Pohrana digitalnih podataka u DNA molekuli

Rajković, Klara

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO MATEMATIČKI FAKULTET

BIOLOŠKI ODSJEK

Seminarski rad

Pohrana digitalnih podataka u DNA molekuli

DNA digital data storage

Klara Rajković

Preddiplomski studij eksperimentalne biologije

Undergraduate study of experimental biology

Mentor: doc. dr. sc. Rosa Karlić

Zagreb, 2021.

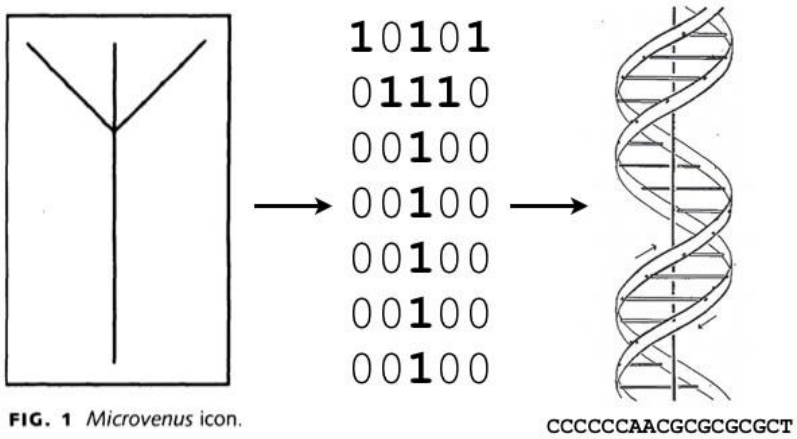
Sadržaj

1. Uvod

U današnje doba rastuće i napredne tehnologije dolazi do povećanja količine digitalnih podataka koji moraju biti fizički pohranjeni. U bližoj budućnosti, postojeći kapaciteti za pohranu podataka neće biti dovoljni zbog eksponencijalnog rasta količine podataka. Stoga znanstvenici neprestano traže bolje i učinkovitije alternativne načine za pohranu tako velikih količina podataka. DNA molekula zbog svojih bioloških svojstava predstavlja obećavajući pristup za dugoročno pohranjivanje digitalnih podataka. Na primjer, DNA nudi pohranu podataka gustoće 10^{18} bajta po mm^3, što je otprilike šest reda veličine gušće od trenutno najgušćeg dostupnog medija. (Zhrinov i sur. 2016). Točnije, svaki gram jednolančane DNA molekule može pohraniti i do 455 eksabajta informacija (Church i sur. 2012). Jedna od prednosti DNA molekule je mogućnost lančane reakcije polimeraze (*Polymerase Chain Reaction,* PCR) koja nudi kopiranje velike količine podataka u vrlo kratkom vremenu s minimalnim troškom. Isto tako, DNA molekula je sadržana u svim organizmima, dakle pristupačna je i lako dohvatljiva što pogoduje pohrani digitalnih podataka. Jedna od pozitivnih strana DNA molekule je ta što je odlično očuvana, pa je kao takva pridonijela filogenetskim istraživanjima jer se DNA mogla izvući iz fosila starih i do nekoliko tisućljeća. Osnovni proces pohrane podataka u DNA uključuje kodiranje digitalnih informacija u DNA sekvence (kodiranje), upisivanje sekvenci u molekule DNA (sinteza), fizičko uvjetovanje i organiziranje podataka u „biblioteku“ za dugotrajno skladištenje, selektivno pristupanje podacima (nasumični pristup), čitanje molekule (sekvenciranje) i pretvaranje tih podataka u digitalne podatke (dekodiranje) (Luis Ceze i sur. 2019). Drugim riječima, digitalne informacije se prvo kodiraju u ATCG sekvence pomoću razvijenih shema kodiranja. Ti nizovi se potom sintetiziraju u oligonukleotide ili dulje fragmente DNA koji omogućuju dugotrajnu pohranu podataka. Kako bi se podaci dohvatili, DNA se sekvencira da se dobije izvorna ATCG sekvenca iz sintetizirane DNA.

2. Povijest pohrane digitalnih podataka u DNA

Pohrana digitalnih podataka u DNA započela je sredinom 1960-ih kada su Norbert Wiener i Mikhail Neiman raspravljali o idejama za „genetičku memoriju“. Tek nakon 20-ak godina se pojavio prvi pokušaj demonstracije pohrane podataka u DNA kada je Joe Davis objavio „Microvenus“ (Panda i sur. 2018). Davis je kodirao 35-bitnu fotografiju u živi soj bakterija *Escherichia Coli*. Fotografija „Microvenus“ predstavlja život i Zemlju (Slika 1.). Fotografija je morala biti prevedena u binarni sustav koji će predstavljati ACGT sustav baza u DNA molekuli.



Slika 1. Proces kodiranja slike „Microvenus“ u molekulu DNA. Fotografija se najprije prevodi u binarni kod koji odgovara sekvenci u DNA molekuli (preuzeto i prilagođeno Panda i sur 2018).

Kasnije se pojavio koncept sakrivanja podataka u DNA molekuli pomoću mikrotočaka. Prikrivanje poruka (steganografija) razvio je profesor Zapp, iako se danas raspravlja čija je točno bila invencija, a to su koristili njemački špijuni u Drugom svjetskom ratu da prijenose tajne informacije (Clelland i sur. 1999). Mikrotočka je bila zapravo jako umanjena fotografija ili poruka koja bi se nalazila iznad točke koja predstavlja kraj rečenice u nekom pismu ili dokumentu. Znanstvenici su mikrotočke primijenili u DNA molekuli tako što se kodirana poruka u DNA prvo zakamuflirala unutar ogromnog ljudskog genoma, a zatim je dodatno prikrivena samim ograničenjem podatka na mikrotočku (Clelland i sur. 1999). Tajna poruka u DNA sadrži kodiranu poruku omeđenu s obje strane PCR početnicama kako bi se kasnije ta tajna poruka mogla povratiti u digitalnu informaciju.

