hálózatot kapjunk, minden kezdeti hosszhoz egyet.
- LON-metrikákat és vizualizációs technikák segítségével elemezzük és jellemezzük a hálózatokban a globális optimum elérésének nehézségeit.

A következőekben a használt metrikákat és technikákat mutatjuk be részletebben.

¹A táblázat és a keresés részletes leírása megtalálható a következő link alatt: <http://www.ericr.nl/wondrous/pathrecs.html>

3.1. Fitnesztáj analízis

3.1.1. Fitnesztájak

A biológiai evolúciós optimalizálás dinamikájának szemléltetésére bevezetett fitnesztáj (angolul *fitness landscape*) koncepciója hasznosnak bizonyult az evolúcióelméletben, mind pedig a kombinatorikus optimalizáló algoritmusok viselkedésének megértésében, és segíthet a teljesítményük előrejelzésében. A keresési teret, vagyis az összes (potenciális) megoldás halmazát tájként tekintve egy heurisztikus algoritmus úgy is felfogható, mint a táj legmagasabb csúcsát kereső hegymászó algoritmus (angolul *hill climbing algorithm*). Ez egy helyi keresési algoritmus, amely folyamatosan felfelé mozog („növekszik”), amíg el nem éri a legjobb megoldást. A „csúcs” elérésekor véget ér az algoritmus. Egy nem optimális, véletlenszerű állapotból indul ki (a hegy „alja”), majd iteratíván megkísérel egy mind jobb megoldást találni minden lépésben, amíg egy bizonyos előfeltétel teljesül. Ennek az előfeltételnek az alapja a heurisztikus függvény. Az iteráció állapotának folyamatos javításának folyamatát mászásnak nevezzük. Ez megmagyarázza, miért nevezik az algoritmust hegymászó algoritmusnak. Egy pont magassága a keresési térben tükrözi a megoldás alkalmasságát/fitnesz értékét.

Formálisabban, egy adott kombinatorikus optimalizálási probléma fitnesztája (S, f, d) a következőkből áll: pontok (megoldások) halmazából (S), egy fitneszfüggvényből ($f : S \rightarrow \mathbb{R}$), amely valós értékű fitnesz értéket rendel a pontok mindegyikéhez, és egy távolságmetrikából (d), amely meghatározza a táj szerkezetét. A fitnesz táj tehát egy csúcs- és élhalmazzal rendelkező gráfként értelmezhető: $G_L = (V, E)$, $V = S$, $E = (s, s') \in S \times S \mid d(s, s') = d_{\min}$, ahol $d_{\min}$ a keresési tér két pontja közötti minimális távolságot jelöli. A táj átmérője ($\text{diam} G_L$) egy másik fontos tulajdonság: ez a keresési tér pontjai közötti maximális távolság (avagy $d_{\max}$).

Bináris kódolású problémák esetén ($S = \{0,1\}^n$) a távolság mértéke pedig a bitkarakterláncok közötti Hamming-távolság. A minimális távolság, $d_{\min}$ 1 (egy bit eltérő értékű), a maximális távolság, $d_{\max}$ pedig n .

A fitnesztáj számos tulajdonsága befolyásolja a heurisztikus optimalizáló algoritmusok teljesítményét. A táj optimumához számos olyan jellemző kapcsolódik, amely meghatározza a fitnesztáj szerkezetét, mint például a modalitás, az episztázis, az egyenetlenség és a megtévesztés/félrevezethetőség szempontjából. A modalitás a keresési térben létező optimumok számára utal, sűrűségüket méri, és összefügg a globális optimum megtalálásának nehézségével. Az episztázis arra utal, hogy a genotípus fitnessse függ több gén kölcsönhatásától. Mind a magas modalitás, mind az episztázis egyenetlenebb fitnesztájhoz vezet. Az ilyen keresési terekben nehezebb keresni, mint egy simább, alacsony episztázisú és modalitású táj esetén[33].

Számos különböző technikát fejlesztettek ki a fitnesztájak jellemzőivel történő

elemzésére.

3.1.2. Fitnesztájak globális jellemzői

A lokális (helyi) módszerek kiegészítéseként - amelyek elsősorban az optimalizálási algoritmusokra gyakorolt azonnali hatásokat elemzik lépésről lépésre -, a következő globális módszerek segíthetnek átfogóbb képet kapni a fitness tájakról.

Fitnesztávolság-korrelációk

A fitnesztávolság-korrelációk (angolul *fitness distance correlation*, FDC) egy ismert és sűrűn használt metrika[34, 35, 36], amelyet először Jones és társai[1] javasoltak a genetikus algoritmusok (angolul genetic algorithms, GA) performanciájának tesztelésére:

$$FDC = \frac{C_{fd}}{S_f S_d} \quad (3.1)$$

ahol $C_{fd} = \frac{1}{n} \sum_{i=1}^n (f_i - \tilde{f})(d_i - \tilde{d})$, $s_f s_d$, $\tilde{f}, \tilde{d}$ f, d standard szórás és átlag, f pedig a fitnessz érték. d a távolság a potenciális megoldás és a globális optimum között. Valójában, FDC az f és d közötti korreláció. Ha az $FDC < 0$, a fitnessz érték növekedik, ahogyan az algoritmus közeledik a globális optimumhoz. Másrészt, ha az $FDC > 0$, a fitnessz érték csökken a globális optimumhoz tartva, ami megtéveszti a keresést a globális optimumtól félrevezetve. Tehát az FDC tükrözi a globális struktúrát, fejleszthetőséget és megtévesztést. Az FDC megvalósítása azonban megköveteli a globális optimum ismeretét, ami miatt kevésbé hasznossá válik az FDC . Sőt, kimutatták, hogy a különböző fitnessz tájakon nagyon hasonló a fitnesztávolság-korreláció[37], ami talán a torz mintáknak köszönhető. A FDC számítási összetettsége $O(n)$, ahol n a minta mérete.

Szórásmutatók

A szórásmutatók (angolul *dispersion metrics*, DM) az átlagos távolságot mérik jó minőségű megoldaspárok között[38]:

$$DISP_{\alpha\%} = \frac{1}{(\alpha n)(\alpha n - 1)} \sum_{i=1}^{\alpha n - 1} \sum_{j=i+1}^{\alpha n} \|x_i - x_j\| \quad (3.2)$$

Ahol a a legnagyobb $\alpha\%$ populációt jelenti. Ezért a DM számítási bonyolultsága $O(\alpha n^2)$. Ha a szóródás növekszik, amikor a mintákat a keresési tér jobb régióira korlátozzuk, akkor a táj gyenge globális szerkezetű, és a probléma nehezen megoldható. Ezért a DM globális struktúrát, fejleszthetőséget és megtévesztést tükröz.

