



IKER
GAZTE
NAZIOARTEKO
IKERKETA EUSKARAZ

VI. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)



Aitortu-PartekatuBerdin 4.0

INGENIARITZA ETA ARKITEKTURA

**Heterogeneotasunaren analisi
azkar eta automatizatua krio-
mikroskopia elektronikoan**

*Oier Lauzirika Zarrabeitia,
José Luis Vilas Prieto,
Mikel Iceta Tena,
José María Carazo Garcia eta
Carlos Oscar Sorzano Sanchez*

109-116 or.

<https://dx.doi.org/10.26876/ikergazte.vi.03.13>

ANTOLATZAILEA



BABESLEAK



EUSKO JAURLARITZA
GOBIERNO VASCO



LAGUNTZAILEAK



Heterogeneotasunaren analisi azkar eta automatizatua krio-mikroskopia elektronikoan

Oier Lauzirika¹, José Luis Vilas¹, Mikel Iceta¹, José María Carazo¹, Carlos Oscar Sorzano¹

Centro Nacional de Biotecnología. C. Darwin, 3, Fuencarral-El Pardo, 28049 Madrid

olauzirika@cnb.csic.es

Laburpena

Krio-Mikroskopia Elektronikoan (CryoEM) laginaren egitura heterogeneotasuna erronka nagusia da. Honetarako, ondoko norabideetatik proiektatutako irudien arteko antzerakotasunaren abantaila hartzen duen metodo bat proposatzen dugu. Honetarako, Osagai Nagusien Analisia (PCA) burutzen da datuen norabide-azpimultzoekin. Analisi lokal hauek modu globalean sinkronizatzen dira Talde Ortogonalen Sinkronizazioaren bidez, matematika aplikatuen arloko eremu berri bat. Heterogeneotasun analisi ikuspegi honek parametro gutxi behar ditu exekutatzeko, eskuzko doipenak murriztuz eta erabilgarritasuna indartuz. Honez gain, parametroek zentzu fisiko-estadistikoa dute, hauen ulergarritasuna eta gardentasuna bermatuz. Beste alde batetik, algoritmoa konputazionalki oso eraginkorra da, ohiko heterogeneotasun-analisi metodoekin alderatuta exekuzio-denbora nabarmen azkarragoak lortuz. Metodoa balioztatzeko, heterogeneotasun konposizionala eta konformazionala erakusten duten datu-simulazio eta esperimentaletan probatzen dugu, egitura-aldakortasuna argitzeko duen eraginkortasuna erakutsiz.

Hitz gakoak: Krio-Mikroskopia Elektronikoa, Irudi prozesamendua, Partikula Isolatuen Analisia, Multzokatzea, Talde ortogonalen sinkronizazioa

Abstract

Understanding structural heterogeneity is a key challenge in Cryo-Electron Microscopy (Cryo-EM). We propose a novel method that leverages the natural similarity between neighboring projection directions to perform a localized Principal Component Analysis (PCA) in directional subsets of the data. These local analyses are then globally synchronized using Orthogonal Group Synchronization, a novel field in applied mathematics. Our approach is designed to pose very few execution parameters, reducing the need for extensive user tuning. The parameters involved have clear physical and statistical interpretations, enhancing their usability and transparency. In addition, it is also computationally efficient, achieving significantly faster execution times compared to conventional heterogeneity analysis methods. We validate our approach on both simulated and experimental datasets exhibiting compositional and conformational heterogeneity, demonstrating its effectiveness in capturing structural variability.

Keywords: Cryo-Electron Microscopy, Image processing, Single Particle Analysis, Clustering, Orthogonal group synchronization

1 Sarrera eta motibazioa

Azken urteetan, Krio-Mikroskopia Elektronikoak (CryoEM) biologia estrukturala irauli du, makromolekula biologikoak bereizmen atomikoarekin bistaratzea ahalbideratuz. Izan ere, ikertzaileek proteinen, birusen eta beste hainbat makromolekulen egitura hiru-dimentsional atomikoa atzeman dezakete egoera ia naturalean (Nogales eta Scheres, 2015).