Sva istraživanja pohrane digitalnih podataka u DNA molekuli su se do 2012. radila *in vivo*. Church i suradnici 2012. koriste sekvenciranje sljedeće generacije (*Next Generation Sequencing,* NGS) u svojim istraživanjima. Pretvorili su html kodirani nacrt knjige koji je sadržavao 53 426 riječi, 11 JPG slika i 1 *JavaScript* program u 5,27 megabitni tok bitova. Zatim su ti bitovi kodirani na 54 898 oligonukleotida, gdje svaki oligonukleotid sadrži 159 nukleotida. Svaki oligonukleotid kodirao je 96-bitni podatkovni blok, 19-bitnu adresu koja određuje lokaciju podatkovnog bloka u nizu bitova i bočne sekvence (PCR početnice) od 22 nukleotida za umnažanje i sekvenciranje (Church i sur. 2012). Prepoznavanje podataka ograničili su na 100 nukleotida kako bi se smanjila greška u sekvenciranju. Ova metoda ima mnogo prednosti u odnosu na dosadašnje pristupe pohrane podataka u DNA molekuli. Kodira se jedan bit po bazi (A ili C za nulu, G ili T za jedan), umjesto dva. To omogućuje kodiranje poruka na više načina kako bi se izbjegle sekvence koje se teško čitaju poput GC parova, ponavljanja ili sekundarne strukture. S obzirom na to da se tok bitova podijelio na blokove podataka, nema potrebe za dugim DNA konstruktima koje je teško sastaviti u tako velikoj mjeri. Isto tako, ovo je *in vitro* pristup kojim se izbjegava problem kloniranja i stabilnosti *in vivo* pristupa. Najveća prednost je korištenje NGS tehnologije koja omogućuje kodiranje i dekodiranje velikih količina informacija što smanjuje troškove za otprilike 100 000 puta u usporedbi s prvom generacijom sekvenciranja (Church i sur. 2012).

U isto vrijeme, Goldman i suradnici koriste već postojeću shemu kodiranja iz 1950-te koju je postavio Huffman. Prije kodiranja u DNA nukleotide, binarni podaci su prvo pretvoreni u trostruki Huffmanov kod, a zatom kodirani u DNA sekvence upućujući na rotirajuću tablicu kodiranja. Svaki bajt dobivenih podataka zamijenjen je s 5 ili 6 trostrukih znamenki koje sadrže samo znamenke „0“, „1“ i „2“ prema Huffmanovom algoritmu (Huffman 1952). Koristeći rotirajuću tablicu kodiranja, eliminiraju se ponavljajući mononukleotidi i izvorni podaci se mogu komprimirati za 25-37,5% (Zing i sur. 2019). Međutim, algoritam ovakvog kodiranja ne može spriječiti veliku količinu GC parova kada se koriste ovakvim binarnim obrascima.

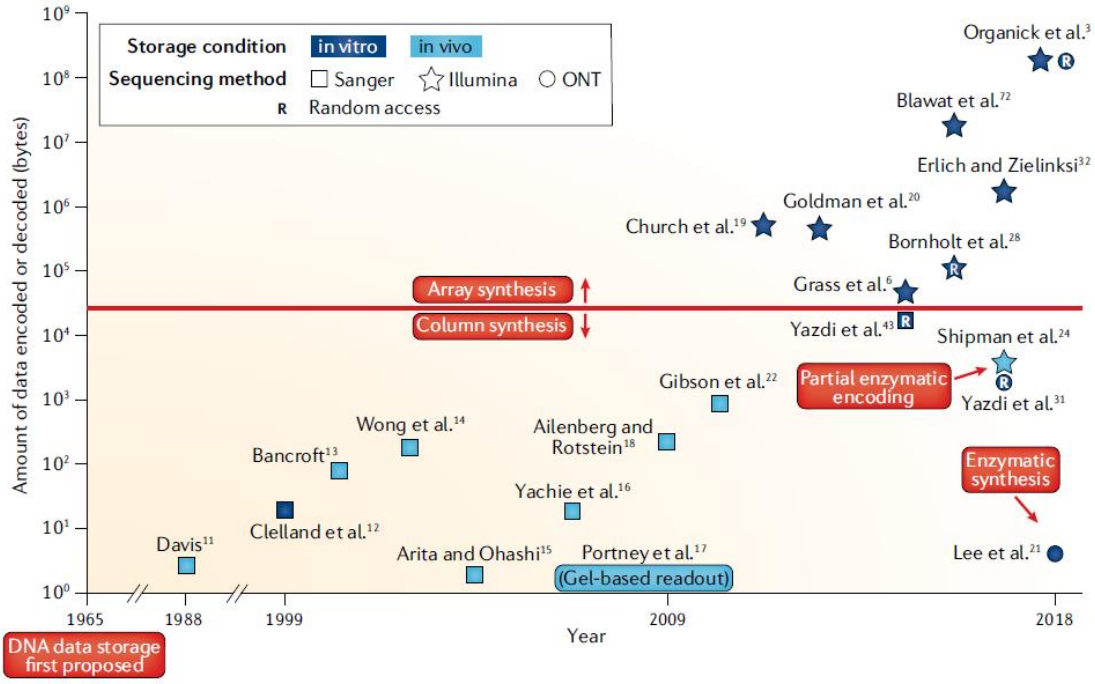
Bornholt i suradnici 2017. unaprjeđuju Goldmanovu shemu *XOR* principom. Svake dvije sekvence generiraju treću suvišnu sekvencu. Stoga se s bilo koje dvije sekvence može lako oporaviti treći slijed. Ova shema uspješno kodira četiri datoteke ukupne veličine 151 kilobajta i oporavlja tri od četiri datoteke bez manualnih intervencija (Shendure i sur. 2017). Nadalje se pojavljuje potreba za nasumičnim pristupom podacima u na temelju DNA molekule. Bornholt i suradnici 2018. stvaraju novu shemu kodiranja gdje je to omogućeno. Dodijeljene su jedinstvene PCR početnice pojedinačnim datotekama, čime se omogućuje nasumičan pristup

podacima. U njihovoj studiji uspješno je pohranjeno i oporavljeno 200 megabajta podataka, što je postavilo novu prekretnicu nadopunjujući izvedivost pohrane velikih podataka DNA (Zing i sur. 2019).