A DM kimutatta, hogy a CMA-ES algoritmus miért nem hatékony multimodális problémák esetén[38]. A diszperziós mérőszámok korlátja, hogy a dimenziók jelentősen megnövekedésekor $1/\sqrt{6}$ -hoz konvergál[39].

3.1.3. Fitnesztájak lokális jellemzői

A fitnesztáj elemzésének ezen formája a közvetlen szerkezetet vizsgálja, amellyel az optimalizáló folyamat „találkozik”. Ezen lokális értékek az azonnali fitness változásokat hivatottak mérni, és nagyon gyakran a véletlenszerű sétával kialakított „útvonalakból” épülnek fel.

Ezek a technikák két csoportra oszthatók, amelyek statisztikai elemzést és információelemzést biztosítanak. A statisztikai elemzési technikákban számos korrelációs mérőszámot javasolnak a táj szerkezetének jellemzésére.

Ebben a dolgozatban a tájban található lokális optimumok (csúcsok) számára, a csúcsok eloszlására a keresési térben és a táj egyenetlenségére, azaz a keresési tér szomszédos pontjai közötti összefüggésre koncentrálunk. A táj egyenetlenségének mérésére és a csúcsok eloszlásának elemzésére statisztikai módszereket javasoltak, de a lokális optimumok számát általában nem lehet meghatározni. A táj egyenetlensége azonban szorosan összefügg a keresési területen belüli lokális optimumok számával. Egy fitnesztájról azt mondjuk, hogy egyenetlen, ha a táj sok csúcsból áll, és ha alacsony a korreláció a szomszédos pontok között. A Weinberger[40] által javasolt autokorrelációs függvények[41, 42] a fitness táj egyenetlenségét mérik. Weinberger azt javasolta, hogy végezzünk véletlenszerű sétákat a táj korrelációs szerkezetének vizsgálatára. A későbbiekben részletesebben kitérünk a véletlenszerű séták fogalmára.

A gyakorlatban használt másik korrelációs mérőszám a korrelációs hossz. Meghatározza azt a távolságot, amelyen túl két csúcs nem korrelál. Ennek a hosszúságnak a nagysága a táj simaságát jelzi. A nagyobb korrelációs hossz simább tájat, míg a kisebb egy durvább, egyenetlenebb tájat jelez. A fitnesztájak elemzését véletlenszerű séta során kapott fitnessértékek sorozatán végzik el. Az alábbi képleteket használjuk a fitnesztáj statisztikai mérőszámainak becslésére[43]:

- a 3.3. képlet az $R(\delta)$ autokorrelációs függvény;
- a 3.4. képlet pedig az autokorrelációs távolság.

$$R(\delta) = \frac{E[(f_i - \bar{f})(f_{i+\delta} - \bar{f})]}{\sigma f^2} \quad (3.3)$$

ahol $\bar{f}$ és σf^2 a véletlenszerű séta sorozatának átlaga és szórása, $E(x)$ pedig x várható értéke.

$$\tau = \frac{1}{\ln(R(1))} \quad (3.4)$$

Jellemzően $R(1)$, az egy lépés utáni autokorrelációt veszik a táj egyenetlenségének mérésére. Ezen kívül egy statisztikailag szignifikánsan korrelált ponthoz az átlagos maximum távolságot is lehet használni, autokorrelációs távolságnak nevezik a metrikát.

Véletlenszerű séták, Markov-lánc

Az irányított gráfon végzett véletlenszerű séta csúcsok sorozatából áll, amelyeket egy kezdőcsúcsból generálunk úgy, hogy kiválasztunk egy élt, áthaladunk az élen egy új csúcsra, és megismételjük a folyamatot.

A véletlenszerű séták (angolul *random walks*) fontos alkalmazási területe a világhálón található oldalak stacionárius valószínűsége alapján történő pagerank meghatározása[44]. A statisztikai irodalomban korábban kidolgoztak egy ekvivalens, Markov-láncnak (angolul *Markov chain*) nevezett koncepciót. Nevét Andrej Markov orosz matematikusról kapta. A Markov-láncnak véges állapothalmaza van. Minden x és y állapotpárhoz tartozik egy valószínűségi átmenet P_{xy} , vagyis az x állapotból az y állapotba, ahol minden x esetén $\sum_y P_{xy} = 1$. Egy véletlenszerű séta a Markov-láncban valamelyik állapotnál kezdődik. Ha az x állapotban van, a következő y állapotot véletlenszerűen választjuk ki P_{xy} valószínűséggel. Egy Markov-lánc ábrázolható irányított gráfként, amelynek csúcsa az egyes állapotokat reprezentálja, és egy P_{xy} súlyú él köti össze az x csúcsot az y -nal. Azt mondjuk, hogy a Markov-lánc összefüggő, ha az irányított gráf erősen összefüggő. Vagyis ha minden csúcsból minden másik csúcsba van irányított él. A P_{xy} -ből álló P mátrixot a lánc átmeneti valószínűségi mátrixának nevezzük. A véletlenszerű séta és a Markov-lánc kifejezéseket felcserélhetően használják.

3.1.4. Lokális optimum hálózatok

A fitness tájak egyik jellemzésére használt technika a lokális optimum hálózatok[3], amelyeket diszkrét keresési terekre terveztek. A lokális optimum hálózatok (angolul *local optima networks*, rövidítve LON) a globális struktúrát reprezentáló keresési tér gráf alapú absztrakciója, ahol a LON minden csomópontja egy lokális optimum, a csomópontok közötti élek pedig az optimum medencéinek szomszédosságát jelentik (a keresés egyik lokális optimumból a másikba való átmenet lehetősége). Tehát a tájak globális szerkezetére összpontosít, amelyek befolyással bírnak a keresés hatékonyságára. Különböző problémák keresési terének globális struktúráját kigenerálva, segíthet megérteni, hogy egy-egy algoritmus miért sikeres, illetve miért vall kudarcot az adott probléma megoldásában.