Testuinguru honetan, Partikula Isolatuen Analisia (SPA) lagin preparazio eta irudi prozesamendu teknika multzoa da, non interesezko proteinaren milaka kopien irudiak transmisio mikroskopia elektroniko baten bidez biltzen diren. Kopia hauetako bakoitzari “partikula” deritzo. Partikula hauek izotz bitreo geruza mehe batean prestatzen dira, mikroskopioko egoera latza iraun ahal izateko. Teknika honek 2017ko Kimika Nobel Saria eskuratu zuen (Cressey eta Callaway, 2017).

Lagina izozterakoan, partikula bakoitzak ausazko orientazio bat hartzen du. Honela, irudi prozesamendu eta teknika konputazionalen bidez, orientazio ezberdinetan lortutako irudiak batzen dira, laginaren 3D egitura bakar bat argituz (Vilas *et al.*, 2022).

CryoEM-en erronka nagusietako bat lagin biologikoen heterogeneotasuna da, non laginaren hainbat konformazio edo konposizio ager daitezke datu-multzo batean (Jonić, 2016). 3D sailkapena eta heterogeneotasunaren analisia erronka honi aurre egiteko bi estrategia dira. Izan ere, teknika hauek datuen 3D izaera aprobetxatzen dute, irudiak jatorrizko egituraren arabera sailkatuz. Hala ere, sailkapen hau ez da erraza, egitura aldaketak ezezagunak baitira. Ondorioz, irudiak irizpide ezkutuen arabera sailkatu behar dira. Honez gain, irudien Seinale-Zarata Ratioa (SNR) oso kaxkarra da, 1/100 ingurukoa. Hau da, seinalea baino 100 aldiz zarata gehiago dute irudiek (Li *et al.*, 2022).

Artikulu honetan heterogeneotasun algoritmo berri bat proposatu nahi dugu. Proposatutako algoritmo hau datu simulatuekin zein datu esperimentalarekin probatuko da, honen eraginkortasuna neurtu ahal izateko.

2 Arloko egoera eta ikerketaren helburuak

Azken hamarkadan aurrerapen garrantzitsuak egon diren arren, egungo 3D sailkapen algoritmoek hainbat ahultasun dituzte. Metodo ohikoenetarikoa bat Txandakako Maximizazioa (AM) da, hau da, modelo estatistiko baten parametro ezezagunak iteratiboki ebazten dituen metodoa (Kimanius *et al.*, 2021). AM 3D sailkapena egiteko erabiltzen denean, irudi bakoitza hainbat erreferentziarekin alderatzen da, modelo bakoitzaren egiantzekotasuna kalkulatzeko. Ondoren, erreferentzi bakoitza irudi egiantzokorrekin ber-erakitzen da. Azkenik, eraiki berritako egitura hauek hurrengo iterazioan eredu gisa erabili daitezke. Prozesu honen hainbat errepikamen egin ostean, erreferentzia bakoitzaren atal bereizgarrienak indartzea espero da. Hala ere, AM metodo lokala da. Beraz, azken emaitzak hasierako proposamenarekiko zehertasuna duela esan omen da literaturan. Aldi berean, hasierako proposamena ausaz aukeratzen da, AM algoritmoaren konbergentzia baldintzatuz (Sorzano *et al.*, 2022; Jonić, 2016).

Problema hau ebazteko beste ohiko ikuspuntu bat gainbegiratze gabeko sare-neuronalak dira (Herreros *et al.*, 2023). Hala ere, metodo hauek beste hainbat arazo propio dakartzate. Lehenik eta behin, konputazio kostu handia dute, datu-bilduma bakoitzerako entrenatu behar baitira. Prozesu honek hainbat ordu edo egun suposa ditzake. Honez gain, sare hauek haluzinazioak izan ditzakete, zentzu biologiko edo esperimental gabeko soluzioak proposatuz. Azkenik, sare-neuronalen izadiagatik, beraien auresanen eta datuen arteko erlazioa ezin daiteke ulertu -kutxa beltza- eta hiper-parametroen esangura ez da interpretagarria.

Arazo hauek ekiditzeko, heterogeneotasun analisi algoritmo berri bat proposatu nahi dugu artikulu honetan. Algoritmo honen helburu nagusia parametro kopurua minimizatzea da, hauen ulergarritasuna baldintzatu gabe. Honela, programaren erabilera erraztea eta prozesamendua automatizatzea lortuko da. Honez gain, baliabide konputazionalen etekin ahalik eta onena lortzea nahi da, gainontzeko algoritmoak baino arinagoa izanik.