Blawat i suradnici su se usredotočili na rješavanje problema nastalih tijekom DNA sekvenciranja, amplifikacije i sinteze. Shema se temelji na dva kodiranja. Osnovni 8-bitni podatkovni blok dodijeljen je slijedu DNA od pet nukleotida gdje se treći i četvrti nukleotid mogu mijenjati. Postoje dva načina kako spriječiti ponavljanje mononukleotida, prva tri nukleotida ne bi trebala biti ista i posljednja dva nukleotida ne bi trebala biti ista. Dakle, 8-bitni podatkovni blok se kodira u 704 različita DNA bloka (Blawat i sur. 2016). Ovom shemom se izbjeglo puno mutacija, no još uvijek problem stvaraju transverzije A u T koje se mogu dogoditi.

Erilich i Zielinski 2017. koriste „fountain code“ u svojoj shemi. „Fountain code“ široko je rasprostranjena metoda kodiranja informacija u komunikacijskim sustavima i dobro je poznat po svojoj robusnosti i visokoj učinkovitosti (Byers i sur. 2002). Koriste se uobičajenom shemom 2 bita u 1 nukleotid. Izvorne binarne informacije segmentirane su na male blokove. Ovi blokovi su unaprijed dizajnirani. Onda se napravi novi podatkovni blok koji je zapravo zbroj selektiranih blokova s nasumičnim dijelovima koji su pritaknuti. Zatim se kodira u nukleotidne blokove prema shemi. Ovom metodom se mogu izbjeći ponavljajući GC sljedovi. Ova studija je povećala teoretsku granicu kodiranja do neviđeno visoke vrijednosti od 1,98 bita po nukleotidu i to je iznimno smanjilo željenu suvišnost za oporavak bez pogrešaka izvorne datoteke, također mehanizam nasumičnog odabira osigurava da dugi mononukleotidni homopolimeri se ne ponavljaju u kodiranome slijedu (Zing i sur. 2019).

Veći napredci pojavljuju se sukladno s razvojem tehnologije, pogotovo novih metoda sekvenciranja. Važno je uočiti da se prilikom otkrića NGS metoda, povećao interes za istraživanja pohrane digitalnih podataka u DNA molekuli (Slika 2.).



Slika 2. Vremenska lenta najrelevantnijih objavljenih radova temeljeni na pohrani digitalnih podataka u DNA molekuli. X-os predstavlja vremenski period, a na y-osi se nalazi količina kodiranih odnosno dekodiranih podataka u bajtovima. Vremenska lenta napravljena je tako da prikaže različite načine koji su se koristili od prvog pokušaja pohrane podataka. Svijetlo plavom bojom označeni su radovi koji su izvodili istraživanje *in vivo*, a tamno plavom bojom označeni su radovi koji su koristili *in vitro* pristup. Kvadratićem su označeni radovi koji su koristili sekvenciranje po Sangeru, zvjezdicom su označeni radovi koji su koristili *Next Generation Sequencing*, a krugom su označeni radovi koji su se koristili *Oxford Nanopore Technology MinION* metodu. Isto tako odvojeni su radovi prema načinu sinteze (preuzeto i prilagođeno Ceze i sur. 2019).

Ovisno o tome koji je cilj pohrane podataka, koriste se različite sheme. Najveću problematiku stvaraju mutacije na jednom nukleotidu te se takve greške trebaju inovativno rješavati, a za to je potrebno složeno kodiranje i dekodiranje.

3. Strategije pohranjivanja digitalnih podataka u DNA

Gledajući povijesno unatrag, vidljivo je da se u početku istraživanja pohrane podataka u DNA molekuli većinski koristio *in vivo* pristup (Slika 2.). Uglavnom su se istraživanja radila na modelnom organizmu *E.Coli*. S obzirom na dugoročno skladištenje podataka, DNA u *in vivo* stanju će se sporije razgraditi nego *in vitro.* Unaprijeđenim tehnologijama, danas se za pohranu podataka koristi i *in vitro* i *in vivo* pristup.

3.1. *In vivo*

U samim počecima koristile su se bakterije kao mediji za pohranu podataka. Danas se koriste drugačiji načini, a jedan od njih je prijenos podataka plazmidom. DNA sekvence najprije su kodirane u plazmid, a zatim prenesene u bakteriju. Kada se plazmidi nađu unutar bakterijske stanice, mogu se ugraditi u genom bakterije, a mogu se samo i zadržati unutar stanice. Problem nastaje prilikom količine podataka koje bakterija može primiti, odnosno veličina plazmida kojega bakterija može primiti. Uz to, mutacije plazmida u bakteriji su jako česte. Tijekom replikacije *E.Coli*, stopa mutacija po nukleotidu po generaciji iznosi $2{,}2 \times 10^{-10}$, odnosno $1{,}0 \times 10^{-3}$ mutacija po genomu po generaciji (Lee i sur. 2012). Budući da se bakterija dijeli svakih 20-30 minuta, iz generacije u generaciju bi bilo previše mutacija.

Drugi pristup temelji se na korištenju rekombinaza koje vežu na dva mjesta prepoznavanja i invertiraju orijentaciju tog segmenta DNA molekule, što odgovara [0,1] stanju bita. Ovom metodom su Bonnet i suradnici realizirali pohranu podataka u živim stanicama. Kasnije su Yang i suradnici izgradili jedanaest ortogonalnih memorijskih prekidača koristeći 34 integraze iz faga i konstruirali su memorijski niz u *E.Coli* koji može zabilježiti 1,375 bajta informacija (Hao i sur. 2021).

Danas najatraktivniji princip za *in vivo* pohranu podataka jest CRISPR-Cas sistem. Bitan je Cas1-Cas2 kompleks koji je zapravo integraza koja ugrađuje egzogenu DNA sekvencu u genom domaćina, takozvanu razmaknicu. Shipman i suradnici su pohranili digitalni film, veličine približno 2,6 kB, u bakterijski genom (Shipman i sur. 2017). To je prvi puta da se velika količina podataka uspjela kodirati i dekodirati *in vivo*. Postoji i strategija koja koristi Cas-9 protein CRISPR-Cas sustava. Princip je da se osmisle *self-targeting guide RNA* molekule koje urezuju DNA lokus koji kodira za te iste molekule kako bi se inducirale mutacije koje će genom domaćina popravljati (Hao i sur. 2021). Tako se omogućuje ugradnja drugih sekvenca u genom domaćina.

