A genom 3D struktúrája

Szalai-Gindl János Márk

Források:

- Hershel Safer: Three-dimensional maps of folded genomes prezentációja (http://acgt.cs.tau.ac.il/group_meeting_presentations/2010/3d%20maps%20of%20folded%20genomes%20-%20hershel%20safer.pptx)
- https://www.quantamagazine.org/genome-architecture-in-twists-of-dna-20150225
- Lieberman-Aiden, Erez, et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* 326.5950 (2009): 289-293.
- Varoquaux, Nelle, et al. "A statistical approach for inferring the 3D structure of the genome." *Bioinformatics* 30.12 (2014): i26-i33.

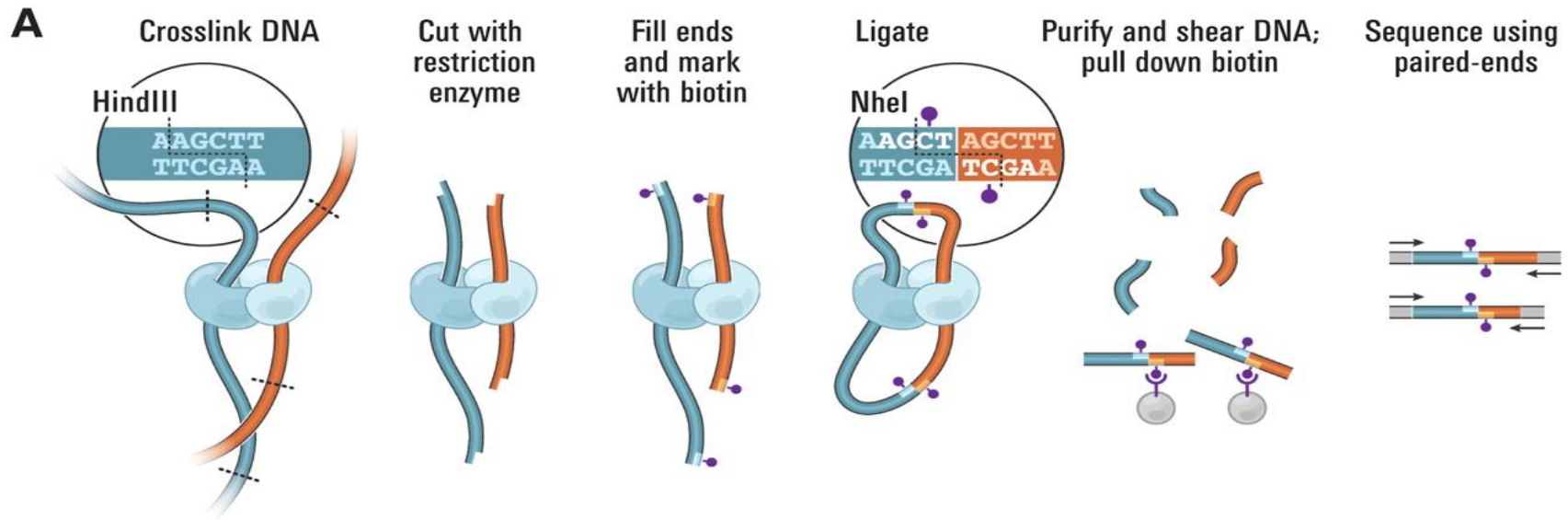
Bevezetés

- Egy élőlénynek majdnem az összes sejtje ugyanazt a DNS-t tartalmazza a sejtmagban
- A sejtek azért különböznek mégis egymástól, mert eltérő típusú sejtekben különféle gének lehetnek „bekapcsolt” vagy „kikapcsolt” állapotokban
- Ehhez viszont ismernünk kell – többek közt – azt is, hogy az egydimenziós szekvencia, hogyan tekeredik, illetve aggregálódik a sejtmagon belül
- Hiszen a sejtmagban lévő DNS funkcionális elemei csak akkor tudnak interakcióba lépni, ha fizikailag közel vannak
- Tehát ismernünk kell a genom háromdimenziós szerkezetét is