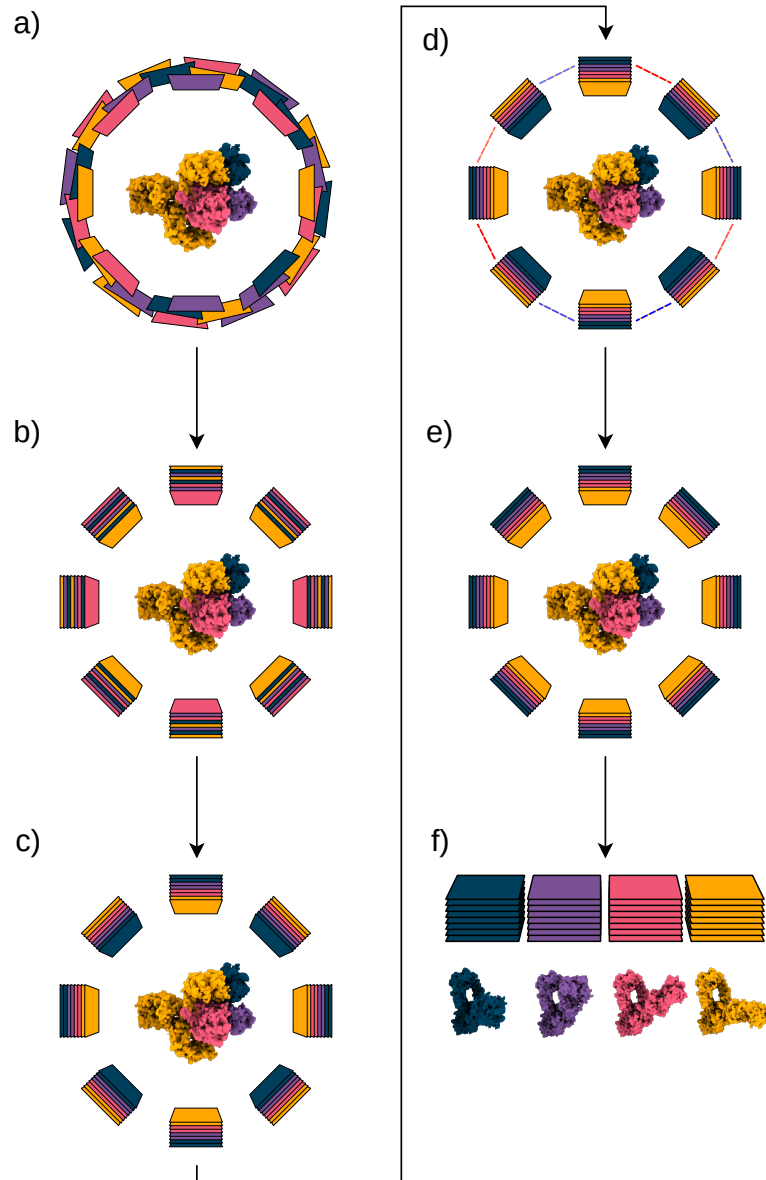
3 Ikerketaren muina

Artikulu honetan heterogeneitate analisi algoritmo berri bat proposatu nahi dugu. Heterogeneitate analisia CryoEM irudi prozesamendu lan-fluxuko azken pausuetako bat da. Beraz, aurreko pausuetan estimatutako hainbat parametroz balia daiteke. Algoritmo honen sarrera irudi datu-multzo heterogeneo bat da. Hau da, laginaren hainbat konposizio edo konformaziotatik lortutako proiektzio nahaste bat. Helburua proiektzio hauek jatorrizko 3D egituraren arabera sailkatzea da. Honetarako tarteko pausu bat dimentsio txikiko konformazio espazio bat lortzea da, non antzeko egiturak hurbil kokatzen dira.

3.1 Algoritmoaren deskribapen zehatza

Metodo hau hurrengo pausuetan banatu daiteke. Pausu hauek xehetasun handiagoarekin deskribatuko dira geroago. Honez gain, 1. Irudiak pausu hauek grafikoki deskribatzen ditu.