3.2. UMDA - Egyváltozós peremeloszlás algoritmus

Az eloszlásbecslési algoritmusok (angolul *estimation-of-distribution algorithms*, EDA) véletlenszerű keresési heurisztika, amely iteratív módon fejleszti ki a keresési tér valószínűségi modelljét. Általában az egyenletes eloszlástól kezdve az EDA mintákat vesz a jelenlegi modelljéből, majd úgy állítja be, hogy a következő iterációban nagyobb valószínűséggel jöjjenek létre jobb megoldások. Ez a finomítási módszer fokozatosan jobb megoldásokhoz vezet, és számos gyakorlati probléma esetén jól teljesít, gyakran felülmúlva a versengő megközelítéseket [45, 46].

Az UMDA (angolul *univariate marginal distribution algorithm*) [47] egy egyszerű EDA, amely megbecsüli a kiválasztott populáció határvalószínűségét. A közelmúltban kimutatták, hogy elég nagy populáció mellett a módszer jól megbirkózik a megtévesztéssel és az episztázisokkal [48]. Ezért az UMDA jó első jelölt a javasolt optimalizálási probléma empirikus feltárására.

Az UMDA valószínűségi modelljeként a valószínűségek (frekvenciák) n hosszúságú p (frekvenciavektora) vektorát használja. Ez a vektor egy egyedi $x \in \{0, 1\}^n$ mintavételére szolgál, amelyet $x \text{ minta}(p)$ -ként jelölünk, úgy, hogy minden $y \in \{0, 1\}^n$.

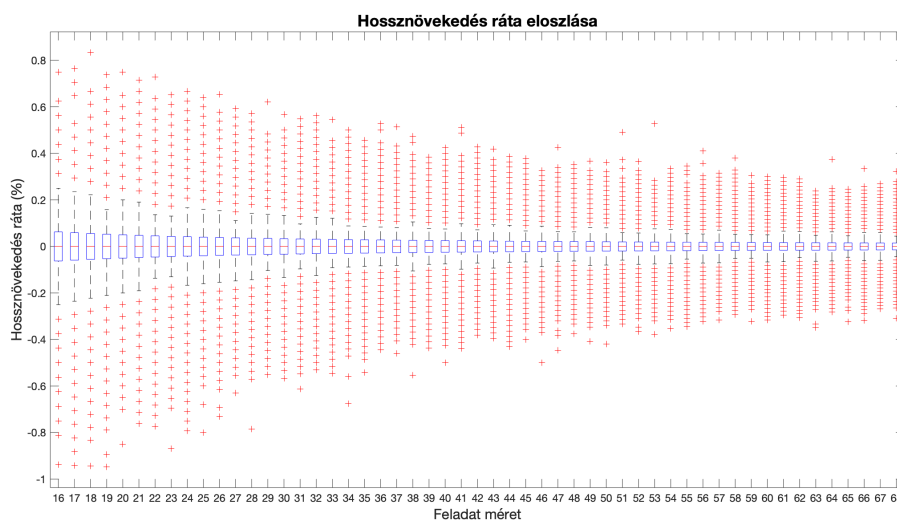
$$\Pr[x = y] = \prod_{\substack{i=1 \\ y_i=1}}^n p_i \prod_{\substack{i=1: \\ y_i=0}}^n (1 - p_i) \quad (3.5)$$

Az UMDA iteratív módon frissíti ezt a vektort a következő módon: először is λ -nyi egyedet mintavételez. Ezután ezek között a λ egyedek közül kiválasztunk μ -nyi egyedet, a legmagasabb fitness értékekkel rendelkezőket, és minden $i \in [n]$ indexnél a p_i frekvenciáját a μ legjobb egyedek között i pozícióban lévő 1-ek relatív számára állítjuk be. Végül, ha a p_i frekvencia $1/n$ alatt van, $1/n$ -ra növelik, és ehhez hasonlóan az $1 - 1/n$ felettieket is $1 - 1/n$ -re állítják be. A frekvenciák $[\frac{1}{n}, 1 - \frac{1}{n}]$ tartományba való bezárása elkerüli, hogy a 0 vagy 1 szélsőértéknél ragadjanak. Végül jelöljük a $t \in \mathbb{N}$ iteráció frekvenciavektorát $p(t)$ -vel.

4. fejezet

Eredmények tárgyalása

Ebben a fejezetben az eddigi vizsgálatok leírását és eredményeit foglaljuk össze. Ahogyan az előző fejezet elején pontokba szedve leírtuk, a kutatást azzal kezdtük, hogy megkerestük mindazon pontokat, avagy a 16 és 32 bit közötti hosszúságú b_x bitsorokat, amelyek fitness értéke - vagyis a 2.1. egyenletetbe behelyettesítve a kapott eredmény - átlépett egy bizonyos küszöbszámot, ezáltal garantálva, hogy ezek a pontok lokális optimumok.

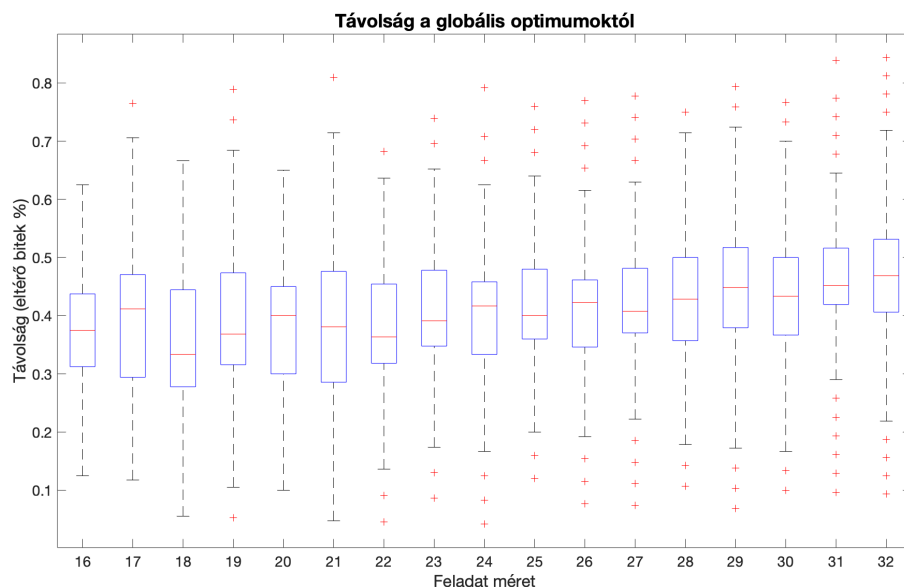


4.1. ábra. A fitness értékek eloszlása

Ezután különböző méréseket, elemzéseket végeztünk a megszűrt pontokon. A 4.1. ábrán láthatjuk a hossznövekedés ráta eloszlását a feladat méretének függvényében - minél nagyobb a feladat mérete, a pontok nagyrésze kevés növekedést ér el (körülbelül kevesebb, mint 10%-os hossznövekedés figyelhető meg), emellett a szórás

és a kiugró értékek is egyre kisebbek.