Bevezetés

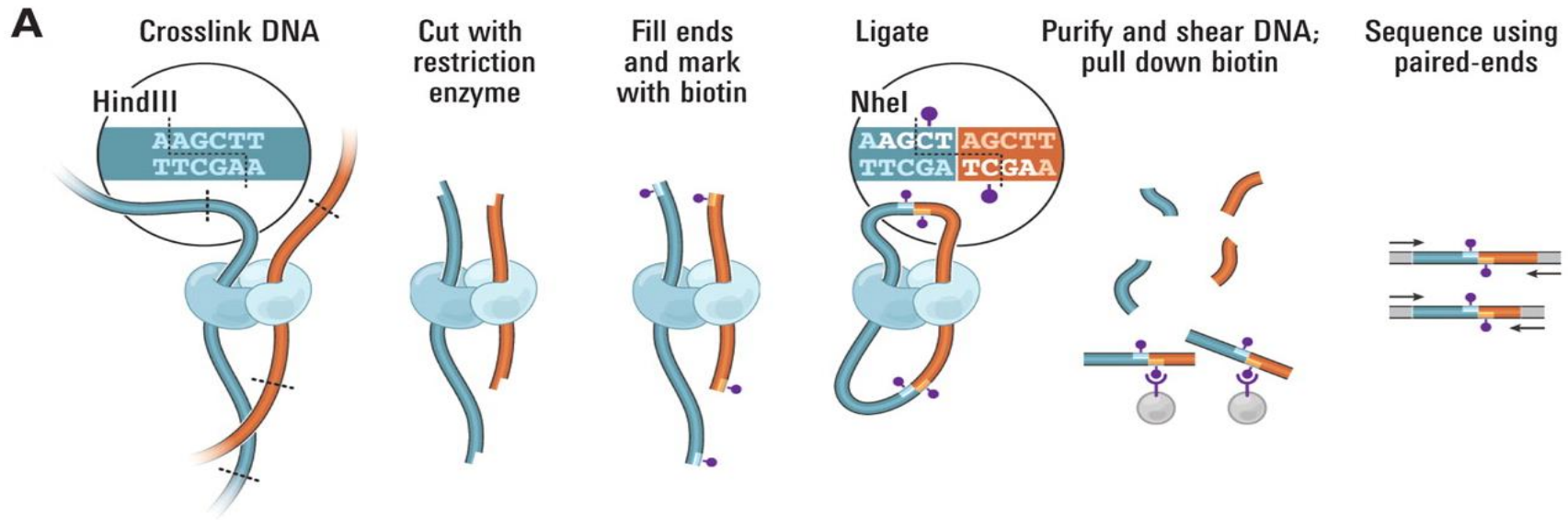
- A 90-es évek elején Katherine Cullen és csapata kifejlesztett egy módszert arra, hogy a sejtmagban közeli DNS részek mesterségesen összeolvadjanak
- Ez lehetővé tette, hogy az „összehajtogatott” DNS struktúrát elemezni lehessen szekvencia leolvasás alapján
- Ez a technika évek alatt továbbfejlődött → az egyik legújabb változat a Hi-C, amely segítségével teljes genomok háromdimenziós feltérképezését lehet megtenni

Biológiai háttér



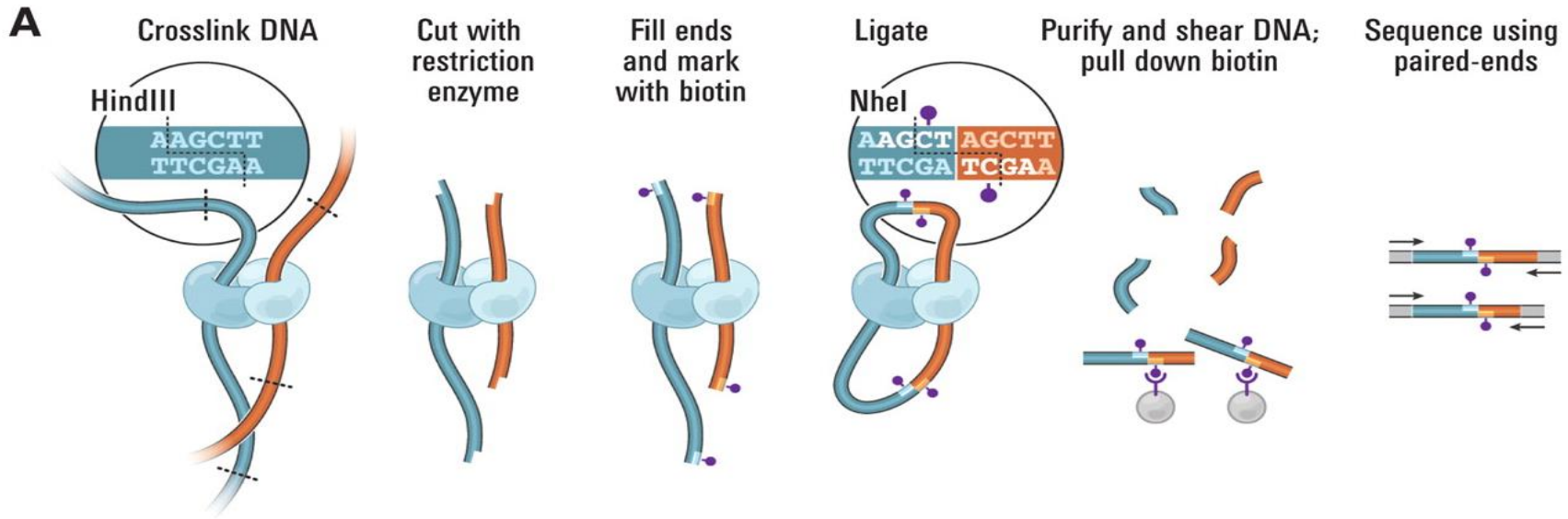
- A Hi-C kísérletnél az első lépése, hogy a mintában szereplő sejtek millióit formaldehiddel kezelik
- Ennek az a kémiai hatása, hogy az egymáshoz közeli DNS szálak között keresztirányú kötődések lesznek