1. Algoritmoaren sarrera irudi datu-multzo heterogeneo bat da. Aurreko pausuek irudi hauen proiektzio norabideak estimatu dituzte (a atala 1 irudian).
2. Irudiak aurretik estimatutako proiektzio norabidearen arabera multzokatu. Multzo bakoitzean, irudiak beraien artean lerrokatu hauen arteko Osagai Nagusien Analisisa (PCA)-ren bidez aldakortasuna estimatu ahal izateko.

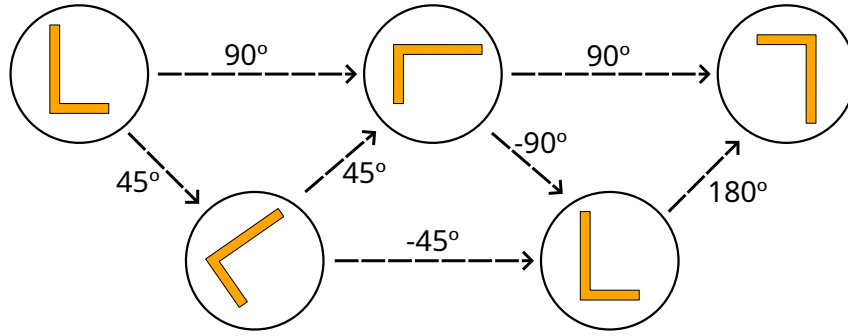


1. Irudia: Metodoaren laburpen grafikoa

(b eta c atalak 1 irudian).

3. Multzo bakoitza ondoko multzoekin alderatu beraien arteko koherentzia ezagutzeko (d atala 1 irudian).
4. Parekako koherentzia ezagutuz, PCA oinarriak globalki sinkronizatu (e atala 1 irudian).
5. Multzo desberdinetako PCA koefizienteak bateratu eta irudiak sailkatu (f atala 1 irudian).

Algoritmo honek ondoko norakoetatik proiektatutako irudien antzekotasunaz baliatzen da. Izan ere, ondoko irudien arteko ezberdintasunak laginaren heterogeneotasunari soilik egotzi dakizkioke. Hori dela eta, lehenengo



2. Irudia: Talde ortogonalen sinkronizazioaren adibide sinplifikatua

pausua irudiak proiektzio norabidearen arabera sailkatzea da. Proiektzio norabide hauek irudi-prozesamenduaren aurreko pausuek estimatu dituzte. Talde hauek osatzeko, garrantzitsua da dibertsitate angeluar txikia izatea, hauen aldakortasun osagai nagusia laginaren heterogeneitatea izateko. Honez gain, geroago argituko diren arrazoiengatik, talde hauek gainjarri egin behar dira. Hau da, irudi bat hainbat taldetan egon daiteke.

Behin taldeak ezarrita daudela, irudiak beraien artean lerrotatu daitezke, beste behin ere, aurreko irudi-prozesamendu pausuek estimatutako proiektzio parametroak erabiliz. Honela PCA analisi bat egin daiteke talde bakoitzean. Honetarako, irudiak bektore gisa adierazi daitezke, pixel lerroak bata bestearen ondoren kateatuz. Lortutako PCA koefizienteek espazio latente bat desestaltzen dute, potentzialki laginaren heterogeneotasuna erakutsiz.

Norabide bakoitzean lortutako PCA espazio latenteak beraien artean antzekoak izan beharko lirarteke, laginaren heterogeneotasuna berdina baita denentzako. Hala ere, PCA-ren soluzioa ez da bakarra, ardatzen noranzkoa ez baitago definituta. Are gehiago, ikusitako heterogeneotasunaren handitasuna norabide proiektzioekiko mendekoa izan daiteke, beraz, ardatzen ordena ere alda daiteke taldeen artean. Hau guztia kontuan izanik, hurrengo pausua PCA talde bakoitzeko PCA ardatzak globalki sinkronizatzea da, beraien koefizienteak espazio bakar batean batu ahal izateko. Honegatik, hasieran talde gainjarriak definitu dira. Bi talde, beraien arteko irudi komunak erabiliz alderatu daitezke. Konkretuki, aldameneko taldeen arteko kobariantza matrizea kalkulatu da.

