

Szegedi Tudományegyetem TTIK Fizikai Tanszék

TDK DOLGOZAT

A DBSCAN klaszteranalízis paraméterfüggésének karakterizálása

Készítette: Trauer János Fizika MSc szakos hallgató

Témavezető: Varga Dániel Tudományos segédmunkatárs

SZTE Optikai és Kvantumelektronikai Tanszék

Szeged 2022

Tartalomjegyzék

Tartalomjegyzék	2
Irodalmi áttekintés	3
Mikroszkópia	3
Fluoreszcens mikroszkópia.....	4
Szuperrezolúciós lokalizációs technika.....	4
TestSTORM	6
rainSTORM	8
Klaszteranalízis.....	9
Célkitűzés	11
Saját eredmények	11
Mérési minta szimulálása	11
A szimulált felvételsorozatok rekonstruálása	12
A lokalizációs adatok klaszteranalízise	13
Eredmények értelmezése	15
Különböző sűrűségű minták vizsgálata	20
Összefoglalás	23
Felhasznált irodalom.....	24
Köszönetnyilvánítás	25

Irodalmi áttekintés

Dolgozatomban bemutatom a megértéshez szükséges alapokat, érintve a mikroszkópiát és a fluoreszcenciát felhasználó fluoreszcens mikroszkópiát. Áttekintem továbbá, az általam felhasznált szuperrezolúciós, lokalizációs képalkotási technikát, a minták szimulálására alkalmas szoftvert és a klaszterek vizsgálatára használt klaszteranalízis módszerek [1], [2] egyikét a DBSCANt.

Mikroszkópia

Az emberiség gyakran akart olyan struktúrákat vizsgálni, amelyeknek mérete kisebb, mint a szem feloldási képessége, ezért további eszközöket kellett használnia, mint a nagyító vagy később a mikroszkóp. Egy mikroszkóp legfőbb tulajdonságai a nagyítás, a feloldás, a képalkotás módja és a fénygyűjtőképessége. Ezek a tulajdonságok felhasználási terület szerint különbözőek, továbbá egymástól nem függetlenek. Azaz ha az egyik tulajdonságot megváltoztatjuk, akkor gyakran más tulajdonsága is változik. [3],[4]

Egy optikai mikroszkóp 4 fő részből épül fel. Első fontos elem a fényforrás. Ennek megválasztásakor érdemes figyelembe venni, hogy az adott fényforrásnak mekkora a teljesítménye, koherens-e, mi a spektruma, stb.

A második az objektív. Egy objektívnek sok fontos tulajdonsága van és ezek alapján lehet kategorizálni azokat. Ilyen specifikációk a különböző korrekciók. A kromatikus aberráció korrekció alapján három fő kategóriába sorolhatjuk az objektíveket. A további fontos tulajdonságok, az objektív nagyítása, a numerikus apertúrája (NA), immerziós közeg típusa (olaj, víz, glicerín), a fedő(üveg)lemez vastagsága (általában 0.17 mm), a munkatávolság (0.21 mm), és hogy végtelenre korrigált-e. A feloldás szempontjából a NA az egyik legfontosabb tulajdonsága a mikroszkópnak. Látható fény esetén a laterális feloldás ~200 nm.

Harmadik alkotóelem a tubus lencse, melynek főbb paraméterei közé tartozik a fókusz távolság és a különböző korrekciók.

A mikroszkóp negyedik eleme a detektor. Legfőbb tulajdonságaik közé tartoznak a spektrális érzékenység, kiolvasás sebessége, a kvantum hatásfok, és az erősítés módja.

Fluoreszcens mikroszkópia

Az ebben az alfejezetben bemutatott technika olyan minták vizsgálatára alkalmas amelyek vagy természetes állapotában vagy mesterségesen, fluoreszcenciára képesek.

A fluoreszcens mikroszkópia [3], [5] esetén, az előző alfejezetben leírt elemeken túl, további alkotórészeket is tartalmaz a mikroszkóp. Ezek a dikroikus tükrök, valamint a gerjesztési- és emissziós szűrő [4]. A dikroikus tükrő egy olyan tükrő amely bizonyos hullámhosszt átenged, azonban a többit egy másik irányba reflektálja. A tükrök felhasználhatók, a különböző hullámhosszúságú gerjesztőlézerek szétválasztására. Ezekre az elemekre azért van szükség, mert a gerjesztő fény intenzitása jóval nagyobb az általunk vizsgált fluoróforok intenzitásánál. Ezeket az elemeket, és a Stokes-féle eltolódási törvényt kihasználva, szeparálhatjuk a gerjesztő fényt a mintából érkező emissziós jelet, ami nagy kontrasztot tesz lehetővé.

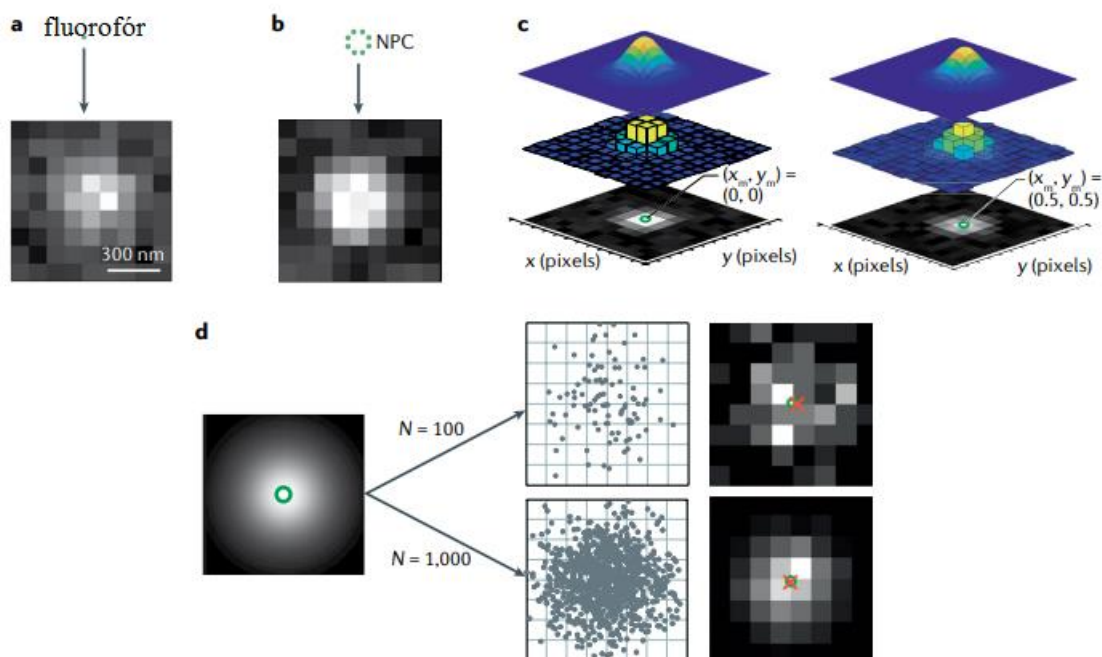
Szuperrezolúciós lokalizációs technika

Az olyan technikákat, amelyek lehetővé teszik a diffrakció limit alatti feloldást szuperrezolúciós mikroszkópos technikáknak hívjuk. Ezen technikák közül a lokalizációs mikroszkópokkal lehet elérni a legnagyobb térbeli feloldást. Ide tartozik a fotoaktiválható lokalizációs mikroszkópia (PALM), PAINT, és a (direkt) sztohasztikus optikai rekonstrukciós mikroszkópia (STORM vagy dSTORM). [4],[5], Ebben az alfejezetben az utóbbival foglalkozom. A lokalizációs technikák azon alapulnak, hogy az egyedi molekulák pozíciója pontosan megadható, ha képük, vagyis a hozzájuk tartozó PSF-ek nem fednek át. A PSF (Point Spread Function) az optikai rendszerre jellemző pontátviteli függvény. A mikroszkópiában: a pontszerűnek tekinthető fényforrás, esetünkben fluoreszcens festék, detektorra leképzett képe. A PSF-ek átfedését időbeli szeparálással érjük el. Erre leggyakrabban bizonyos szerves festékek azon fotokémiai tulajdonságát használjuk fel, hogy kapcsolhatóak egy fluoreszcensen aktív és egy sötét állapot között. A STORM esetén a fluoreszcens molekuláknak három állapota van. Az egyik az aktív (ON) állapot, mikor fényt bocsájtanak ki. A másik a kikapcsolt állapot (OFF), mikor sötét van, a molekula nem bocsájt ki fényt, de még véletlenszerűen visszatérhet aktív állapotba. A harmadik pedig a kiégetett/kifakult (bleached) állapot amikor már nem fog kibocsájtani fotont a molekula, nem tér vissza aktív állapotba. Ennél a technikánál fontos, az egyidőben aktív állapotú molekulák számát egy optimális értéken tartani, ezzel elérve azt, hogy a felvillanások (és ezzel a detektál képük) elszeparáltak mint időben, mint térben. Azonban

ahhoz, hogy ezt teljesíthessük, szükségünk van egy puffer oldatra. A detektálás pontosságát nagyban befolyásolja a molekula által emittált fotonok száma. Minél nagyobb a kibocsájtott fotonok száma (és ezzel együtt a detektálté is) annál pontosabban lehet meghatározni a pozícióját a molekulának. Az egyik legszélesebb körben használt festék, az Alexa Fluor 647, és a kifakulás előtt ~10 alkalommal lehet aktiválni. Ez azt eredményezi, hogy egy fluorofór képe nem egy pont lesz, hanem egy klaszter. Ennek a klaszternek a mérete sok paramétertől függ. Csak, hogy párat említsünk: a festéket és a fehérjét összekapcsoló kötés hossza, az epitóp körüli forgás sebessége, vagy a lokalizációs precizitás. Ezek alapján is megállapítható, hogy egy minta optimális állapotban tartása, nagy kihívás.