Biológiai háttér



- Ezután felvagdalják a genomot, learatják az összekötődött darabkákat, amelyeket megszekvenálnak (ún. „paired-end” szekvenálással)
- A szekvenált darabkák olyanok, mint egy közeli fotó a DNS-DNS érintkezésekről a 3D-s genomban

Biológiai háttér



- Ezeket a darabokat hozzáillesztik egy referencia genomhoz, amelynek segítségével listázni lehet a genom érintkezési pontjait
- Végül egy olyan mátrixot kapunk: *DNS kapcsolat térképet*, amely a DNS pozíciói között jelöli az interakciók gyakoriságát egy adott felbontásnál

Módszerek

- Az alapvető kérdés: hogyan rekonstruáljuk a genom 3D struktúráját ebből a kapcsolat térképből?
- Két általános megközelítés van erre:
  - *Konszenzus módszerek*: egyértelmű átlag struktúrára következtetnek, amely reprezentatív az adatokra nézve
  - *Együttes módszerek*: struktúrák populációját hozzák

Módszerek

- Konszenzus módszerek:
  - A kapcsolat térkép gyakoriságait páronkénti távolságokra: *kívánt távolságokra* konvertálják (különböző biofizikai modelleket használva)
  - Következtetnek arra a térbeli elhelyezkedésre, amely a legjobban illeszkedik a páronkénti távolságokra megoldva egy *multidimensional scaling (MDS)* feladatot
  - A konvertálás azonban változhat pl. a különböző felbontások alkalmazásánál, különböző organizmusoknál stb.

Módszerek

- Konszenzus módszerek:
  - A probléma enyhítésére született meg a ChromSDE módszer (Zhang et al. (2013)), amely képes egyszerre optimalizálni a 3D struktúrát és annak a függvénynek a paramétereit, amely gyakoriságokat térbeli távolságra képzi
  - Ben-Elazar et al. (2013) egy olyan módszert javasolnak, - amely rokon jellegű a *nem-metrikus MDS*-ekhez, - és ahol a 3D struktúra és a kívánt távolságok váltakozva lesznek optimalizálva

Módszerek

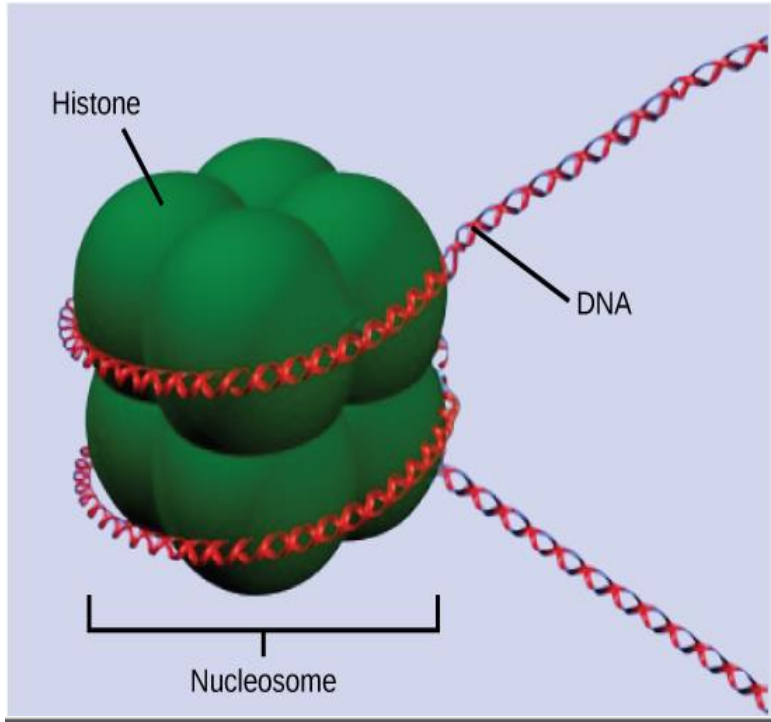
- Együttes módszerekre példa:
  - Hu et al. (2013) és Rousseau et al. (2011) két formális valószínűségi modellt írnak le az érintkezési gyakoriságokra, és azoknak a fizikai távolságokkal való kapcsolataira
  - Ezután egy Markov-lánc Monte Carlo (MCMC) mintavételezési eljárást alkalmaznak, hogy a 3D struktúráknak egy együttesét produkálják, amelyek konzisztensek a megfigyelt érintkezési számokkal