Talde batzuen arteko kobariantza jakinik, (1) optimizazio problema eratu daiteke globalki bateragarria den soluzio bat aurkitzeko.

$$\max_{\hat{Q}} f(\hat{Q}) = \sum_{i,j \in \mathcal{G}} \langle \Gamma_{ij}, \hat{Q}_i \hat{Q}_j^\dagger \rangle \quad (1)$$

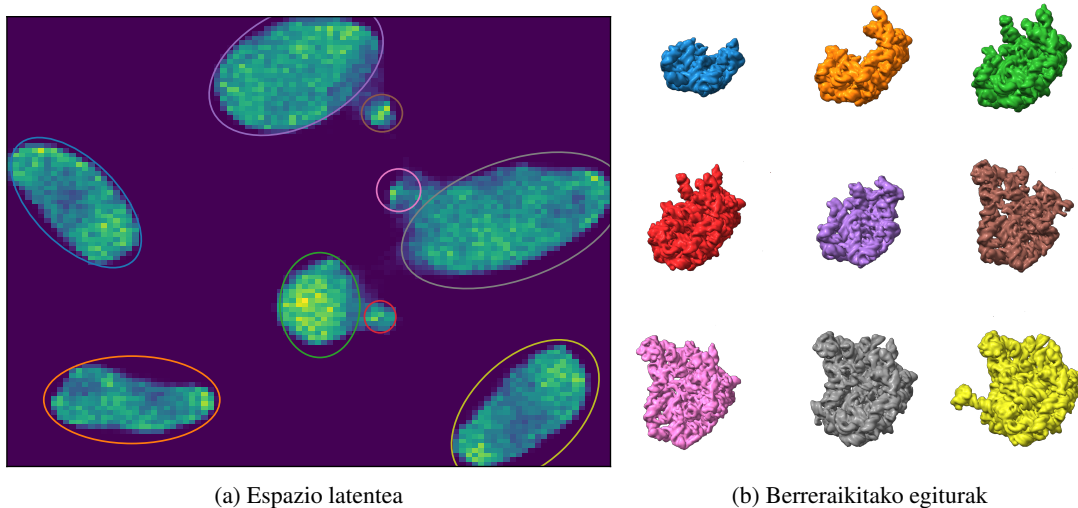
ondoko baldintzapean:

$$\hat{Q}_i \in O(d)$$

non $\hat{Q}$ -ren blokeak PCA ardatzak sinkronizatzeko transformazio ortogonalak dira, Γ_{ij} i eta j multzoen arteko kobariantza matrizea da eta $\mathcal{G}$ aldameneko proiektzio norabideen grafoa da. Optimizazio arazo hau “Talde ortogonalen sinkronizazio” modura ezagutzen da matematika aplikatuaren eremuan (Liu *et al.*, 2023). Arazo honen soluzioa erabiliz, norabide bakoitzean lortutako PCA koefizienteak sinkroniza eta bateratu daitezke, espazio latente bakar bat lortuz.

Sinkronizazio hau hobeto ulertzeko, 2 irudiko adibide sinplifikatua erabiliko da. Suposa dezagun forma baten hainbat banako ditugula, ausazko orientazioarekin. Honez gain, hainbat banakoren arteko orientazio erlatiboa neurtu dezakegu. Talde ortogonalen sinkronizazioaren helburua parekako informazio honen bidez banako guztiak globalki lerrotatzea da.

Azkenik, espazio latente hau hainbat modutara zatikatu daiteke, irudiak konformazio edo konposizioaren arabera bereiziz. Adibidez, laginak heterogeneitate konposizionala badu, konformazio espazioak talde diskretuak izatea espero da. Honela, Nahasketa Gaussiarraren Modelo (GMM) bat erabili daiteke talde hauek bereizteko. Beste alde batetik, laginak heterogeneitate konformazionala badu, espazio koformazionalak barietate bat erreprezentatuko du. Barietate hau k -Auzokide Hurbilenak (kNN) bidez banatu daiteke laginaren hainbat konformazioen ereduak lortzeko.