A lokalizációs technikák (SMLM – Single Molecule Localization Microscopy) feladata az, hogy a detektor (pl. CCD kamera) által létrehozott pixelizált képekből kinyerjük a felvillanó molekulák pontos pozícióját, és ezek segítségével létrehozunk egy nagyobb feloldású képet a vizsgált struktúráról. Ebben az alfejezetben ennek a folyamatnak a különböző fázisait mutatom be.

Egy mérés során nyert képen, a fluoreszcens molekula egy több pixelt elfoglaló, zajos foltként jelenik meg (PSF, 1.ábra **a**), amelynek mérete 200-300 nm-es tartományba esik. Egy pont elmosódását okozhatja az, hogy egymáshoz nagyon közel lévő fluoreszcens molekula egyszerre emittál, így a képük átfednek egymással, így a pixelizált képen egy folt látható, azonban két vagy több molekula van jelen az adott területen (1.ábra **b**). Ezek szétválasztására és a foltokhoz tartozó molekulák pontos helyzetének megállapítására felhasználjuk azt, hogy tudjuk, egy ilyen felvillanás helyének legjobb közelítése az Airy-féle elhajlási kép közepe. Ilyen illesztések felhasználásával megkaphatjuk a lokalizáció x , y koordinátáját a pixelen belül. Azonban, számítási igények alacsonyan tartása céljából, a közelítésre használhatunk Gauss függvényt az Airy függvény helyett (1.ábra **c**). Az illesztés több módon is történhet, használhatjuk a maximum likelihood becslést vagy a legkisebb négyzetek módszerét. Az illesztéssel kapott maximumhelyet nagyjából 20 nm precizitással lehet meghatározni fluoreszcens molekulák esetén. Mint egy korábbi alfejezetben kitértem rá, a lokalizáció precizitását nagyban elősegíti a fluoreszcens molekula által kibocsájtott fotonok számának (N) magasabb értéke is. Minél magasabb ez a szám annál pontosabb a mérés is (1.ábra **d**). [10]



1. ábra A SMLM technikák alapvető lépései. a) A fluorofór képe, b) Egy gyűrűs struktúra képe, c) A Gauss függvény pontra illesztése és koordináta meghatározása, d) Az emittált fotonok számának (N) jelentősége. Az ábra nem saját, szerkesztett.

[10]

Az így detektált felvillanások, illesztésből megkapott koordinátáit a program elmenti egy táblázatba, majd ezeket az adatokat felhasználva létrehoz egy nagyfeloldású képet. Ehhez a feldolgozáshoz mi, a kutatócsoport által megírt lokalizációs szoftvert használtuk, a rainSTORM-ot.

TestSTORM

A TestSTORM [6],[7], az SZTE OKT AdOptIm kutatócsoport tagjai által, MatLab-ben megírt szimulációs szoftver, amely arra lett kifejlesztve, hogy szimulálja az egész mérési folyamatot. A program könnyű használatát egy felhasználói felület segíti (2.ábra). A szoftverben beállíthatunk különböző mérés típust, például polarizáció érzékeny, vagy két színű mérést, attól függően melyet szeretnénk szimulálni. A következő beállítási lehetőség a PSF típusa. Ezután jön a legfontosabb beállítási lehetőség a minta típusának megadása. A TestSTORM-ban 8 darab mintatípusból lehet választani, amelyek között vannak valódi biológiai struktúrák alakját szimulálók is. Továbbá, választhatunk különböző festékek közül és azok paramétereit is megadhatjuk. Beállíthatjuk továbbá, hogy a minta elmozdul-e a mérés során (drift). Legvégül

beállíthatjuk a szimulálni kívánt optikai rendszer és a kamera paramétereit is. (Az általunk használt minta megadásának módját a következő fejezetben taglalom.)

TestSTORM: test sample generator program for localization

Type of Measurement

☐ Polarization sensitive

☐ Dual-Colour

Crosstalk:

☐ Astigmatic 3D

Cylindrical l. f. (m):

Magnification:

Tube lens f. (cm):

☒ Fix seed for rand. numbers

Patterns

disks
disks

Modify Delete

Add New Pattern

Import from parameters file

Dye Parameters

☒ Dye 1 ☐ Dye 2

Select:

Emission WL:

Char. ON time (s):

Char. OFF time:

Bleaching constant (s):

Emitted:

Mean bonding angle (°):

SD of bonding angle (°):

Mean N of labels/epitope:

Var. N of labels/epitope:

Length of linkers (nm):

Non-spec. l. dens. (1/um3):

Acquisition Parameters

Frame size (px):

Number of frames:

Frame rate (1/s):

Exp. time:

Pixel size:

Av. BG level:

Struct. BG strength:

RI of immersion:

RI of sample:

Numerical aperture:

Electrons/count:

Pre-amplification:

Actual EM:

Quantum efficiency:

PSF Type

☒ Gaussian

☐ Scalar

☐ Vectorial

Gaussian PSF Parameters

Opt. coll.

Drift Parameters

Drift type:

Std of acc (nm/s):

Drift velocity (nm/s):

Damping coeffs (1/s, 1/nm):

Save As

Search

Export parameters

Import parameters

Generate

Recalculate

Plot labels

Progress_report:

AdOptim Group, University of Szeged, Hungary, 2014-2018©

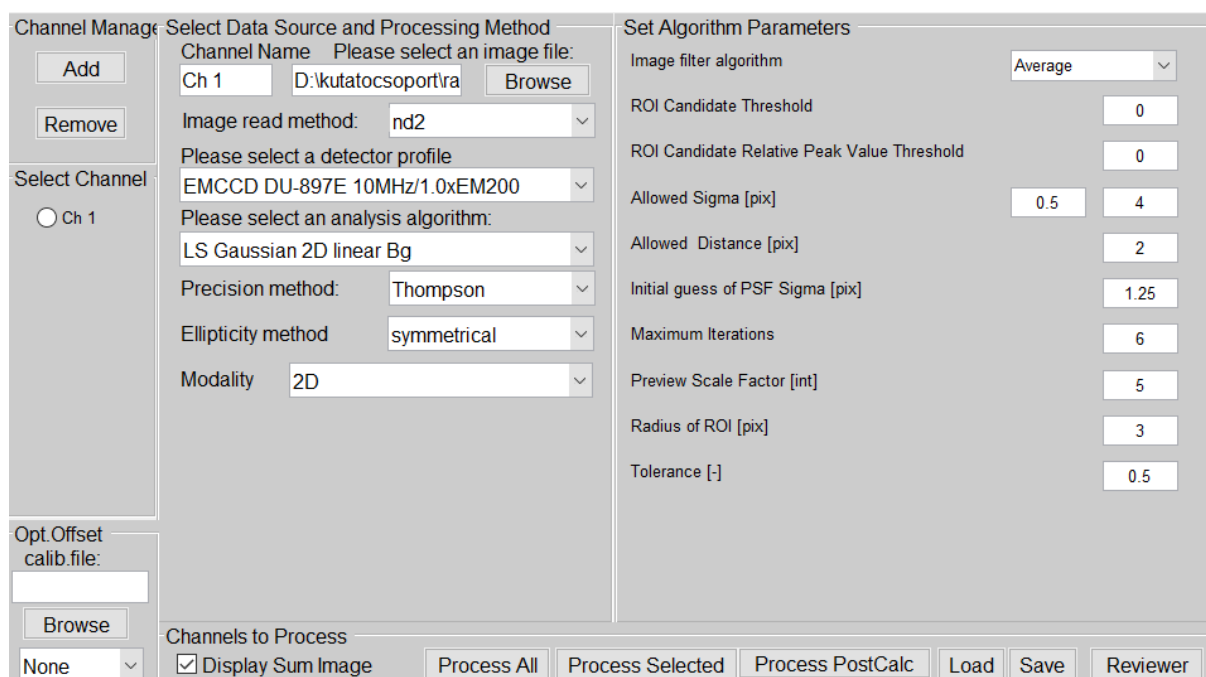
2. ábra A TestSTORM felhasználói felülete

A TestSTORM algoritmus három főbb lépést hajt végre:

- A szoftver először a kiválasztott mintához kapcsolódó fluoreszcens molekulák 3D-s koordinátáit generálja le. Ezek véletlenszerűen kapcsolódnak a megadott struktúrához.
- Az algoritmus második szakaszában történik a minta viselkedésének szimulálása, a festékparaméterek alapján.
- A harmadik fázisban jön létre paramétereknek megfelelő pixelizált kép, képsorozat.