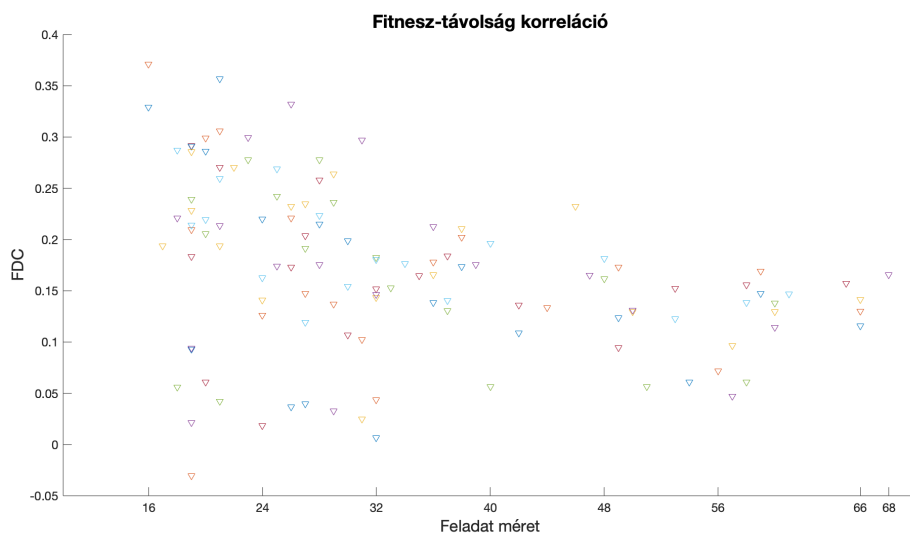
A 4.2. ábrán is hasonlóképp láthatjuk a pontok távolságát a globális optimumtól a feladat méret és a távolság százalékos arányában. Távolság alatt a Hamming-távolságot értjük, amely két azonos hosszúságú bináris szó eltérő bitjeinek a számát jelenti. A grafikonon látható, hogy a feladat méretével megegyezően átlagosan a távolság is megnő. Tehát minél nagyobb a feladat, annál több a globális optimumtól Hamming-távolságban messze lévő lokális optimumjaink lesznek, vagyis egyre nehezebb kikerülni a lokális optimumok vonzáskörzetéből, hogy eljuthassuk a globális optimumhoz.



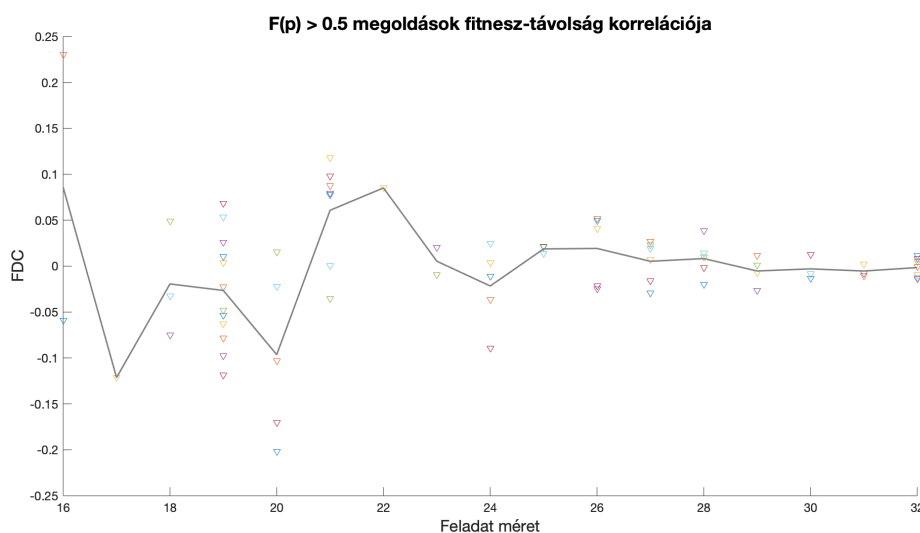
4.2. ábra. Lokális optimumok átlagos Hamming-távolsága a globális optimumtól százalékban kifejezve

A következő két ábrán, 4.3-4.4. ábrákon láthatjuk az előző fejezetben ismertetett fitness-távolság-korrelációkat szintén a feladat méretének tekintetében. A 32 feletti értékeket az előző fejezetben említett pálya rekordok segítségével számoltuk ki. A 4.4. ábrán csak a jó minőségű pontokat ábrázoltuk (ahol a pontok fitness értéke nagyobb, mint 0.5), az átlag -7.0346×10^{-4} , 0-hoz közeli szám, vagyis elmondhatjuk, hogy nincs korreláció.

Továbbiakban végeztünk statisztikai elemzéseket is, mint például a bitek valószínűségi eloszlása. A 4.5. ábrán látható, hogy minden egyes jó minőségű megoldás esetén (küszöbérték=0,4) a legelső bit 1-es, ez nem is olyan meglepő, hiszen a legelső jelentős számjegy mindenképp egyes lesz - az első jelentős számjegy definíciója az az első számjegy, amely nem 0. Ez azért fontos, mert bármilyen számot át lehet