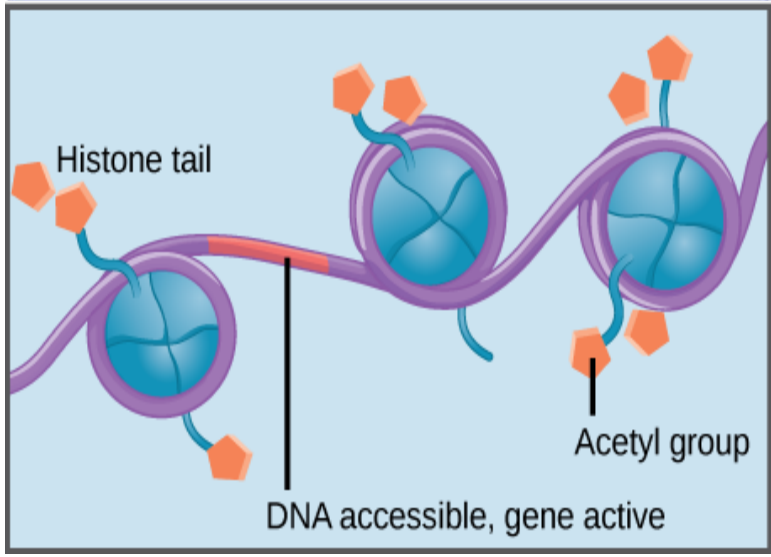
Módszerek

- A konszenzus- és az együttes módszereknek is megvannak az előnyei és korlátjai:
  - Az együttes módszerek biológiailag pontosabbak, mert a Hi-C adatok a sejtek egy populációjából származnak, amelynek mind megvan a saját 3D struktúrája
  - Viszont az interpretálhatóság szempontjából kérdéses
  - A konszenzus módszerek ezzel szemben egy egyedüli struktúrát nyújtanak, amellyel a vizuális vizsgálat és elemzés könnyebb
  - Továbbá bizonyos tulajdonságok megegyeznek a különböző sejttípusok között is
- A konszenzus módszerekre fogunk a továbbiakban fókuszálni

Módszerek

- A kromoszómákat „gyöngyfüzér” struktúraként modellezzük:
  - Azaz mindenegyes gyöngy egy adott hosszúságú genomikai ablakot reprezentál
  - $x_i \in \mathbb{R}^3$ az i. gyöngy koordinátái
  - Legyen n az összes gyöngy száma a genomban
  - $\boldsymbol{X} = (x_1, \ldots, x_n)$ a szerkezet koordináta mátrixa
- A Hi-C adatok egy $n \times n$-es $\boldsymbol{c}$ mátrixban vannak összegezve:
  - ahol minden sor és oszlop megfelel egy genomikai helynek,
  - és a mátrix $c_{i,j}$ eleme az érintkezési gyakoriság





Kép forrása: https://archive.cnx.org/contents/112da94a-d347-4695-af35-9c97314f3e84@6/eukaryotic-epigenetic-gene-regulation

MDS-alapú módszerek – Metrikus MDS

- A metrikus MDS egy klasszikus módszer, hogy következtessünk a pontok koordinátáira úgy, hogy páronként adottak a hozzávetőleges Euklideszi távolságok
- Esetünkben minden (i, j) gyöngypárhoz szeretnénk egy $\delta_{i,j}$ fizikai kívánt távolságot rendelni a $c_{i,j}$ elemek alapján
- Biofizikai megfontolások alapján a $\delta_{i,j} = \gamma {c_{i,j}}^{-3}$ (ha $c_{i,j} > 0$) transzformációt szokták alkalmazni