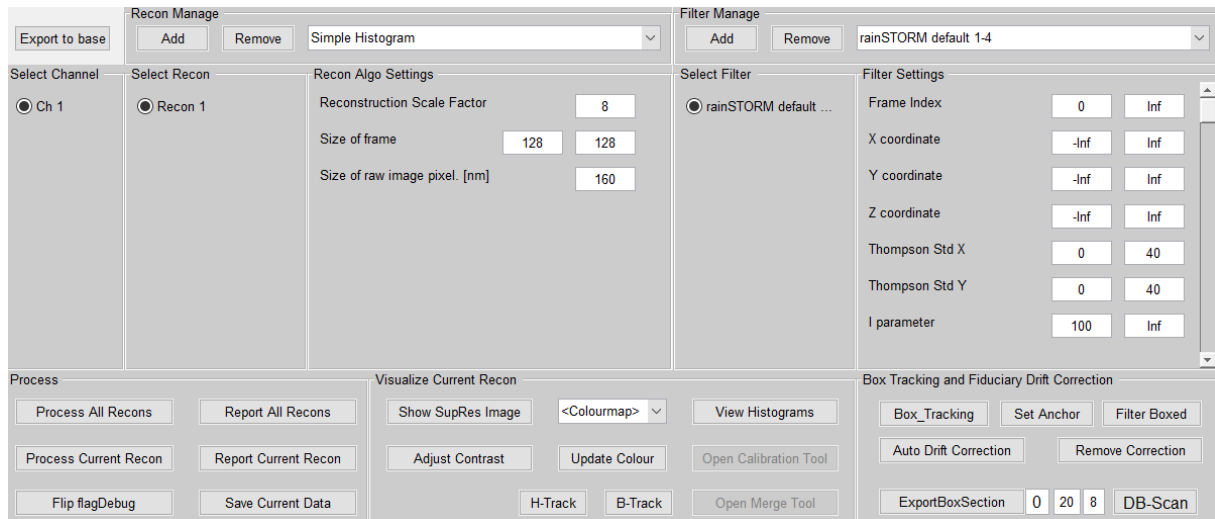
rainSTORM

A rainSTORM [8], [9] a TestSTORM-hoz hasonlóan egy MatLab-ben megírt szoftver, amely a lokalizációs mikroszkópai mérések (PALM vagy (d)STORM) vagy szimulációk (TestSTORM) által készített felvételsorozatok kiértékelésére lett kifejlesztve. A program használatakor két felhasználói felület segíti a felhasználót (3. és 4. ábra).



2. ábra A rainSTORM első felhasználói felülete

Az elsőn adhatjuk meg a feldolgozás módját. A szoftver képes 6 különböző fájltypust (tiff, tiff-folder, raw, nd2, dat és csv) beolvasni és automatikusan kiválasztani az ennek megfelelő képbeolvasási metódust is. A következő beállítási lehetőség a detektor típusának megadása. Itt 10 darab lehetőség közül lehet választani. Ezután megadhatjuk a 12 darab, illesztési algoritmus közül azt, amelyet használni szeretnénk. Továbbá, lehetőségünk van megváltoztatni az algoritmus paramétereit, ha az előre definiált értékek nem lennének megfelelőek. Abban az esetben, ha több képsorozatot szeretnénk kiértékelni a rainSTORM-mal, akkor arra is van lehetőség, plusz csatorna (channel) hozzáadásával. Minden paraméter megadása után a "Process All" gombra kattintva lefuttathatjuk a kiértékelést, majd annak befejeztével a "Reviewer" megnyomásával előhívhatjuk a második felületet.



3. ábra A Reviewer felhasználói felülete

Ezen a felületen tudjuk elmenteni a feldolgozott adatokat, például a koordinátákat tartalmazó táblázatot. Továbbá előhívhatjuk és elmenthetjük a rekonstruált nagyfeloldású képet és a hozzá tartozó fontosabb hisztogramokat is. Lehetőség van drift korrekció végrehajtására és különböző szűrők beállítására (például csak a legpontosabb lokalizációkat ábrázoljuk), hogy jobb képet kapjunk. Valamint lefuttathatjuk a következőkben taglalt DBSCAN klaszteranalízist is.

Klaszteranalízis

Azokat az algoritmusokat melyek során valamilyen módszer alapján csoportosítjuk a lokalizációkat klaszteranalízisnek nevezzük. Erre több módszer is létezik, például a DBSCAN [11], [12]. A DBSCAN (Density-based spatial clustering of applications with noise) egy sűrűségeen alapuló, klaszteranalízisre használt algoritmus,

A DBSCAN azon az elven alapszik, hogy egy klaszteren belül a lokalizációk sűrűsége magasabb, mint a klaszteren kívül lévő zajok sűrűsége. Ahhoz, hogy ezt a megkülönböztetést eltudja végezni, és megtalálja a sűrűbb területeket (klasztereket), két paramétert kell megadnunk az algoritmusnak, az ε -t és az N -t. Az ε a vizsgálni kívánt pont köré megrajzolt kör sugara, N pedig az a minimális pontmennyiség aminek jelen kell lennie az ε által meghatározott kör területében. Azokat a pontokat amelyekre ez a feltétel teljesül mag pontoknak (core points) nevezzük. A klaszter definiáláshoz további fogalmakat kell bevezetnünk.

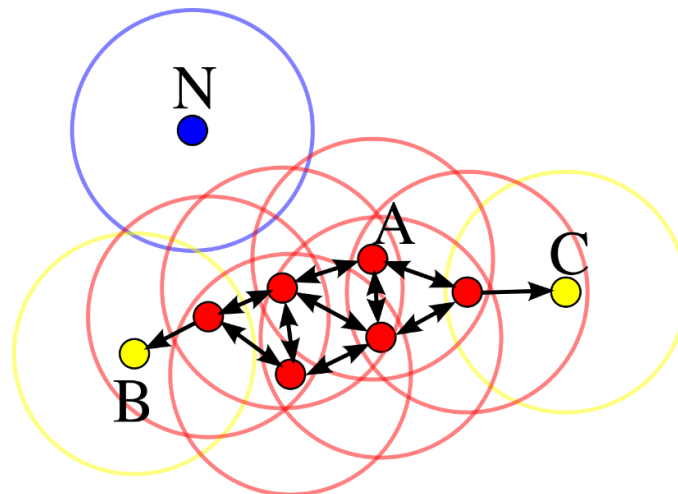
- Egy pont ε -szomszédjának hívjuk azokat a pontokat amelyek ε távolságra vagy annál közelebb helyezkednek el. Ezen pontok összességét N_ε -el jelöljük.

- Egy b pont közvetlenül elérhető (directly density-reachable) egy q pontból, ha a q pont Eps-szomszédjai közé tartozik (b eleme $N_\epsilon(q)$ -nak) és $N_\epsilon(q)$ értéke nagyobb vagy egyenlő N -nel.
- Egy b pont elérhető (density-reachable) a q pontból, ha összeköthetők olyan p pontok láncolatával amelyekre igaz, hogy $N_\epsilon(p)$ nagyobb vagy egyenlő N . Ennek a feltételnek nem kell teljesülnie a b pontra. Ebből kifolyólag az elérhetőség nem szimmetrikus.
- Egy b pont kapcsolódik (density-connected) a q ponthoz, ha van egy olyan o pont amelyből mind a b mind a q pont elérhető. Ezért a kapcsolódás tulajdonság szimmetrikus.

Ezen fogalmak segítségével definiálni tudjuk, hogy mi számít klaszternek és mi zajnak.

- Klaszter azon pontok halmaza amelyek egymásból elérhetők, valamint kapcsolódnak is egymáshoz.
- Zaj azon pontok halmaza amelyek egy klaszterhez sem tartoznak.

Azokat a pontokat amelyek nem zaj pontok és nem is mag pontok azok a határ pontok. Ezek még a klaszter részei, azonban nincs elegendő pont körülöttük, ezért az ezeken túl lévő pontokra már nem fut le az algoritmus.



4. ábra A DBSCAN működése. A pirossal jelölt pontok (pl. A) a mag pontok, mivel elegendő mennyiségű Eps-szomszédjuk van (4 db). A B és C pontok a határ pontok. N egy zaj pont. [12]

Célkitűzés

Dolgozatomban a DBSCAN algoritmussal foglalkoztam, ami egy széles körben használt módszer. Előnye, hogy a klasztereket egyedileg lehet vele jellemezni és tudja kezelni a zajt, valamint a klaszter szélén lévő pontokat. Hátránya viszont, hogy a felhasználótól kéri be a szükséges paramétereket, és ezek pontos beállítása lényegesen befolyásolja az eredmények helyességét. Dolgozatomban biológiailag releváns, több skálán klasztereződést mutató minták elemzésével foglalkoztam. Megvizsgáltam, hogy a DBSCAN klaszteranalízis eredménye hogyan függ a kezdeti paraméterek beállításától.