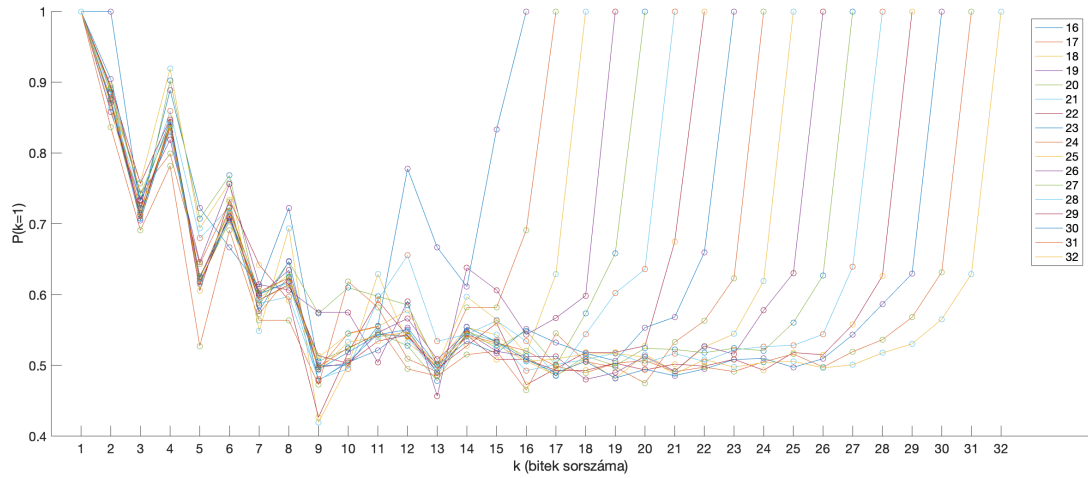


4.3. ábra. Fitness-távolság-korreláció



4.4. ábra. Fitness-távolság-korreláció - csak jó minőségű megoldások. Átlag = $-7.0346e-04$, nincs korreláció

írni 0-val kezdődőre. Például a 010 000010-ként is írható. Ami viszont érdekesebb az az, hogy az esetek 100%-ban az utolsó bit is 1-es, vagyis a lokális optimumjaink páratlan számok. érdekes, repcés mintát vesz fel az ábra, a 3., 5., 7., 9. sorszámú, helyiértékű bitek nagyobb valószínűséggel lesznek 0-sok, mint az előtte lévő számok.



4.5. ábra. Bitek valószínűségi eloszlása jó minőségű megoldások esetén ($F(p) > 0,4$)

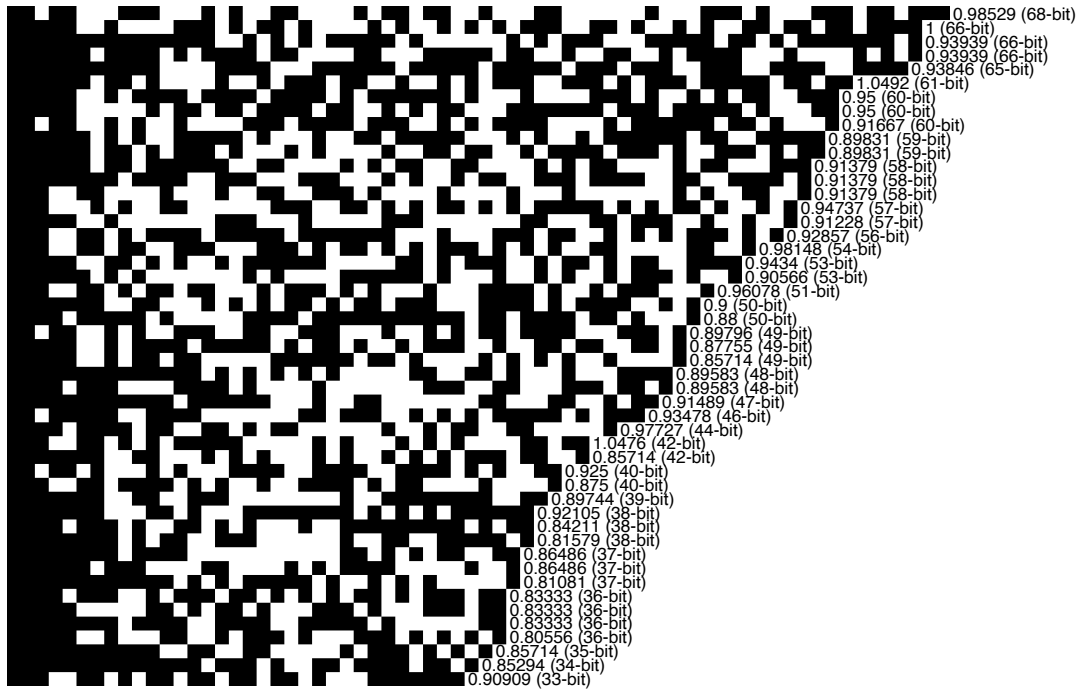
Az előző fejezetben ismertett pálya rekordokat felhasználva, binárisan ábrázoltuk a legjobb eredményeket 33 és 68 bit között, ezt láthatjuk a 4.6. ábrán. Az egyesek és nullások száma szinte fele arányban van jelen a megoldások esetén (fekete szín jelöli az 1-est, fehér a 0-ás jelenlétét). Ugyanez megfigyelhető az általunk vizsgált tartományban is, a 4.7. ábrán láthatjuk 29-32 bites tartományban ábrázolva a legjobb 32 darab megoldást. Ez a megfigyelés későbbiekben fontos szerepet fog játszani az optimalizálásban.

A genetikus algoritmus (angolul *generic algorithm*, rövidítve GA) egy keresési heurisztika, amelyet rutinszerűen használnak optimalizálási és keresési problémák hasznos megoldásainak generálására. Azt megoldásokat generál az optimalizálási problémákra a természetes evolúció által inspirált technikák segítségével, mint például az öröklődés, a mutáció, a szelekció és keresztezés. A genetikai algoritmusok az egyik legjobb módja egy olyan probléma megoldásának, amelyről keveset tudunk. Genetikus algoritmusunk azonban hamar beragadt lokális optimumokban, mivel elég hamar megtanulta, hogy a bitszingünk első fele csupa egyesekből álljon (legtöbb lokális optimum esetén ezt áll fent, megfigyelhető akár a 4.7. ábrán a kevésbé jó megoldások esetén is).