3. Irudia: “Ribosembly” datu simulatuekin lortutako emaitzak

3.2 Emaitzak

Algoritmoaren baliagarritasuna egiaztatzeko, hainbat datu-multzorekin balioztatu da. Alde batetik, datu simulatuak erabili dira, aldagai esperimentalak kontrolpean izateko. Beste alde batetik, biltegi publikoetan gordailututako egiazko datuak erabili dira emaitza errealistak lortu ahal izateko. Honez gain, heterogeneitate konposizionala zein konformazionala duten datuak hautatu dira, bi kasuetako balioa ezagutzeko.

Lehenik eta behin CryoBench frogako “Ribosembly” datu multzoarekin probatu da. Datu multzo simulatu honek Ribosmaren 16 muntai egoera ditu, heterogeneotasun konposizionalaren froga on bat izanik (Jeon *et al.*, 2024). Gure algoritmoa erabiltzean, 3. Irudian bereiz daitekeen bezala, 9 egoera esanguratsu aurkitu dira. Gainontzeko 7ak populazio oso txikia dute edo aldaketa oso txikiak deskribatzen dituzte. Emaitza hauen arabera, metodo honek egoera garrantzitsuenak lehenesten dituela ondorioztatu daiteke.

Beste alde batetik, froga berdineko “IgG-1D” datu multzo simulatua erabili da heterogeneotasun konformazionala ikertzeko (Jeon *et al.*, 2024). 4. Irudian ikus daitekeen moduan, espazio latentea eratzun formako barietate bat da. Izan ere, IgG-1D antigorputzaren Fab eremuak birakako mugimendu bat deskribatzen du. Eratzun honen ebaketak hartuta eta partikula hauekin eraikitzen, biraketa honetako konformazio ezberdinak bereiz daitezke 4. Irudiko b atalean.

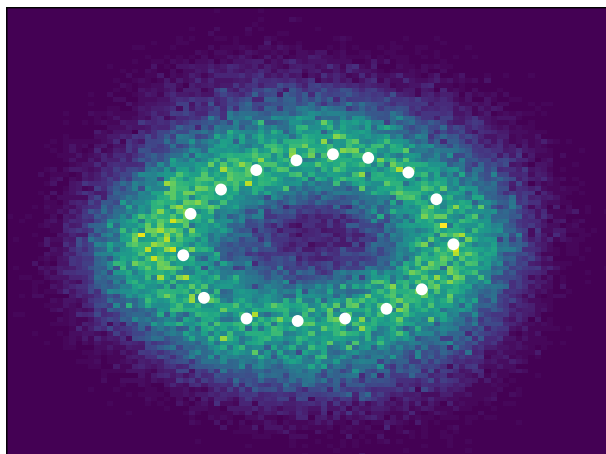
Datu esperimentalei dagokionez, lehen froga EMPIAR-10076 datu-multzo publikoarekin egin da (Davis *et al.*, 2016). Datu-multzo honek 50S erribosomaren hainbat konposizioen nahastea dauka. 5. Irudian ikus daitekeen moduan, gure algoritmoa erabiliz, 4 konformazio esanguratsu aurkitu dira, nahiz eta benetako datu-multzoak 6 izan. “Ribosembly” datu-multzoan bezala, nabaritu ezin izan diren 2 konposizioak oso populazio txikia dute.

Azkenik, EMPIAR-10180 datu-multzo esperimentalean, Splizeosomaren (Plaschka *et al.*, 2017) “buruaren” flexibilitatea antzeman dugu, espero diren eraldaketa biologikoekin lerrotatuz. Honek, gure irizpidea makromolekula dinamikoak aztertzeke egokia dela erakusten du.

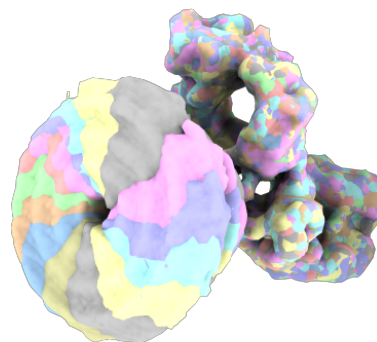
Eraginkortasun konputazionalari dagokionez, emaitza hauen prozesamendurako ordu bat baino gutxiago behar izan da datu multzo bakoitzarentzako 2xIntel Xeon E5-2687W eta 2x NVIDIA GTX1080 dituen makina batean. Honekin alderatuz, egungo algoritmo gehienek hainbat ordu behar dituzte antzeko emaitzak burutzeko.