Saját eredmények

Mérési minta szimulálása

Első feladatomban a minták előállítása volt, ami, az előző fejezetben már említett, biológiai minták szimulálására alkalmas szoftver, a TestSTORM használatával történt. Mivel a TestSTORM több típusú mintának szimulálására alkalmas, azonban én csak egy típust használtam, így ebben a fejezetben, csak az általam használt szimulációk paramétereit/beállításait mutatom be.

A PSF beállításánál a fényforrások intenzitáseloszlását a Gauss nyalábokhoz hasonlatosnak tételeztem fel, így ezt a típust választottam. Minden beállított paraméterrel több szimulációt is lefuttattam.

Add Disks Pattern	
Length (nm):	680
Radius (nm):	250
Grating (nm):	[3400;0]
Dens (1/um):	[40;40]
Fi (°):	-15
Theta (°):	0
X (nm):	-250
Y (nm):	0
Z (nm):	0

☐ Star Pattern

☐ Array Pattern

☐ Vesicles Pattern

☐ Rods Pattern

☐ Lines Pattern

☐ Octagons Pattern

☐ Axons Pattern

☒ Disks Pattern

☐ Read Lines from File

☐ Read Mesh from File

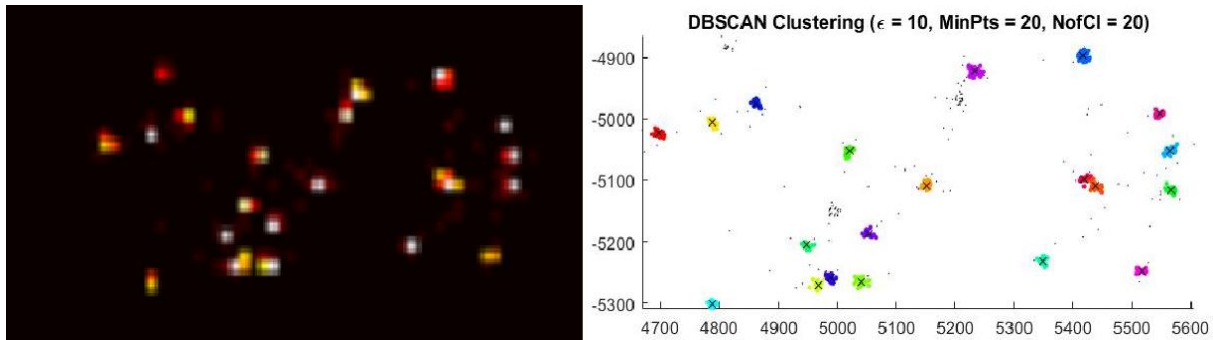
ok

6. ábra A minta típusának és paramétereinek megadása. A lehetséges mintatípusok (jobb oldal). A korong struktúra paramétereinek beállítása (bal oldal)

A következő fontos lépés, a minta megadása. Erre, a számomra legalkalmasabb mintatípus, a korong (disks), amely esetén egy korongon belül véletlenszerűen helyezkednek el az epitópok. Azért ezt a mintát használtam, mert ennek van biológiai relevanciája, pl. a DNS kettős szálú törésekben elhelyezkedő javító fehérjék hasonló struktúrában helyezkednek el. Vizsgálataimhoz kettő, közvetlen egymás mellett elhelyezkedő korongot tartalmazó mintákat használtam, amelyeknél a korongok sugara 250 nm, középpontjaik távolsága pedig 500 nm. A vizsgálatok során több különböző sűrűségű mintát szimuláltam, amelyeknek az eddig leírt tulajdonságuk azonos. Az epitópok sűrűsége a különböző szimulációknál 30, 40, 50, 60, 80 $1/\mu\text{m}$, az egy epitóphoz csatlakozó festékek száma 5 festék /epitóp. A festék paramétereinél, az alapértelmezett Alexa Fluor 647-est állítottam be. A festék, az optikai rendszer és a detektor szimulált paramétereit a valós méréseknek megfelelően állítottam be. A drift szimuláció paramétereit az alapértelmezett értékeken hagytam, mivel esetemben nincs szükség drift beállítására. Mindezen paraméterek beállítása után a generálást (Generate) elindítva, a szoftver létrehoz egy 3000 képkockából (frame) álló képsorozatot, amit tif formátumban ment el. Minden, 40, 60, 80 $1/\mu\text{m}$, sűrűségű minták esetén, 10-10 db szimulációt készítettem.

A szimulált felvételsorozatok rekonstruálása

Az így kapott fájlból, a korábban bemutatott rainSTORM nevű szoftver felhasználásával nyerhetjük ki a lokalizációk koordinátáit. Ehhez ki kell választanunk, a megfelelő analízis algoritmust, ami esetünkben a "Least-Squares Gaussian 2D linear Bg". Ezt kiválasztva a szoftver a felvillanások képéből egy lineáris háttérrel meghatároz és levon, majd a felvillanásra egy szimmetrikus Gauss függvényt illeszt, és az így kapott illesztés középpontjába helyezi a lokalizációt. Ezt az illesztést minden képkockán (frame) elvégzi a program. A háromezer képkockán lefuttatott vizsgálatok során kapott lokalizációs adatok (a fényforrások koordinátái) alapján a rainSTORM rekonstruálja a nagyfeloldású képet. Ezt a képet a reviewer felhasználói felületen belül jeleníthetjük meg, valamint a lokalizációk koordinátákat is elmenthetjük csv formátumban, a további, más programmal való vizsgálatok elvégzéséhez (mint például esetünkben). Továbbá a szoftveren belül alkalmazható a DBSCAN is. A programba beépített DBSCAN akkor a leghasznosabb, mikor bizonyos N és ε értékek esetén, ábrázolni szeretnénk a talált klasztereket (erre mutat példát a 7. ábra jobb oldalán látható kép).



7. ábra A nagyfeloldású kép (bal oldal), és a DBSCAN által visszaadott eredmény képe (jobb oldal) A színek csak a klaszterek megkülönböztetésére szolgálnak. A fekete pontok zajnak lettek jelölve.

Azonban a mi célunk a különböző N és ε értékek esetén talált klaszterszám vizsgálata, ezért egy olyan programot kellett írnom, amely a rainSTORM-ban lévő DBSCAN függvényt használja, oly módon, hogy egy megadott ε és N tartományon végzi el a vizsgálatot, valamilyen tetszőlegesen megadható lépésközzel.

A lokalizációs adatok klaszteranalízise

Az ilyen típusú vizsgálathoz használt programot MatLab-ben írtam meg, és a következő főbb lépéseket hajtja végre:

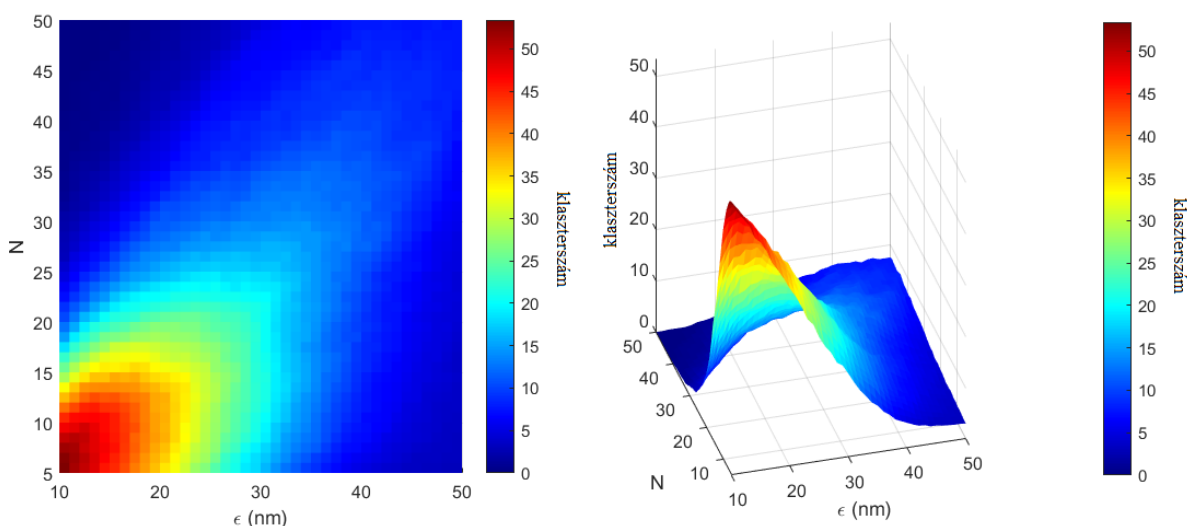
1. Beolvassa a rainSTORM által létrehozott és elmentett csv fájlt.
2. A fájlból kiolvassa a lokalizációk x , y és z koordinátáit nm-ben. Ezek a csv fájlban egy-egy oszlopot jelentenek, ezért a megfelelő oszlopok beolvasása szükséges.
3. Beolvassa az általunk megadott ε és N tartományt valamint a lépésközt. A tartományt úgy adtuk meg, hogy definiáltam egy kezdő és egy vég értéket mind a két paraméternél, továbbá megadtam egy lépésközt, figyelve arra, hogy a vég értékre is lefusson a program.
4. A megadott intervallumon, minden ε - N értékpár esetén kiszámolja a talált klaszterek számát a DBSCAN függvényt felhasználva, majd ezeket a számokat elmenti egy mátrixba. Ezekben a mátrixokban szereplő elemek, a talált klaszterek száma. Az oszlopok az N , a sorok pedig az ε értékeket jelentik. Ezek indexe nem egyezik meg az adott paraméter értékével, amire a későbbiekben fontos emlékezni.
5. A kapott eredményeket az adott tartományon ábrázolja és el is menti a létrehozott ábrát.