4.1. A táj félrevezető jellegéről

Az előző fejezetben tárgyalt véletlen séták autokorrelációs függvénnel történő vizsgálata alapján kimutatható, hogy az optimumok nem korrelálnak egymással, azaz a kapott fitnessztájak erősen egyenetlenek. A 4.8. ábrán látható is 16, 20, 24, 28

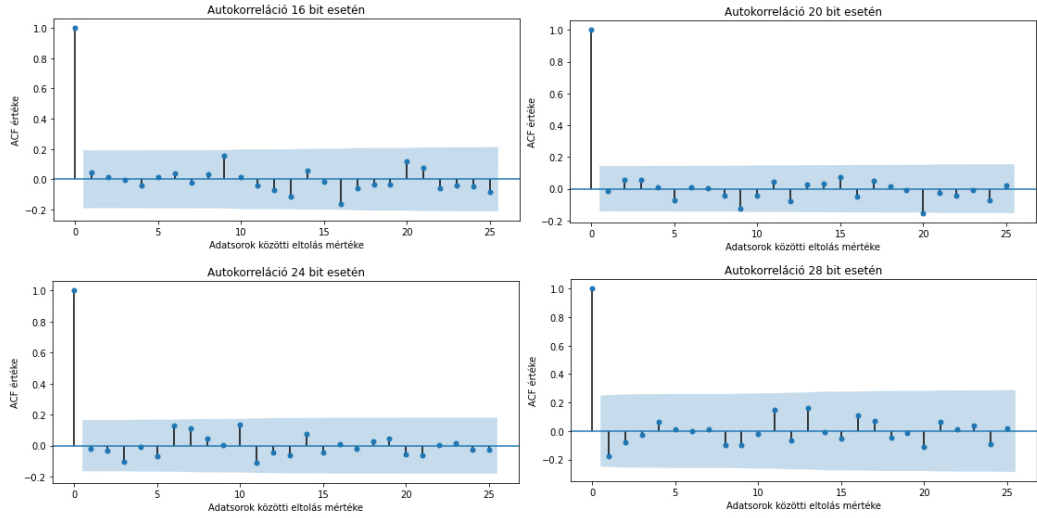
Pálya rekordok bináris ábrázolása



4.6. ábra. Az OEIS A006884 szekvenciájának pálya rekordjai binárisan ábrázolva (<https://oeis.org/A006884>)

bites számok esetén a korrelációs vizsgálat eredménye. Az x-tengelyen az eltolás van ábrázolva ($k \in \{1, \dots, 25\}$), a y-tengelyen pedig a Pearson-féle korrelációs együttható minden egyes eltolás esetén. A kék árnyalatú terület egy konfidenciaintervallum. Ha a pálcikák magassága ezen a tartományon kívül esik, az azt jelenti, hogy a korreláció statisztikailag szignifikáns. Esetünkben ez nem igaz, tehát a fitness értékek nem korrelálnak egymással.

Fitnesztájaink nemcsak erősen egyenetlenek, hanem ráadásul multimodálisak (32 bit esetén például 132 érték ér el 0,7%-os növekedést, a meghatározott küszöbértékünket pedig 108.030 érték múlta felül) és félrevezetőek. Ahogyan ez előző szekcióban is tárgyaltuk, rengeteg lokális optimum első fele kevés 0-ás bitet tartalmaz, így az algoritmus elég hamar megtanulja, hogy ezeket az értékeket 1-esekre állítsa be. Valójában ez megvezeti a keresést, és a globális optimumtól távolabbra viszi. A 4.7. ábrán láthattuk, hogy a legnagyobb hossznövekedéssel rendelkező értékek körülbelül egyforma arányban tartalmaznak 0-ásokat és 1-eseket. Kereső algoritmusunkat tehát módosítani kellett ennek függvényében, hogy ezt a félrevezető tendenciát ki tudjuk küszöbölni.



4.8. ábra. Bitek valószínűségi eloszlása jó minőségű megoldások esetén ($F(p) > 0,4$)

UMDA jól megbirkózik a megtévesztésekkel, illetve az episztázisokkal, mint az evolúciós algoritmusok.

Ezen megfigyeléseket felhasználva a továbbiakban ezekkel a módszerekkel vizsgáltuk tovább a problémát.

4.2. Klaszterezett UMDA

A továbbiakban klasztreztük a populációt n csoportba az első $n/2$ bit szerint. Számoltunk egy p_c valószínűségi modellt minden c klaszterre. Mindezt amiatt, hogy elkerüljük, hogy a genetikus algoritmusunk ne ragadjon bele egy lokális optimumba, azaz kiküszöböljük a főleg csupa 1-es kezdőbitekhez való konvergenciáját.

100 futtatás esetén a sikerarány 72%, ezzel szemben a „sima” UMDA 5%-ot ért el. A 4.9. ábrán látható a klaszterezett UMDA konvergencia-sebessége 32 bit esetén.

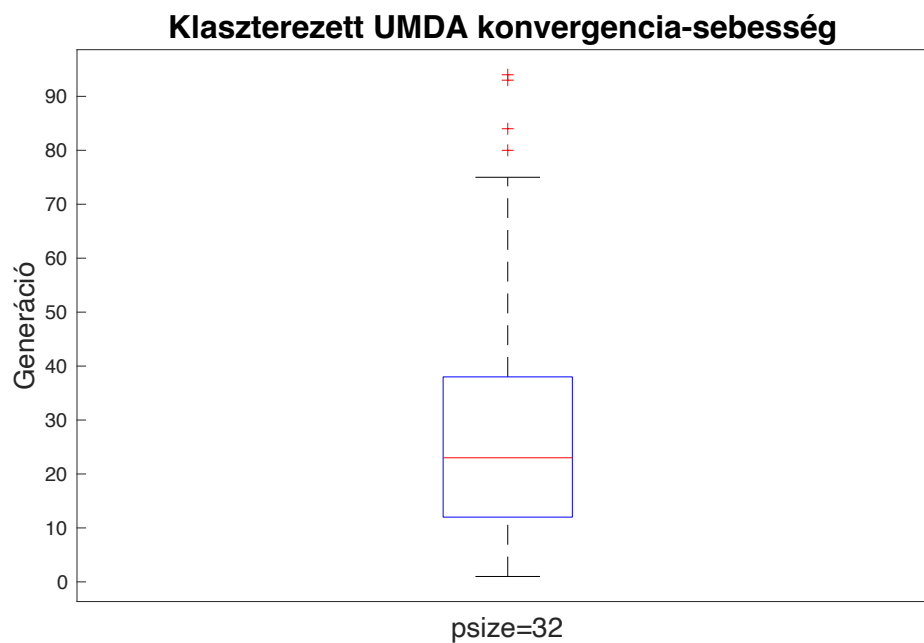
Az UMDA algoritmust menete a következő:

- 1: $t \leftarrow 0$
- 2: $p^{(t)} \leftarrow (\frac{1}{2})_{i \in [n]}$
- 3: **while** megállási feltétel nem teljesül **do**
- 4: **for** $i \in [\lambda]$ **do**
- 5: $x^{(i)} \sim \text{sample}(p^{(t)})$
- 6: **end for**

```

7:   $y^{(1)}, \dots, y^{(\mu)}$  jelölje a legjobb  $\mu$  egyedet  $x^{(1)}, \dots, x^{(\lambda)}$ -ből
8:  for  $i \in [n]$  do
9:     $p_i^{(t+1)} \leftarrow \frac{1}{\mu} \sum_{j=1}^{\mu} (y_i^{(j)})$ 
10: end for
11:  $p^{(t+1)}$ -et szűkítjük le az  $[\frac{1}{n}, 1 - \frac{1}{n}]$  intervallumra
12:  $t \leftarrow t + 1$ 

```



4.9. ábra. A klaszterezett UMDA konvergencia-sebessége 32 bit esetén

5. fejezet

Konklúzió

A fitness-táj elemzés hatékony technikákat kínál az evolúciós algoritmusokkal megközelített optimalizálási probléma nehézségének meghatározására. Számos különböző technikát és metrikákat (autokorreláció, korreláció hossza, információtartalom, információstabilitás stb.) fejlesztettek ki a fitness-tájak szerkezeti jellemzőinek elemzésére, amelyek befolyásolják az evolúciós algoritmusok teljesítményét. Ezek különösen lehetővé teszik a táj egyenetlenségének mérését, modalitásának és a lokális optimumok sokféleségének, a véletlenszerű séták rendszerességének mértékét.

A tanulmányunk során a 16 és 32 biten ábrázolt számok Collatz pályáit vizsgáltuk statisztikai megközelítéssel, bebizonyítottuk, hogy a belőlük kinyert fitness-tájak egyenetlenek, multimodálisak és félrevezetőek. A továbbiakban egy olyan kereső algoritmust fejlesztettünk ki, amelyek mindezen megfigyeléseket számba veszik és lényegesebben jobban teljesítenek, mint „sima” változataik.

A kutatás a továbbiakban újabb optimalizálási paraméterek keresésére koncentrálna, illetve a Collatz-probléma feltárása az ismert és megállapított határok felett: az algoritmusok felhasználása 68 bitnél hosszabb kezdeti értékű bitsorozatok keresésére.

Irodalom

- [1] Terry Jones, Stephanie Forrest és tsai. «Fitness distance correlation as a measure of problem difficulty for genetic algorithms.» *ICGA*. 95. köt. 1995, 184–192. old. (hiv. old. 2, 9, 11).
- [2] Helge Rosé, Werner Ebeling és Torsten Asselmeyer. «The density of states—a measure of the difficulty of optimisation problems». *International Conference on Parallel Problem Solving from Nature*. Springer. 1996, 208–217. old. (hiv. old. 2, 9).
- [3] Gabriela Ochoa, Sébastien Verel, Fabio Daolio és Marco Tomassini. «Local optima networks: A new model of combinatorial fitness landscapes». *Recent advances in the theory and application of fitness landscapes*. Springer, 2014, 233–262. old. (hiv. old. 2, 13).
- [4] David Iclanzan, Fabio Daolio és Marco Tomassini. «Data-driven local optima network characterization of QAPLIB instances». *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*. 2014, 453–460. old. (hiv. old. 2).
- [5] David Iclanzan, Fabio Daolio és Marco Tomassini. «Learning inherent networks from stochastic search methods». *European Conference on Evolutionary Computation in Combinatorial Optimization*. Springer. 2014, 157–169. old. (hiv. old. 2).
- [6] Réka Albert és Albert-László Barabási. «Statistical mechanics of complex networks». *Reviews of modern physics* 74.1 (2002), 47. old. (hiv. old. 2).
- [7] Jeffrey C Lagarias. «The $3x+1$ problem and its generalizations». *The American Mathematical Monthly* 92.1 (1985), 3–23. old. (hiv. old. 3).
- [8] Jeffrey C Lagarias. *The ultimate challenge: The $3x+1$ problem*. American Mathematical Soc., 2010 (hiv. old. 3).
- [9] Riho Terras. «A stopping time problem on the positive integers». *Acta Arithmetica* 3.30 (1976), 241–252. old. (hiv. old. 3).
- [10] Terence Tao. «Almost all orbits of the Collatz map attain almost bounded values». *arXiv preprint arXiv:1909.03562* (2019) (hiv. old. 3).

- [11] David Barina. «Convergence verification of the Collatz problem». *The Journal of Supercomputing* 77.3 (2021), 2681–2688. old. (hiv. old. 3, 4).
- [12] Emre Yolcu, Scott Aaronson és Marijn JH Heule. «An automated approach to the collatz conjecture». *Automated Deduction–CADE 28* (2021), 468. old. (hiv. old. 3).
- [13] Liesbeth De Mol. «Tag systems and Collatz-like functions». *Theoretical Computer Science* 390.1 (2008), 92–101. old. (hiv. old. 3).
- [14] Paul Andoloro. «On total stopping times under $3x+1$ iteration». *Fibonacci Quarterly* 38.1 (2000), 73–78. old. (hiv. old. 5).
- [15] Jose Capco. *Odd Collatz sequence and binary representations*. Techn. jel. hal-02062503. HAL archives-ouvertes.fr, 2019 (hiv. old. 5).
- [16] Thoma Cloney, Eric Goles és Gerard Y Vichniac. «The $3x+1$ problem: a quasi cellular automaton». *Complex Systems* 1.2 (1987) (hiv. old. 5).
- [17] M Bruschi. «Two Cellular Automata for the $3x+1$ Map». *arXiv e-prints* (2005), nlin-0502061. (hiv. old. 5).
- [18] Jeffrey C Lagarias és A Weiss. «The $3x+1$ problem: two stochastic models». *The Annals of Applied Probability* (1992), 229–261. old. (hiv. old. 5).
- [19] S Wright. «The roles of mutation, inbreeding, cross-breeding and selection in evolution, p 356–366». *Proceedings of the Sixth Annual Congress of Genetics*. 1. köt. 1932 (hiv. old. 5).
- [20] Katherine Mary Malan. «A survey of advances in landscape analysis for optimisation». *Algorithms* 14.2 (2021), 40. old. (hiv. old. 6).
- [21] Katherine M Malan és Andries P Engelbrecht. «A survey of techniques for characterising fitness landscapes and some possible ways forward». *Information Sciences* 241 (2013), 148–163. old. (hiv. old. 6).
- [22] Sébastien Verel, Gabriela Ochoa és Marco Tomassini. «Local optima networks of NK landscapes with neutrality». *IEEE Transactions on Evolutionary Computation* 15.6 (2010), 783–797. old. (hiv. old. 6).
- [23] Olaf Mersmann, Bernd Bischl, Heike Trautmann, Mike Preuss, Claus Weihs és Günter Rudolph. «Exploratory landscape analysis». *Proceedings of the 13th annual conference on Genetic and evolutionary computation*. 2011, 829–836. old. (hiv. old. 6).
- [24] Rachael Morgan és Marcus Gallagher. «Length scale for characterising continuous optimization problems». *International Conference on Parallel Problem Solving from Nature*. Springer. 2012, 407–416. old. (hiv. old. 6).