3.3 Kodearen eskuragarritasuna

Deskribatutako kodea Xmipp (<https://github.com/I2PC/xmipp>) (Sorzano *et al.*, 2004) mikroskopia irudi prozesamendu paketea garatu da. Xmipp GPLv3 lizentziarekin banatzen da, beraz, software hau askatasunez erabili, aldatu eta banatu daiteke.

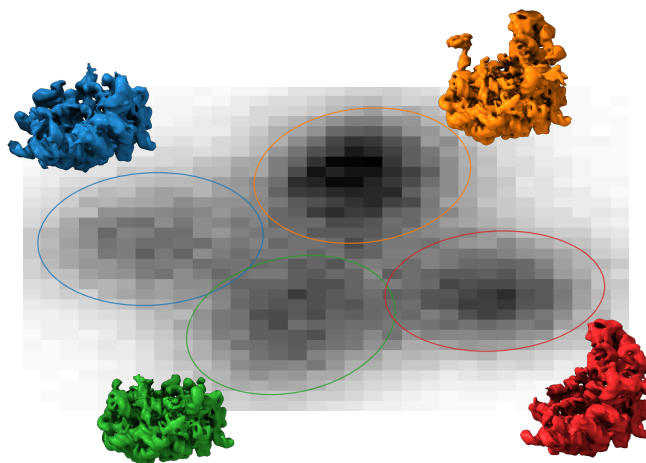


(a) Espazio latentea



(b) Berreraikitako egiturak

4. Irudia: IgG antigorputzaren datu simulatuekin lortutako emaitzak



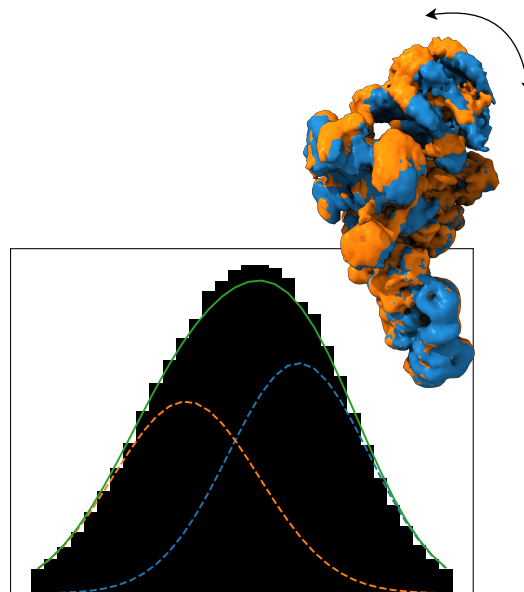
5. Irudia: L17 gabeko 50S Erribosomaren bitartekariekin lortutako emaitzak

4 Ondorioak

Aurreko atalean erakutsitako emaitzek metodo honen eraginkortasuna frogatzen dute hainbat laginen aldakortasun estrukturala kapturatzeko. PCA eta Talde Ortogonalen Sinkronizazioa erabiliz, ikuspegi honek emaitza bizkor eta sendoak burutzen ditu. Honez gain, parametro kopurua txikia izanez eta hauen esangura fisiko eta estatistikoa bermatzean, oso erabilgarria da biologoentzako. Datu simulatuen zein esperimendalekin egindako frogek konposizio eta konformazio aldagarritasuna bereizteko gaitasuna berresten dute, biologia estrukturalaren ikerketarako tresna indartsu izateko ahalmena nabarmenduz.

5 Etorkizunerako planteatutako norabidea

Testu hau Partikula Isolatuen Analisia (SPA) teknikaren inguruan orientatu da. Hala ere, CryoEM-en beste korrante nagusi bat dago, Subtomogramen Batezbestekoa (STA) deritzona. Teknika honetan, lagina hainbat inklinazio angelutan eskuratu da, irudien arteko hainbat murrizketa geometriko ezarri. Beraz, hurrengo pausua metodo hau STA teknikara zabaltzea da, honek dakartzan inbariantza geometrikoak kontuan hartuz. Honez gain, froga zehatzagoak egin nahi dira, algoritmoaren indartasunak eta ahultasunak ezagutzeko sarrerako datuei dagokionez.