Vizsgálataink során 10-es ε és 5-ös N kezdőértéket, 1-es lépésközt, valamint 50-es végértéket használtam.

Mielőtt azonban rátérnénk az ábrázolásra és az azok által kapott eredmények vizsgálatára, érdemes átgondolni, milyen eredményeket, vagy legalábbis tendenciákat várunk, adott N és ϵ értékek esetén.

A DBSCAN előző fejezetben kifejtett működését átgondolva, elsőre azt várnánk, hogy adott N értéknél, az ϵ érték növelésével a talált klaszterek száma csökken, mivel minél nagyobb egy pont köré írt kör sugara, annál valószínűbb, hogy a minimális pontmennyiséget (N db pontot) tartalmazni fogja és a pontthalmazt egy klaszternek veszi. Azonban egy nagyobb N esetén, ϵ -t növelve eleinte növekedő klaszterszámra kell számítanunk, ugyanis kisebb sugarakra nem fog teljesülni a minimális pontmennyiség. Azaz várható minden kiértékelés során, hogy lesz egy olyan N érték, amelynél megjelenik ez a kezdeti növekedési tendencia a megtalált klaszterek számában, majd egy adott ϵ érték után ez a tendencia csökkenésbe megy át.

Az adott ϵ értékek esetén is hasonló tendenciákra számíthatunk. Kisebb sugárméretek esetében az N növekedésével a talált klaszterek száma csökken, mivel egy bizonyos elvárt minimális pontmennyiség fölött egyre kevesebb az esély arra, hogy teljesül a feltétel. Ehhez hasonló tendenciát várhatunk a nagyobb ϵ értékek esetén is, hiszen a kisebb N -ek biztos teljesülnek, azonban a feltétel szigorodásával csökkenés megy végbe a talált klaszterek számában, bár nagyobb N értéknél kezdődik ez a tendencia.

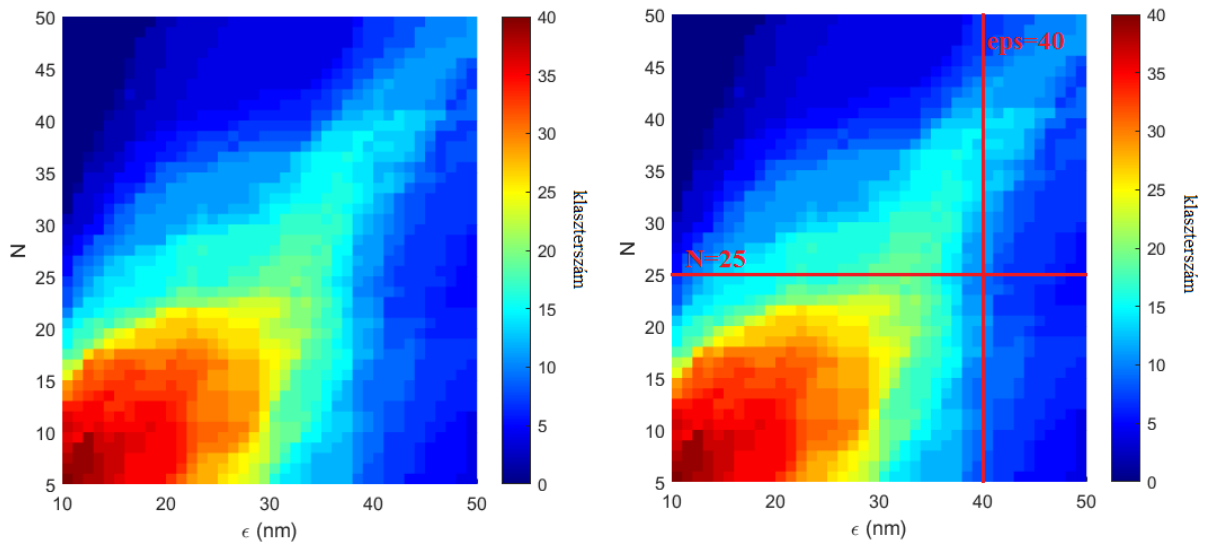


8. ábra A klaszteranalízis eredményének ábrázolása. Az eredeti háromdimenziós grafikon (jobb oldal), és a kétdimenziós, színskálázott térkép (bal oldal)

Az ábrázolás során arra voltam kíváncsi, hogy egy adott ε - N értékpárnál mekkora a talált klaszterszám. Ezért egy háromdimenziós grafikont hoztam létre. Az ábra x tengelyén az ε , az y tengelyen pedig az N értékek szerepelnek. A harmadik irányban pedig a talált klaszterek száma szerepel. Mivel egy háromdimenziós grafikon átlátása, értelmezése elég nehézkes, ezért az ábrázolás során a klaszterszámot nem csak a z tengely irányában ábrázoltam, hanem egy színskálával is jelöltem a darabszámot. Ezzel kaptam egy úgynevezett talált klaszterszám térképet, amely sokkal átláthatóbb módon közli az információt. A színskálát úgy érdemes megválasztani, hogy elég széles spektrumot öleljen fel, ezzel elérve azt, hogy a különböző klaszterszámokhoz tartozó tartományokat jobban el tudjuk különíteni. A 8.ábrán mutatok példát a kétdimenziós térképre és a háromdimenziós grafikonra.

Eredmények értelmezése

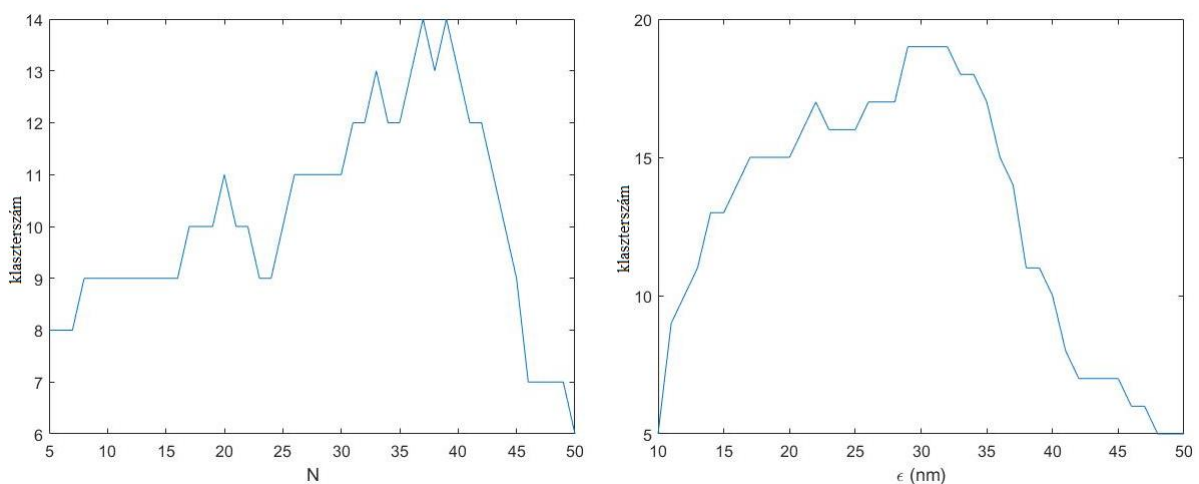
Az így kapott ábrákat megvizsgálva (például az 9.ábrán látható térképet) láthatjuk, hogy a korábbi alfejezetben taglalt, adott N értékre vonatkozó várt tendenciák teljesültek. Hasonlóan a kis ε értékekre érvényes gondolatmenet is igazolást nyert. Azonban észrevehetjük, hogy egy adott ε érték fölött a várt tendencia nem teljesül. (a térképeken egy maximális érték sáv jelenik meg, ami a térkép bal alsó sarkától a jobb felső sarkáig nyúlik) Azt láthatjuk, hogy egy széles ε tartományon eleinte a klaszterszám növekszik, majd egy maximális érték elérése után újra csökkenni kezd. Ez az eddig meggondolások alapján kissé ellentmondásosnak tűnhet a korábbi gondolatmenettel összehasonlítva, ahol abból indultunk ki, hogy a minimális pontmennyiség feltétel szigorodásával a klaszterszám csökken. Azt is megfigyelhetjük, hogy a vizsgált tartományban az utolsó 5 vizsgált (45-50) ε értéknél az N növelésével a talált klaszterek száma is nő.