3.2. *In vitro*

U novijim studijama *in vitro* pohrana podataka temeljena na DNA viđa se češće nego *in vivo*. Osnovni način je korištenje „biblioteke“ oligonukleotida. Tijekom procesa sinteze, svakom oligonukleotidu se dodjeljuje kratka oznaka kako bi se razlikovali za vrijeme samog procesa, ali i za kasnije sekvenciranje. Trenutna tehnika sinteze oligonukleotida može generirati najviše 200-mere, s visokom točnosti i čistoćom (Ping i sur. 2019).

DNA nanostrukture su različite strukture izrađene od DNA, koje djeluju i kao strukturni i kao funkcionalni element, a mogu poslužiti kao kostur za stvaranje složenijih struktura (Kang i sur. 2021). Zbog velike programabilnosti, DNA nanostrukture pružaju različite načine za pohranu digitalnih podataka. Jedna od nanostruktura je ukosnica koja se može hibridizirati na određenim položajima unutar DNA molekule. Chen i suradnici su upotrijebili motiv ukosnice od 8 i od 16 nukleotida duž dvolančane DNA za kodiranje digitalnih informacija i razvili su integrirani sustav nanopora visoke razlučivosti koji ima sposobnost otkrivanja ukosnica DNA od duljine od oko 3 nm za dekodiranje digitalnih informacija (Chen i sur. 2018). Promatrajući dalje, Chen i suradnici su kombinirali mehanizam nanopora i *Toehold-mediated DNA strand replacement*. *Toehold-mediated DNA strand replacement* temelji se na neenzimatskoj zamjeni jednog lanca DNA ili RNA s drugim lancem DNA ili RNA. Ta metoda se široko koristi za mijenjanje konformacije DNA nanostruktura, a te dvije strukture (prije i poslije) se prevode u binarni sustav kao „0“ i „1“ (Hao i sur. 2021).

Zhang i suradnici su razvili pristup kodiranja zasnovan na takozvanom DNA origamiju. Programiranjem DNA origami oblika i točkastih uzoraka na origami oblicima, postigli su kodiranje tekstualnih poruka, glazbe i slikovnih datoteka u DNA origamiju. Razvili su DNA origami kriptografiju (*DNA Origami Cryptography*) koja koristi preslagivanje virusne okosnice M13 u nanometarske uzorke koji su zapravo Brailleovo pismo za sigurnu komunikaciju koja može stvoriti poruku veličine preko 700 bitova (Zhang i sur. 2019).

Osim DNA nanostruktura, za pohranu podataka koristi se i DNA koja je fluorescentno modificirana, metilirana DNA i DNA urezi modificirani enzimatskim urezivanjem. Fluorescentno označena DNA istraživana je kao element za pohranu s fluorescentnim signalom u „ON“ stanju kao binarni bit „1“ i signal u „OFF“ stanju kao binarni bit „0“ (Nguyen i sur. 2019). Kako bi se pohranilo što više podataka, nužno je koristiti različite fluorescentne boje, a to dovodi do problema pri uporabi, jer se spektri boja nekada preklapaju. Lin i suradnici razvili su barkodirana fluorescentna stanja gdje su kodirali 36 različitih boja koristeći samo 16 DNA

lanaca, tako što su različito pozicionirali fluorofore na površinu DNA molekule (Lin i sur. 2012).

Zanimljiv pristup s korištenjem ureza duž dvolančane molekule DNA koristili su Tabatabaei i suradnici koji su uveli DNA bušene kartice (*DNA punch cards*), mehanizam za pohranu makromolekula u koji se podaci zapisuju u obliku ureza na unaprijed određenim položajima na okosnici dvolančane DNA (Hao i sur. 2021).

4. Metode sinteze DNA molekula

Jedna od najučestalijih metoda sinteze DNA je oligonukleotidna sinteza bazirana na fosforamiditu (*phosphoramidite-based oligonucleotide synthesis*). Ova metoda sintetizira DNA lanac tako da se ugrađuje mononukleotid po mononukleotid. Cilj ove sinteze je ne stvarati homopolimere, a to se uspješno izbjegava jer je mononukleotid zapravo zaštićen fosfitom koji sprečava da se za njega veže drugi nukleotid. Kad se spajaju mononukleotidi, onda se treba ukloniti ta zaštitna skupina, a to se radi dodatkom kiseline. Oligonukleotidnom sintezom su se znanstvenici bavili još 1950-ih, ali je krajem 1960. Lestinger objavio rad gdje je usavršio metodu koja se i danas primjejuje u laboratorijima (Hogrefe 2014). Ova metoda je vrlo efikasna i točna, te se greške javljaju u malom postotku, a to se uglavnom događa na zadnjem dodavanju mononukleotida ili prilikom završetka sinteze.

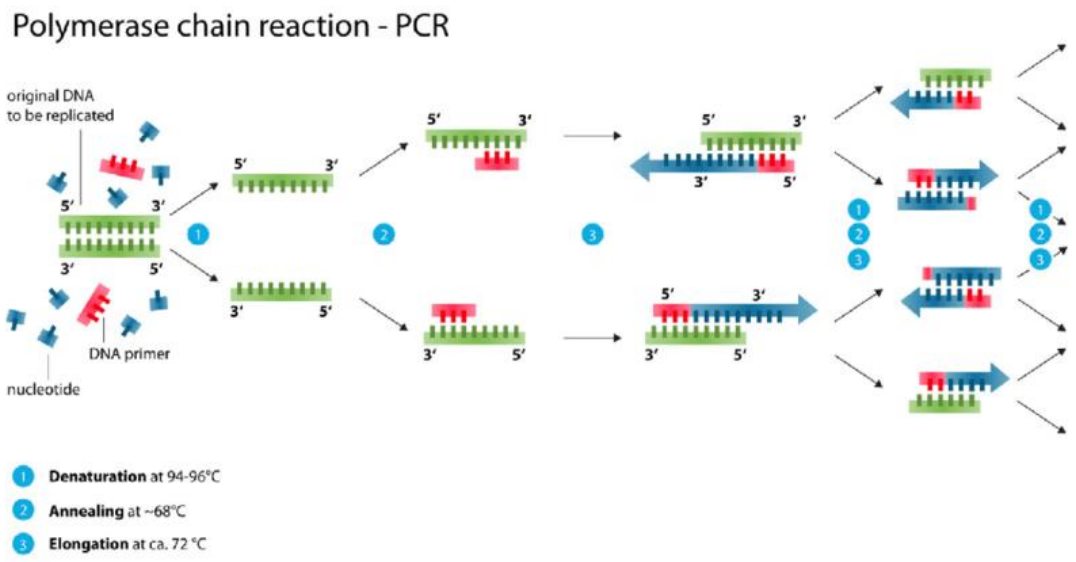
Postoji i *de novo* sinteza DNA, odnosno enzimatska sinteza DNA. U ovom procesu enzimi polimerizacije ugrađuju nukleotide bez kalupa. Veliki izazov je kontorliranje adicije mononukleotida, s obzirom na to da npr. deoksinukleotid transferaza (TdT) ima tendenciju katalizirati adiciju više nukleotida u jednom ciklusu (Ceze i sur. 2019). Ova metoda još uvijek nije komercijalno pristupačna za korištenje, no ima potencijala za jeftin i točan način sinteze DNA. Nedavno su Lee i suradnici razvili multipleks enzimatsku DNA sintezu upravljanu fotonima (*photon-directed multiplexed enzymatic DNA synthesis*) i to baš za pohranu digitalnih podataka. Tom metodom uspjeli su kodirati 12 jedinstvenih DNA oligonukleotida s glazbom iz video igrica, što je ekvivalent 110 bita podataka (Lee i sur. 2020).