6. Irudia: Splizeosoma pre-katalitikoaren flexibilitate analisiaren emaitzak

Erreferentziak

- Cressey, Daniel, eta Ewen Callaway. 2017. Cryo-electron microscopy wins chemistry nobel. *Nature* 550.167–167.
- Davis, Joseph H., Yong Zi Tan, Bridget Carragher, Clinton S. Potter, Dmitry Lyumkis, eta James R. Williamson. 2016. Modular assembly of the bacterial large ribosomal subunit. *Cell* 167.1610–1622.e15.
- Herreros, D., R. R. Lederman, J. M. Krieger, A. Jiménez-Moreno, M. Martínez, D. Myška, D. Strelak, J. Filipovic, C. O. S. Sorzano, eta J. M. Carazo. 2023. Estimating conformational landscapes from cryo-em particles by 3d zernike polynomials. *Nature Communications* 14.
- Jeon, Minkyu, Rishwanth Raghu, Miro Astore, Geoffrey Woollard, Ryan Feathers, Alkin Kaz, Sonya M. Hanson, Pilar Cossio, eta Ellen D. Zhong. 2024. Cryobench: Diverse and challenging datasets for the heterogeneity problem in cryo-em.
- Jonić, Slavica. 2016. Cryo-electron microscopy analysis of structurally heterogeneous macromolecular complexes. *Computational and Structural Biotechnology Journal* 14.385–390.
- Kimanius, Dari, Liyi Dong, Grigory Sharov, Takanori Nakane, eta Sjors H. W. Scheres. 2021. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochemical Journal* 478.4169–4185.
- Li, Hongjia, Hui Zhang, Xiaohua Wan, Zhidong Yang, Chengmin Li, Jintao Li, Renmin Han, Ping Zhu, eta Fa Zhang. 2022. Noise-transfer2clean: denoising cryo-em images based on noise modeling and transfer. *Bioinformatics* 38.2022–2029.
- Liu, Huikang, Man-Chung Yue, eta Anthony Man-Cho So. 2023. A unified approach to synchronization problems over subgroups of the orthogonal group. *Applied and Computational Harmonic Analysis* 66.320–372.
- Nogales, Eva, eta Sjors H.W. Scheres. 2015. Cryo-em: A unique tool for the visualization of macromolecular complexity. *Molecular Cell* 58.677–689.
- Plaschka, Clemens, Pei-Chun Lin, eta Kiyoshi Nagai. 2017. Structure of a pre-catalytic spliceosome. *Nature* 546.617–621.
- Sorzano, C. O. S., A. Jiménez-Moreno, D. Maluenda, M. Martínez, E. Ramírez-Aportela, J. Krieger, R. Melero, A. Cuervo, J. Conesa, J. Filipovic, P. Conesa, L. del Caño, Y. C. Fonseca, J. Jiménez-de la Morena, P. Losana, R. Sánchez-García, D. Strelak, E. Fernández-Giménez, F. P. de Isidro-Gómez, D. Herreros, J. L.

- Vilas, R. Marabini, eta J. M. Carazo. 2022. On bias, variance, overfitting, gold standard and consensus in single-particle analysis by cryo-electron microscopy. *Acta Crystallographica Section D* 78.410–423.
- Sorzano, C.O.S., R. Marabini, J. Velázquez-Muriel, J.R. Bilbao-Castro, S.H.W. Scheres, J.M. Carazo, eta A. Pascual-Montano. 2004. Xmipp: a new generation of an open-source image processing package for electron microscopy. *Journal of Structural Biology* 148.194–204.
- Vilas, Jose Luis, Jose Maria Carazo, eta Carlos Oscar S. Sorzano. 2022. Emerging themes in cryoem-single particle analysis image processing. *Chemical Reviews* 122.13915–13951. PMID: 35785962.

6 Eskerrak eta oharrak

Emaizta hauek “La Caixa” Fundazioaren (ID 100010434) laguntzari esker lortu dira. Laguntza honen kodea B006435 da. Honez gain, lan honek MICIU/AEI/10.13039/501100011033-k finantzatutako PID2022-136594NB-I00 Bekaren laguntza ere izan du.