9. ábra Egy 60 1/um sűrűségű mintához tartozó térkép (bal oldal). A választott értékek ($N=25$ és $\epsilon=40$) megjelölése (jobb oldal).

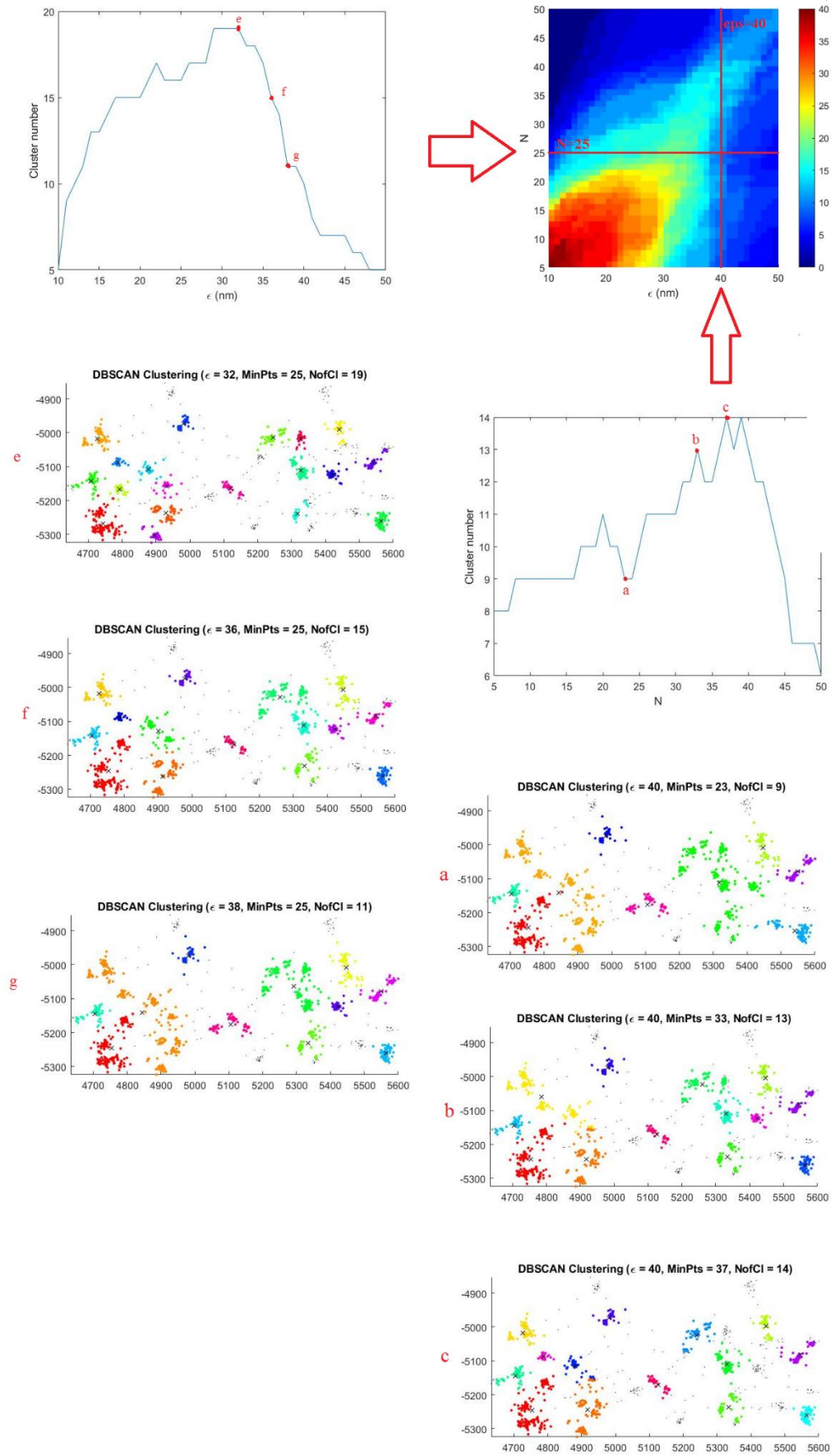
Ezeknek az elsőre ellentmondásosnak tűnő jelenségekre való magyarázatok megtalálásához és megértéséhez, további vizsgálatokra van szükség. Mivel az eredeti programom ilyen irányú vizsgálatot nem végez el, így egy ehhez szükséges programrészlettel egészítettem ki a meglévő szkriptet. Itt annyi történik, hogy kiválasztok egy adott ϵ vagy N értéket és az ahhoz tartozó N és ϵ értékek függvényében ábrázolom a klaszterszámot.

A vizsgálat bemutatásához az 9.ábra bal oldalán látható térképet használom. Ehhez olyan értékeket választottam, amelyeknél a kérdéses tendenciák figyelhetők meg. Ezekhez az $N=25$ és az $\epsilon=40$ értékeket választottam. Ezeket megjelöltem a 9.ábra jobb oldalán. Ezekre az értékekre lefuttattam az előbb leírt programot. Ennek eredményeit láthatjuk a 10.ábrán.

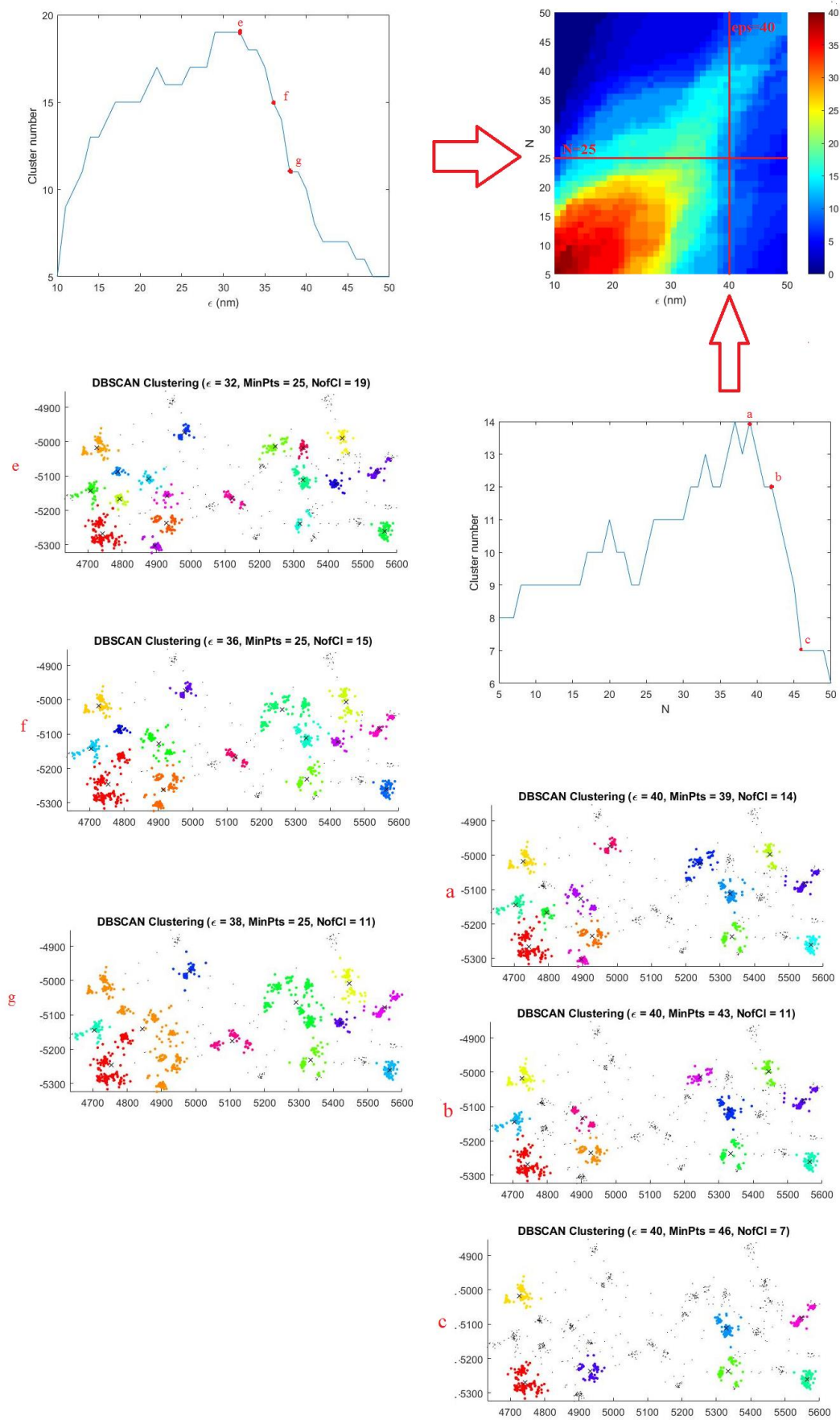


10. ábra Az $\epsilon = 40$ (bal oldal) és az $N = 25$ (jobb oldal) tartozó klaszterszámváltozás