Većina znanstvenika koristi oligonukleotide dugačke između 150 i 230 nukleotida za pohranu podataka (Ceze i sur. 2019). Sinteza duljih oligonukleotida predstavlja izazov budući da je važno imati što točniju sintezu bez uvođenja grešaka, a ispostavilo se da kada se pokušaju napraviti dulji lanci, da su greške češće. Stoga se dulje sekvence dobivaju slaganjem više oligonukleotida, što opet vremenski i troškovno otežava posao.

5. Metode sekvenciranja DNA molekule

Sam postupak čitanja molekule DNA započinje odabirom ciljnog oligonukleotida čiji se slijed treba analizirati. Za većinu tehnika sekvenciranja, ciljni lanac treba biti umnožen, što se može postići lančanom reakcijom polimeraze (*Polymerase chain reaction,* PCR). PCR je temperaturno osjetljiv kružni mehanizam gdje se ciljni lanac eksponencijalno umnaža, a nukleotidi se fluorescentno označuju kako bi bili vidljivi na optičkoj analizi (Mullis i sur. 1986).

Ciljni dio DNA molekule koju se želi umnožiti određuje se kratkim oligonukleotidnim sekvencama - početnicama, koji su komplementarni krajevima sekvence DNA od interesa. Ove početnice su pokretači serije reakcija pomoću enzima DNA polimeraze, koja na kalupu jednog lanca DNA sintetizira novi, komplementarni lanac, pri čemu veličina sintetiziranog dijela DNA molekule odgovara dužini koju omeđuju izabrane početnice. Na Slici 3. prikazana je shema PCR metode.



Slika 3. Shematski prikaz PCR metode. (1) Denaturacija (*Denaturation*): dvolančana DNA molekula se razdvaja na dva lanca, postiže se temperaturom od otprilike 95ºC. (2) Komplementarno vezanje početnica (*Annealing*): prilikom hlađenja naprbližno 68ºC, početnice se komplementarno vežu. (3) Sinteza komplementarnog lanca (*Extension*): povišenjem temperature na približnu 72 postiže se optimum za djelovanje Taq polimeraze koja dodaje komplementarne baze i stvara dvolančanu molekulu. (preuzeto i prilagođeno Hochstetter 2020.)

Umnožene sekvence dalje se trebaju analizirati, odnosno sekvencirati. Jedna od učestalijih metoda je Sangerovo sekvenciranje ili dideoksi metoda. Tijekom replikacije DNA molekule, u smjesu se dodaju dideoksinukleotidi (ddNTP) od kojih se svaki spaja s određenom bazom te

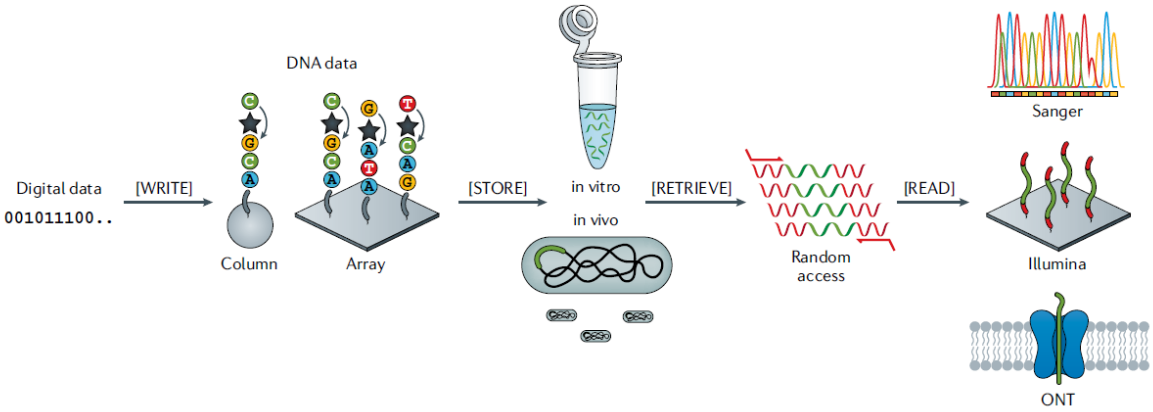
nakon njegove ugradnje, replikacija završava i dobivaju se lanci nukleotida (Berg i sur. 2002). Za svaku bazu se radi zasebna reakcija. Nakon toga se elektroforezom utvrđuju nizovi DNA različite veličine te se tako razlikuju baze u sekvenci.

Novija metoda sekvenciranja je sekvenciranje sljedeće generacije (*Next Generation Sequencing*, NGS). NGS su metode masovnog paralelnog sekvenciranja koje nude iznimo visoku točnost, brzinu i prilagodljivost. Nasprem Sangerovog sekvenciranja, NGS pokazuje točnost od ~99% (Baudhuin i sur. 2015). Jedna od NGS metoda je pirosekvenciranje, a temelji se na principu "sekvenciranja sintezom", u kojoj se sekvenciranje vrši otkrivanjem nukleotida ugrađenog u DNA polimerazu. Pirosekvenciranje se oslanja na emisiju svjetlosti na temelju lančane reakcije kada se oslobađa pirofosfat (Ansorge 2009). Metoda može ispisati do 250 pb i znatno je brža od Sangerovog sekvenciranja, iako radi s kraćim sekvencama (Schuster 2008.)