- [25] Hendrik Richter. «Codynamic fitness landscapes of coevolutionary minimal substrates». *2014 IEEE Congress on Evolutionary Computation (CEC)*. IEEE. 2014, 2692–2699. old. (hiv. old. 6).
- [26] Fabio Caraffini, Ferrante Neri és Lorenzo Picinali. «An analysis on separability for memetic computing automatic design». *Information Sciences* 265 (2014), 1–22. old. (hiv. old. 6).
- [27] Katherine M Malan, Johannes F Oberholzer és Andries P Engelbrecht. «Characterising constrained continuous optimisation problems». *2015 IEEE Congress on Evolutionary Computation (CEC)*. IEEE. 2015, 1351–1358. old. (hiv. old. 7).
- [28] Shinichi Shirakawa és Tomoharu Nagao. «Bag of local landscape features for fitness landscape analysis». *Soft Computing* 20.10 (2016), 3787–3802. old. (hiv. old. 7).
- [29] Yuan Sun, Michael Kirley és Saman K Halgamuge. «Quantifying variable interactions in continuous optimization problems». *IEEE Transactions on Evolutionary Computation* 21.2 (2016), 249–264. old. (hiv. old. 7).
- [30] Mang Wang, Bin Li, Guofu Zhang és Xin Yao. «Population evolvability: Dynamic fitness landscape analysis for population-based metaheuristic algorithms». *IEEE Transactions on Evolutionary Computation* 22.4 (2017), 550–563. old. (hiv. old. 7).
- [31] Arnaud Liefoghe, Fabio Daolio, Sébastien Verel, Bilel Derbel, Hernan Aguirre és Kiyoshi Tanaka. «Landscape-aware performance prediction for evolutionary multiobjective optimization». *IEEE Transactions on Evolutionary Computation* 24.6 (2019), 1063–1077. old. (hiv. old. 7).
- [32] Anna S Bosman, Andries P Engelbrecht és Mardé Helbig. «Progressive gradient walk for neural network fitness landscape analysis». *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 2018, 1473–1480. old. (hiv. old. 7).
- [33] Vesselin K Vassilev, Terence C Fogarty és Julian F Miller. «Information characteristics and the structure of landscapes». *Evolutionary computation* 8.1 (2000), 31–60. old. (hiv. old. 10).
- [34] Wei Li, Xiang Meng és Ying Huang. «Fitness distance correlation and mixed search strategy for differential evolution». *Neurocomputing* 458 (2021), 514–525. old. (hiv. old. 11).
- [35] Yu Wang és Bin Li. «Understand behavior and performance of real coded optimization algorithms via NK-linkage model». *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*. IEEE. 2008, 801–808. old. (hiv. old. 11).

- [36] Marco Tomassini, Leonardo Vanneschi, Philippe Collard és Manuel Clergue. «A study of fitness distance correlation as a difficulty measure in genetic programming». *Evolutionary computation* 13.2 (2005), 213–239. old. (hiv. old. 11).
- [37] Leila Kallel, Bart Naudts és Marc Schoenauer. «On functions with a given fitness-distance relation». *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*. 3. köt. IEEE. 1999, 1910–1916. old. (hiv. old. 11).
- [38] Monte Lunacek és Darrell Whitley. «The dispersion metric and the CMA evolution strategy». *Proceedings of the 8th annual conference on Genetic and evolutionary computation*. 2006, 477–484. old. (hiv. old. 11, 12).
- [39] Rachael Morgan és Marcus Gallagher. «Sampling techniques and distance metrics in high dimensional continuous landscape analysis: Limitations and improvements». *IEEE Transactions on Evolutionary Computation* 18.3 (2013), 456–461. old. (hiv. old. 12).
- [40] Edward Weinberger. «Correlated and uncorrelated fitness landscapes and how to tell the difference». *Biological cybernetics* 63.5 (1990), 325–336. old. (hiv. old. 12).
- [41] Peter F Stadler. «Landscapes and their correlation functions». *Journal of Mathematical chemistry* 20.1 (1996), 1–45. old. (hiv. old. 12).
- [42] Boryana Dimova, J Wesley Barnes és Elmira Popova. «Arbitrary elementary landscapes & AR (1) processes». *Applied Mathematics Letters* 18.3 (2005), 287–292. old. (hiv. old. 12).
- [43] Peter Merz és Bernd Freisleben. «Fitness landscape analysis and memetic algorithms for the quadratic assignment problem». *IEEE transactions on evolutionary computation* 4.4 (2000), 337–352. old. (hiv. old. 12).
- [44] Fan Chung és Wenbo Zhao. «PageRank and random walks on graphs». *Fete of combinatorics and computer science*. Springer, 2010, 43–62. old. (hiv. old. 13).
- [45] Martin Pelikan, Mark W Hauschild és Fernando G Lobo. «Estimation of distribution algorithms». *Springer handbook of computational intelligence*. Springer, 2015, 899–928. old. (hiv. old. 14).
- [46] Carsten Witt. «Theory of estimation-of-distribution algorithms». *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*. 2020, 1254–1282. old. (hiv. old. 14).
- [47] Heinz Mühlenbein és Gerhard Paass. «From recombination of genes to the estimation of distributions I. Binary parameters». *International conference on parallel problem solving from nature*. Springer. 1996, 178–187. old. (hiv. old. 14).

- [48] Benjamin Doerr és Martin S Krejca. «The univariate marginal distribution algorithm copes well with deception and epistasis». *Evolutionary Computation* 29.4 (2021), 543–563. old. (hiv. old. 14, 20).
- [49] David Iclănzan. «Global Optimization of Multimodal Deceptive Functions». *Evolutionary Computation in Combinatorial Optimisation*. Szerk. Christian Blum és Gabriela Ochoa. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, 145–156. old. ISBN: 978-3-662-44320-0 (hiv. old. 20).