MDS-alapú módszerek – Metrikus MDS

- A metrikus MDS ezek után a gyöngyöket úgy fogja 3D térbe elhelyezni, hogy az Euklideszi távolság az i. és j. gyöngy között: $d_{i,j}(\boldsymbol{X}) = \|x_i - x_j\|$ annyira közel legyen $\delta_{i,j}$-hez, amennyire csak lehet
- Ha $\mathcal{D}$ jelöli az indexeknek azt a részhalmazát, amelynek a távolságaira szeretnénk kényszereket (tipikusan, ahol $c_{i,j} > 0$), a metrikus MDS az alábbi célfüggvényt fogja optimalizálni:

$$\min_{\boldsymbol{X}} \sum_{(i,j)\in\mathcal{D}} (d_{i,j}(\boldsymbol{X}) - \delta_{i,j})^2$$

MDS-alapú módszerek – Metrikus MDS

- Az előző változattal az a probléma, hogy a nagy értékek dominálják, ahol igazából kicsi az érintkezési gyakoriság
- De ezek az értékek kevésbé megbízhatóak, mint ahol nagyobb az érintkezési gyakoriság (azaz ahol közel vannak egymáshoz), ezért érdemes súlyozni:

$$\min_{\boldsymbol{X}} \sum_{(i,j)\in\mathcal{D}} {\delta_{i,j}}^{-2} \left(d_{i,j}(\boldsymbol{X}) - \delta_{i,j}\right)^2$$

MDS-alapú módszerek – Nem-metrikus MDS

- Az előző módszernél láttuk, hogy erős feltételezésünk van a DNS fizikájáról, amikor az érintkezési gyakoriságokat átvittük a kívánt távolságokra
- A nem-metrikus MDS (NMDS) egy alternatívát nyújt: egyedül azon a feltevésen nyugszik, hogy ha két genomikai pozíció között nagyobb az érintkezési gyakoriság, akkor közelebb vannak egymáshoz 3D-ben
- Ekkor a feladat: adott $c_{i,j}$ hasonlóságoknak egy halmaza (azaz az érintkezési gyakoriság $i.$ és $j.$ gyöngy között), keressük $\boldsymbol{X} \in \mathbb{R}^{3\times n}$ úgy, hogy:

$$c_{i,j} \geq c_{k,l} \Leftrightarrow \left\|x_i - x_j\right\|_2 \leq \|x_k - x_l\|_2$$

MDS-alapú módszerek – Nem-metrikus MDS

- A fentit meg lehet úgy oldani, hogy az alábbit minimalizáljuk:

$$\min_{X,\Theta} \sum_{i,j} \frac{\left(\left\|x_i - x_j\right\|_2 - \Theta(c_{i,j})\right)^2}{\Theta(c_{i,j})^2}$$

- Ahol Θ egy csökkenő függvény
- Az algoritmus, amely megoldja ezt, két lépésből áll:

1. Rögzítjük Θ-t és minimalizáljuk a célfüggvényt $\boldsymbol{X}$-re vonatkozóan
2. Illesztjük Θ-t az új $\boldsymbol{X}$ konfigurációhoz (izotóniás regresszió módszerrel)

Módszerek – Poisson modell

- Az MDS módszereken kívül még létezik például olyan módszer is, ahol a struktúrára következtetés maximum likelihood problémaként van feltüntetve
- Ehhez a az érintkezési gyakorisságokra kell egy valószínűségi modellt felállítani, amely a 3D struktúrával van paraméterezve, amelyre az érintkezési gyakorisságok megfigyeléseiből szeretnénk következtetni
- Az érintkezési gyakorisságok független Poisson valószínűségi változókként vannak modellezve: $\mathbb{P}\left(c_{i,j} = k\right) = \frac{\lambda^k}{k!} e^{-\lambda}$, ahol $\lambda = \beta d_{i,j}(\boldsymbol{X})^{\alpha}$ (valamely $\beta > 0$ és $\alpha < 0$)

Módszerek – Poisson modell

- Ki tudjuk fejezni a likelihoodot az alábbi módon:

$$\ell(\boldsymbol{X}, \alpha, \beta) = \prod_{i,j} \frac{(\beta d_{i,j}(\boldsymbol{X})^{\alpha})^{c_{i,j}}}{c_{i,j}!} \exp(-\beta d_{i,j}(\boldsymbol{X})^{\alpha})$$

- Ezt kell elvileg maximalizálni $\boldsymbol{X}, \alpha, \beta$-ben (valójában α-nál apriori tudást is felhasználhatunk, és választhatjuk konstansnak)
- A módszerek összehasonlításánál szimulációs adatoknál általában a Poisson modellt használó módszerek felülmúlják a többit a valódi szerkezet és a becsült szerkezet összehasonlításánál (Varoquaux et al. (2014))

Köszönöm a figyelmet!