Još jedna od metoda sekvenciranja je tehnologijom nanopora. Radi na principu da se ciljna DNA pusti kroz nanometarske pore koje su pod naponom (Lin i sur. 2021). Kako DNA molekula prolazi kroz pore, događaju se fluktuacije u električnom naponu što se onda prevodi u sekvencu. Velika prednost sekvenciranja nanoporama je očitavanje podataka u stvarnome vremenu, dakle kako DNA prolazi kroz pore, odmah se očitava sekvenca. Glavni izazov u ovoj metodi je kako smanjiti učestalost pogrešaka. Neovisno o tome, ova metoda je korištena u mnogo studija, posebno uređaj *MinION*. *MinION* je mali ručni sekvencer koji je 2014. godine predstavila tvrtka *Oxford Nanopore Technologies* (ONT). Ova inovacija značajna je zbog nekoliko razloga: trošak sekvenciranja po bazi je malen u usporedbi s tehnologijama sekvenciranja druge generacije, ima sposobnost sekvenciranja dugih lanaca što je odlično za slaganje cijelih genoma, kompaktan je za nositi i ima sposobnost generiranja podataka u stvarnom vremenu (Mikheyev i Tin, 2014). Danas se točnost sekvenciranja putem *MiniION* povećava i nije više toliko velika greška.

6. Koraci u pohranjivanju podataka

Digitalna pohrana podataka u DNA molekulu sastoji se od 4 koraka prikazana na Slici 4.



Slika 4. Pregled glavnih koraka pohrane digitalnih podataka u DNA. Prvo algoritam računala prebacuje nizove bitova u sekvence (*write*). DNA sekvence su onda sintetizirane, čime se stvaraju brojne kopije svakog slijeda. Sinteza se može provoditi na koloni ili mikročipovima. Nakon toga, dobiveni DNA materijal može se klonirati i pohraniti *in vivo* ili češće *in vitro* – zamrznut u otopini ili osušen kako bise zaštitio od okoliša (*store*). DNA podaci koji se trebaju čitati se mogu selektivno uzeti iz baze DNA postupkom koji se zove nasumični pristup (*retrieve*). Nasumični pristup se postiže PCR-om. Naposlijetku radi se sekvenciranje koje može biti izvršeno različitim instrumentima (*read*). Najčešća metoda je Sangerovo sekvenciranje i Illumina, ali sve češća je uporaba *Oxford Nanopore Technologies*. (preuzeto i prilagođeno Ceze i sur. 2019.)

6.1. Kodiranje i sinteza podataka (*Write*)

Proces počinje kodiranjem. Algoritam računala prebacuje nizove bitova u sekvence DNA. Dobivene sekvence DNA se tada sintetiziraju i tako generiraju puno fizičkih kopija svakog slijeda. Sekvence DNA su proizvoljne, ali su određene duljine takve da se nizovi bitova mogu razdvojiti na manje dijelove koji se kasnije ponovo trebaju sastaviti u izvorne podatke. Da bi se ponovo sastavljanje moglo dogoditi, potrebno je ili indeksirati svaki manji dio ili pohranjivati dijelove različitih DNA sekvenca koje se preklapaju (Goldman i sur. 2013). Ako se želi postići primamljiva količina informacija, treba se sintetizirati velika količina DNA sekvenci pa je zato za sintezu bolje koristiti mikročipove što omogućuje istovremenu sintezu više sekvenci (Kosuri i Church 2014).

6.2. Pohranjivanje podataka (*Store*)

Nakon sinteze, dobivena DNA se treba pohraniti. Jedna fizički izolirana baza DNA može pohraniti 1012 bajtova (Organick i sur. 2019). Za takve podatke potrebna je „biblioteka“ kako bi se mogli proširiti sustavi skladištenja.

6.3. Dekodiranje podataka (*Retrieve*)

Nakon što je zatražen određeni podatak, potrebno je dohvatiti i uzorkovati te podatke iz baze DNA. Kako bi se izbjeglo čitanje svih podataka, potreban nam je nasumični pristup odnosno mogućnost odabira točno određenog podatka koji će se čitati iz cjelokupne baze. Nasumični pristup može se izvesti selektivnim postupcima kao što su to magnetska ekstrakcija kuglicama sa sondama koje odgovaraju određenim podacima ili PCR-om korištenjem početnica koji su komplementarni podacima koji su se tijekom kodiranja označili (Bornholt i sur. 2016).

6.4. Čitanje podataka (*Read*)

Nakon što je uzorak DNA odabran, idući korak je sekvenciranje. Očitanja koja su dobivena sekvenciranjem se dekodiraju natrag u izvorne digitalne podatke. Uspjeh ovog koraka ovisi o pokrivenosti sekvence i stope pogreške tijekom cijelog procesa (Bornholt i sur. 2016).