Ezek a grafikonok azt láthatjuk, hogy ahogy nő az N értéke, úgy kezdetben a klaszterszám növekszik, majd egy érték után erősen csökkeni kezd. Hasonló tendencia figyelhető meg az epsilon változásakor. Utóbbi esetben ezt a jelenséget megtudtuk magyarázni korábban is, azonban az előbbi nem. Ezért ezen grafikonon választottam néhány N értéket a növekvő és csökkenő tartományon is, ahol az adott ϵ - N értékpárhoz tartozó DBSCAN eredményt ábrázoltam. Ehhez a rainSTORMba beépített DBSCANt használtam. A 11. és 12. ábrán láthatjuk, hogy mi történik az $\epsilon=40$ értéknél. Eleinte (11. ábra **a** részábrája) a klaszterszám 10-11 értéken van, majd ahogy közeledünk az $N=37$ ponthoz (11. ábra **c** részábrája) a klaszterszám 12, 13 és 14 értéket vesz fel. Az ábrákon láthatjuk, hogy mi okozza ezt a klaszterszám növekedést. Az $N=23$ értékhez tartozó ábrán (11. ábra **a** részábrája) azt figyelhetjük meg, hogy egy klasztert általában kettő, három kisebb klaszter alkot. A narancssárgával és a zöld színnel jelölt klasztereket, viszont ennél több. Ahogy növeljük N értékét, úgy ezeket a nagyobb összefüggő klasztereket elkezd "feldarabolni" a program több klaszterre (11. ábra **b** részábrája). Ennek a magyarázata az, hogy azokat a pontokat, amik a kisebb N értékeknél az áthidalás szerepét töltötték be, a program elkezd zajnak kezelni. (Áthidalás történik két sűrűbb terület között, ha a köztes pontokra továbbra is teljesül az N feltétel, így azok összekötik a sűrűbb területeket. Ezt a DBSCAN működésének leírásánál említett elérhetőség (density-reachable) definíció fejezi ki. Ennek lesz az eredménye az, hogy az eddig egybefüggő klaszterek több kisebb klaszterre tagolódnak. Például az $N=23$ -nál (11. ábra **a** részábrája) narancssárga klaszterből $N=37$ értékre (11. ábra **c** részábrája) négy darab klaszterre darabolódik. A zöld esetén ugyan ez a helyzet.



11. ábra A $40=\epsilon$ értékhez tartozó klaszterszámértékek változásának növekvő tendenciája



12. ábra A $40 = \epsilon$ értékhez tartozó klaszterszámtérképek változásának csökkenő tendenciája

A 12. ábrán a csökkenő tendenciához tartozó pontok vizsgálatának eredménye látható. Ezen a tartományon bekövetkező jelenség sokkal egyszerűbb, ugyanis a klaszterszám csökkenését a szigorúbb feltétel okozza. Itt már nem történik meg nagyobb klaszterek többre bontása, csak az eddig talált klaszterek száma lesz kisebb, mivel egyre több lokalizáció lesz zajpontnak besorolva.

Röviden, az ellentmondást a DBSCAN működését leíró alfejezetben már megemlített áthidalási jelenség jelenléte okozza, amit a "várható tendenciák" részben nem vettünk figyelembe.

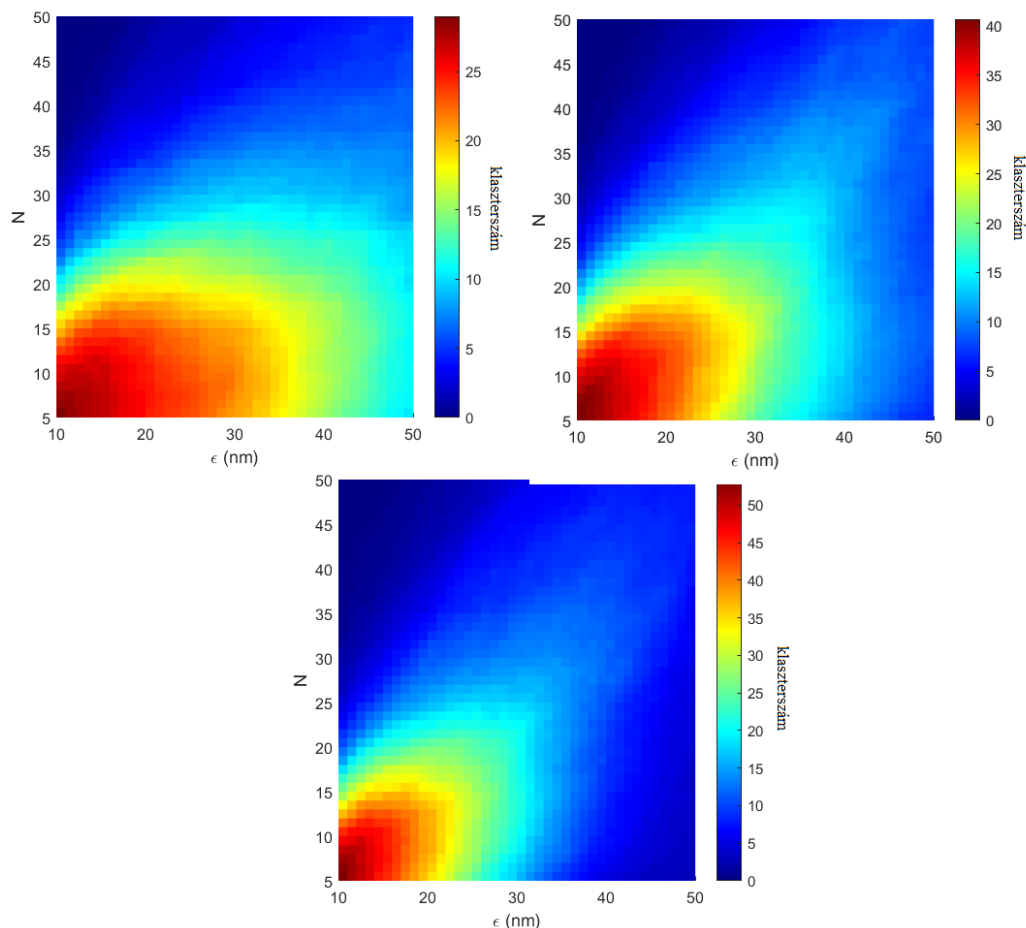
Különböző sűrűségű minták vizsgálata

Miután az előző alfejezetben leírt jelenséget megmagyaráztam, egy másik megfigyelt tendenciát is vizsgálni kezdtem.

Az általam készített 40, 60 és 80 $1/\mu\text{m}$ sűrűségű minták esetén, azt figyeltem meg, hogy a különböző sűrűségű mintákhoz tartozó térképeken megjelenő, korábban maximális érték sávnak nevezett, sáv meredeksége függ a szimulált minta sűrűségétől. Minél nagyobb az epitópok sűrűsége, annál meredekebb a megfigyelhető sáv. Azonban ahhoz, hogy ennek a megállapításnak a helyességét vizsgálhassam, az egy sűrűség értékhez tartozó tíz darab minta adatait átlagoltam, mivel az egyedi mintákra kapott eredményeket befolyásolta, hogy a véletlenszerűen elhelyezett epitópok pontosan hova is kerültek. Az átlagolással így egy általánosabb tendenciát tudtam elemezni. Ehhez összeadtam a különböző szimulációkra kapott adatokat tartalmazó mátrixokat majd elosztottam a mátrixok számával, azaz tízzel. Így kaptam egy átlagolt mátrixot. (Megjegyezném, hogy ezzel csak a talált klaszterszámok értékét átlagoltuk, a vizsgált tartomány változatlan maradt.) Ezt mindegyik sűrűség esetén elvégeztem, és az így kapott eredményeket térképeken ábrázoltam (13.ábra). Látható, hogy a 40-es sűrűséghez tartozó ábrán (13.ábra bal felső) ez a sáv pár fokot zár be az x tengellyel, majdnem vízszintes. A 60-as értéknél (13.ábra jobb felső) már egy egyértelmű meredekség figyelhető meg. Hasonlóan a 80-as értéknél (13.ábra alsó).