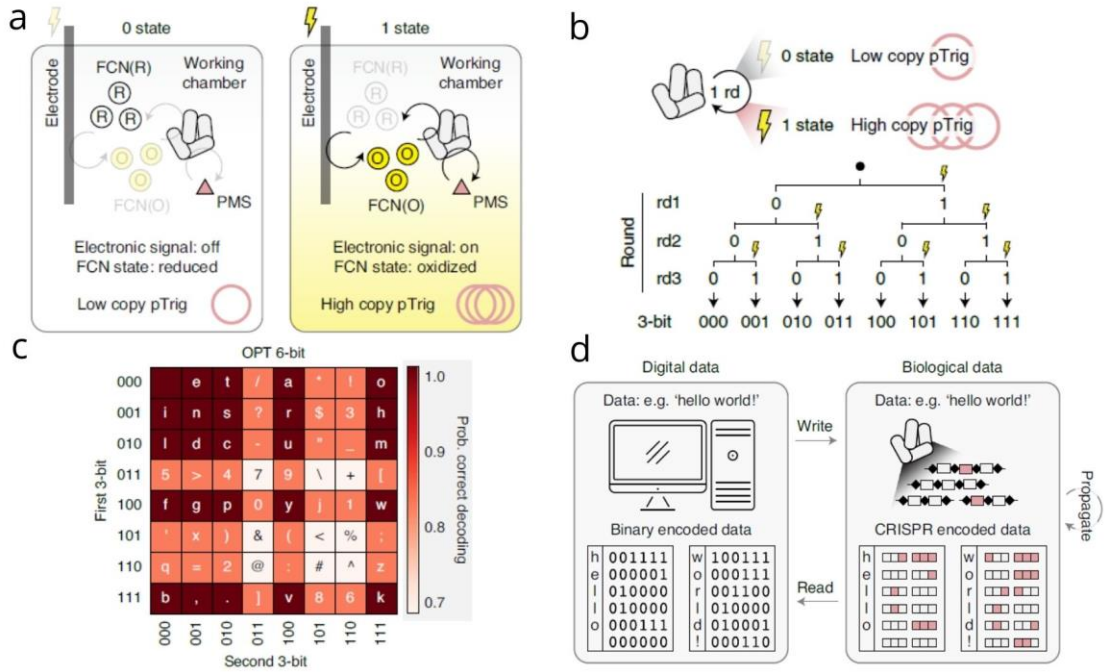
7. Primjer pohrane podataka *in vivo*

Yim i suradnici su ove godine izdali rad u kojemu su opisali novi elektrogenetski okvir za izravnu pohranu digitalnih podataka u živim stanicama upotrebljavajući metodu CRISPR. Unijeli su binarne podatke sastavljene od 3-bitnih jedinica uz pomoć CRISPR-a u stanice bakterije električnom stimulacijom. Pokazali su multipleksno kodiranje podataka u barkodirane populacije stanica kako bi dobili smislenu pohranu podataka i kapacitet do 72 bita, koji se može tijekom mnogo generacija očuvati u prirodnom okruženju (Yim i sur. 2021). Ovo je temeljni rad za budući napredak *in vivo* pohrane digitalnih podataka u DNA molekuli.

Opisali su prilagodljivu i izravnu strategiju *DRIVES* (*data recording in vivo by electrical stimulation*) za kodiranje digitalnih podataka u genome živih stanica bez potrebe za sintetiziranjem DNA *in vitro*. Korištenjem električnih signala za podešavanje redoks biomolekula i senzora stanicama, omogućuje se izravan prijenos digitalnih podataka s računala u žive stanice. Podaci pohranjeni na tim „živim tvrdim diskovima“ stabilno se održavaju kroz više generacija stanica od vanjskih okruženja u kojima bi se inače razgradila gola DNA molekula (Yim i sur. 2021).

Kako bi električni signal mogli koristiti za biološko prepoznavanje, istražili su SoxRS regulon kako bi potaknuli ekspresiju gena mijenjajući elektrokemijsko stanje stanice (uvođenje oksidativnog stresa), što utječe na broj kopija plazmida. Oksidativni stres se može izazvati s različitim dozama fenazin metosulfata (PMS), a koristili su i željezov (III) heksacijanoferat (oksidirani – FCN (O); reducirani – FCN (R)) kao zamjenski elektron akceptor. Radili su u anaerobnim uvjetima zbog lakšeg kontroliranja redoks reakcija. Konstruirali su elektrokemijski redoks kontroler s 24 komore koji može davati električne impulse zasebno svakoj komori. Kad je bio isključen to odgovara 0V, a kad je uključen to odgovara +0,5V. U Signal +0,5V oksidira FCN (R) i PMS, koji se aktivira soxS promotor za povećanje broja kopija pTrig, kako se i vidi na slici 5a, i time se broj kopija pTrig-a povećao više od 400 puta. U skladu s time, nove razmaknice koje potječu iz pTrig-a bile su 34 puta učestalije u slučaju s električnim signalom, nego bez. Radili su eksperiment kodiranja u kojem su stanice bile izložene različitim električnim signalima kroz tri uzastopna kruga, testirajući svih osam mogućih binarnih kombinacija kao što je pokazano na slici 5b. S većim elektroničkim signalom, CRISPR područje genoma se povećao tijekom eksperimenta, a prema tome i udio razmaknica iz pTriga. (Yim i sur. 2021).

Odlučili su kodirati tekstualne poruke u žive stanice. Koristili su strategiju kodiranja u kojoj se svaki „bajt“ prevodi u 6-bitni kod (za 26 ili 64 moguća znaka) izgrađen od dvije spojene 3-bitne podatkovne jedinice promatrane kroz dvije populacije stanica. Ispitali su različite sheme kodiranja kako bi optimizirali preslikavanje znakova u bajtove. Istražene su bile dvije sheme: (1) klasična DEC 6-bitna tablica kodiranja za 64 osnovna ASCII znaka i (2) optimizirana (OPT) 6-bitna tablica kodiranja, dizajnirana da uzme u obzir učestalost uporabe slova i performansu klasifikatora vidljivo na slici 5c. Da bi provjerili performanse ovih shema kodiranja, kodirali su 12-bajtnu tekstualnu poruku, „hello world!“, izravno u stanice *E.Coli*. Za svaki eksperiment kodiranja, tekst je podijeljen na 12 pojedinačnih 6-bitnih znakova, pri čemu je svaki dodijeljen dvjema populacijama stanica s barkodiranim stanicama koje sadrže po 3-bitne podatke kao što je prikazano na slici 5d. Svih 24 barkodiranih populacija privremeno je inducirano s dodijeljenim 3-bitnim signalima paralelno na elektrokemijskom redoks kontroleru. Tijekom kodiranja, profili broja kopija pTrig točno su odgovarali binarnim profilima unosa za svaku barkodiranu populaciju. Dekodiranje podataka sekvenciranja iz stanica kodiranih OPT-om uspješno je vratilo izvornu poruku „hello world!“, dok je dekodiranje podataka iz stancia kodiranih DEC-om vratilo poruku „xello world!“ zbog pogrešne klasifikacije prvih 3-bitnih ‘101’ kao ‘111’. Kako bi riješili ovaj nedostatak, slijedili su implementaciju strategije ispravljanja grešaka pomoću jednostavne provjere pariteta. S obzirom na to da je posljednji bit binarnih podataka (najnovije generiran u CRISPR nizu) uvijek najpouzdaniji za klasifikaciju, koristili smo posljednji bit od svakih 6 bita kao kontrolni zbroj za prethodnih 5 bitova. Nakon početne klasifikacije ulaza, alat za ispravljanje pogrešaka broji broj '1' u prvih pet klasificiranih bitova, a zatim očekuje '0' ili '1' za vrijednost kontrolne sume u šestom bitu na temelju brojanja. Kad vrijednost klasificirane kontrolne sume ne odgovara očekivanoj vrijednosti, klasifikator označava da je došlo do pogreške tijekom klasifikacije znaka, a pogreška se zatim ispravlja na temelju vjerojatnosti zabune klasifikatora. Nadalje su kodirali tekst 'synbio@cu' pomoću protokola za kodiranje/dekodiranje OPT u stanice i otkrili da su 2 od 54 bita u početku pogrešno klasificirana, ali su pogreške otkrivene i uspješno ispravljene kako bi se vratila ulazna poruka. Iako ispravljanje pogrešaka još uvijek nije savršeno, strategija OPT značajno smanjuje stopu pogrešaka u prosjeku na 0,79%. Zajedno, ovi rezultati pokazuju sposobnost kodiranja i pohrane značajnih količina informacija izravno u žive stanice samo pomoću električne stimulacije i pokazuju da pažljivo osmišljavanje kodiranja informacija i strategije ispravljanja grešaka mogu značajno poboljšati točnost rekonstrukcije pohranjenih podataka (Yim i sur. 2021).