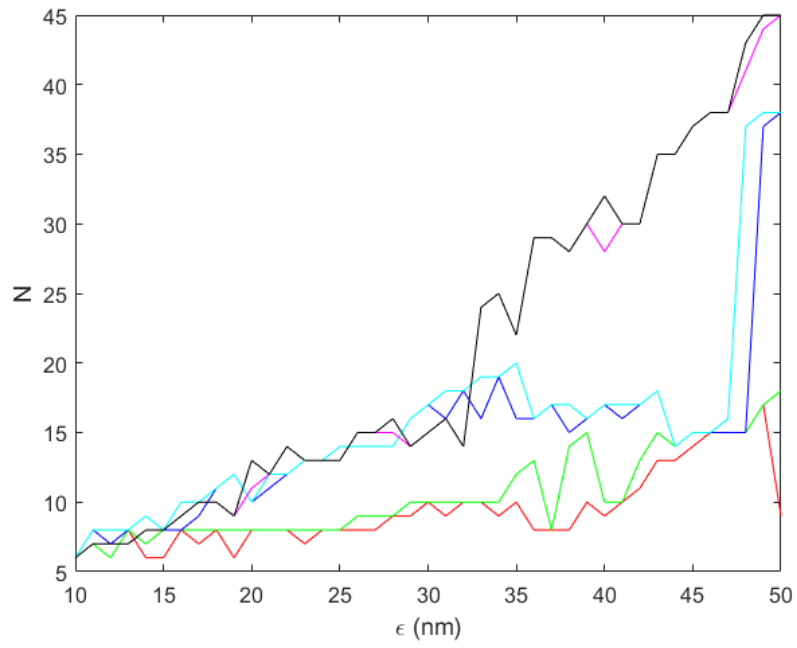
Mivel a vizsgálni kívánt sáv némely értéknél elég széles, és egy azon maximális érték több ε - N értékpárhoz is tartozhat, ezért ezt figyelembe kellett vennem. Azonban az ehhez használt program és az eredmény értelmezéséhez, fontos megemlíteni az adatokat tartalmazó átlagolt mátrixok felépítését. Mint már korábban leírtam, a mátrixok a klaszterszámokat tartalmazzák, viszont ezeknek az elemeknek a helyét különböző indexek jelölik és nem az ε és N értékek.

Például a legmagasabb értékhez az 1-1 indexpár tartozik (ez a mátrix első eleme). Az oszlopok a különböző N , a sorok az ε értékeket jelképezik. (40*45 mátrixok.)



13. ábra A különböző sűrűségekhez tartozó átlagolt térképek, Bal felső a 40-es, jobb felső a 60-as, az alsó pedig a 80-as sűrűséghez tartozó térképet ábrázolja

A program minden sorban (ε) megkeresi a legnagyobb értéket és az ahhoz tartozó indexeket (N értékeket). Az így kapott intervallumból kivesszi a legkisebb és legnagyobb indexet (N értéket), majd ezeket ábrázolja az ε függvényében. Az így kapott eredményeket a 14. ábrán láthatjuk. Észrevehető, hogy a megállapításunk, ezen minták alapján, helyesnek bizonyult, azonban a magabiztosabb állításhoz értelemszerűen több mintán való vizsgálat szükséges. Továbbá érdemes megemlíteni, hogy a 13 és 14 ábra összehasonlítása elsőre nehéz, mivel a klaszterszám értékek átlagolásának hatására, az átlagolt mátrixban megjelentek egymástól csak tizedekkel eltérő értékek is, amit a térképeken, a színazonosság miatt, nem tudunk megkülönböztetni, a program viszont igen.



14. ábra A maximálisérték sáv meredeksége különböző sűrűség értékek esetén. 40-es értékhez tartozik a zöld (legnagyobb index) és a piros (legkisebb index), a türkiz (legnagyobb index) és a sima kék (legkisebb index) a 60-as, míg a fekete (legnagyobb index) és a lila (legkisebb index) a 80-as sűrűséghez tartozó meredekségek.

Összefoglalás

Napjainkban a biológiai kutatások fontos eszköze a fluoreszcens mikroszkópia, melynek segítségével nagy kontraszttal tudunk láthatóvá tenni különböző sejtalkotókat. A kutatásoknak lendületet adott a nagyfeloldású technikák megjelenése, melyek lehetővé tették a diffrakció limit alatti struktúrák leképezését. Ezen technikák közül a lokalizációs (SMLM: single-molecule localization microscopy) technikákkal lehet elérni a legnagyobb feloldást (~10-20 nm). A lokalizációs technikákkal viszont a hagyományos pixelizált képektől eltérően a nyers adat a fluoreszcens molekulák koordinátái. Ezek a koordináták jó pontossággal megadják az általuk megjelölt fehérjék pozícióját. Mivel a sejtfunkciókat gyakran ezen fehérjék térbeli elrendeződése szabályozza, ezért az általuk létrehozott klaszterek jellemzése egy fontos feladat.

Ezen klaszterek jellemzésére használható klaszteranalízis algoritmusok egyikét, a DBSCAN algoritmust vizsgáltam. A dolgozatot áttekintése után arra a megállapításra juthatunk, hogy a DBSCAN erősen érzékeny az ε és N paraméterekre. Alacsony értékek használatakor a valódinál több és kisebb klasztert, magasabb értékek esetén kevesebb klasztert detektálhat az algoritmus. Esetünkben a legtöbbször, a DBSCAN, magasabb ε és alacsonyabb N értékpárok esetén találta meg és választotta szét helyesen a szimulált mintákat.

Továbbá vizsgáltam a minta sűrűségének hatását a klaszteranalízis eredményére, és azt a megállapítást tettem, hogy az ε - N térképen a maximális talált klaszterszám által kijelölt sávmeredeksége a sűrűséggel együtt nő. Azt is megállapítottam, hogy magasabb sűrűségű minták esetén alacsonyabb N és ε érték beállításával visszakaphatjuk az eredeti minta struktúráját (helyesen választódnak szét a klaszterek).

A szimulációs adatsorok klaszteranalízise lehetővé tette a DBSCAN algoritmus mélyebb megértését és az eredmények elemzése hozzásegít a megfelelő paraméterek kiválasztására valós mérési fájlok esetén is.

Felhasznált irodalom

- [1.] Ismail M. Khater, Ivan Robert Nabi and Ghassan Hamarneh(2020): *A Review of Super-Resolution SingleMolecule Localization Microscopy Cluster Analysis and Quantification Methods*) <https://doi.org/10.1016/j.patter.2020.100038>
- [2.] M. Marendá, E. Lazarova S. van de Linde, N. Gilbert and D. Michieletto (August 2020): *SuperStructure: a Parameter-Free Super-Structure Identifier for SMLM Data*, t doi: <https://doi.org/10.1101/2020.08.19.254540>
- [3.] Michael W. Davidson, Mortimer Abramowitz: *Optical microscopy*
- [4.] Gajdos, T. (2020) Superresolution localization microscopy using multiple modalities. Doctoral dissertation, University of Szeged, doi: <http://doktori.bibl.u-szeged.hu/10417/>
- [5.] Peter J. Verveer (2016): *Advanced fluorescence microscopy Methods and Protocols*. Springer
- [6.] Sinkó, J. et. al. TestSTORM: Simulator for optimizing sample labeling and image acquisition in localization based super-resolution microscopy, Biomed. Optics Express 5, 778-787 (2014)
- [7.] Tibor Novák, Tamás Gajdos, József Sinkó, Gábor Szabó and Miklós Erdélyi: TestSTORM: Versatile simulator software for multimodal super-resolution localization fluorescence microscopy, Sci.Rep., doi: 10.1038/s41598-017-01122-7 (2017)
- [8.] <http://laser.cheng.cam.ac.uk/wiki/index.php/Resources>
- [9.] Rees, E. J., Erdelyi, M., Schierle, G. S. K., Knight, A., & Kaminski, C. F. (2013). *Elements of image processing in localization microscopy*. Journal of Optics, Special Issue: High Resolution Optical Imaging, doi: 10.1088/2040-8978/15/9/094012
- [10.] Mickaël Lelek, MelinaT. Gyparakí, Gerti Beliu, Florian Schueder, Juliette Griffié, Suliana Manley, Ralf Jungmann, Markus Sauer, Melike Lakadamyali and Christophe Zimmer (June 2021):*Single-molecule localization microscopy*. *Nature Reviews Methods Primers*
- [11.] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. CiteSeerX 10.1.1.121.9220. ISBN 1-57735-004-9.
- [12.] Schubert, Erich; Sander, Jörg; Ester, Martin; Kriegel, Hans Peter; Xu, Xiaowei (July 2017). "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN". *ACM Trans. Database Syst.* **42** (3): 19:1–19:21. doi:10.1145/3068335. ISSN 0362-5915. S2CID 5156876.

Köszönetnyilvánítás

Ezúton szeretném megköszönni Varga Dániel témavezetőmnek minden nemű segítségét, amit a dolgozat megírása közben adott. Valamint köszönettel tartozom Dr. Erdélyi Miklós Tanár Úrnak és az AdOptIm kutatócsoport összes tagjának, hogy egy kellemes környezetben végezhettem vizsgálataimat.