Slika 5. Izravno pohranjivnje digitalnih podataka u žive stanice CRISPR metodom. (a) U stanju 0 ne primijenjuje se električni signal (0,0 V) kako bi bilo što manje FCN (R) i PMS, a prema tome i kako bi broj kopijs pTrig-a bio nizak. U stanju 1 primijenjuje se električni signal (0,5 V) koji oksidira FCN (R) i PMS, aktivirajući soxS promotor koji povećava broj kopija pTrig. FCN (R) – ferocijanid; FCN (0) – fericijanid; PMS – fenazin metosulfat. (b) Stanice su bile izložene električnim signalima tijekom tri uzastopna kruga, što je testiralo svih osam mogućnosti 3-bitnih binarnih profila podataka. (c) Optimizirana (OPT) 6-bitna tablica znakova koja uzima u obzir učestalost upotrebe slova i performans klasifikatora. 6-bitni binarni podaci za svaki znak podijeljeni su u dvije barkodirane populacije stanica. (d) Prikazan je cjelokupni proces pohrane i dohvaćanja podataka. Digitalne informacije mogu se izravno kodirati u CRISPR područje u genomima bakterijske populacije pomoću električnih signala. Populacija stanica tada se može arhivirati za dugotrajno skladištenje, propagirati za pojačavanje podataka i sekvencirati za dohvaćanje podataka. Primjer je kodirana poruka „hello world!“ koja je uspješno ukodirana i kasnije dohvaćena. (preuzeto i prilagođeno, Yim i sur. 2021)

8. Primjer pohrane podataka *in vitro*

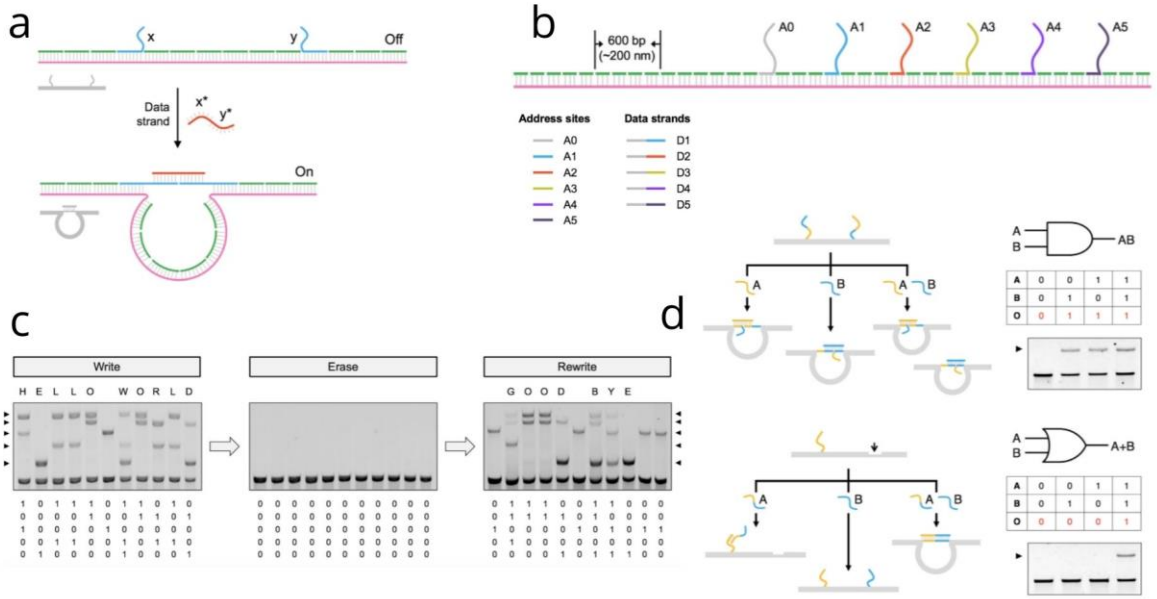
Chandrasekaran i suradnici su osmislili sistem kodiranja podataka u različite konformacije nanoprekidača nakon čega se podaci dekodiraju gel-elektroforezom. Sistem je 5-bitni sustav koji može pisati, brisati i prepisivati binarne prikaze alfanumeričkih simbola i kompatibilan je s logičkim operacijama „OR“ i „AND“. Ovakav pristup je jednostavan i brz za izvođenje, pa se zamišlja da će pomoći pri sigurnoj razmjeni informacija i pomoći u napretku prethodno razvijenih strategija šifriranja (Chandrasekaran i sur. 2017).

Nanoprekidači imaju petlje koje se mogu prostorno programirati tako da se u DNA ugrađuju privjesci na različitim mjestima. U prijašnjim radovima su se nano prekidači koristili za proučavanje interakcija između molekula, a ovdje nanoprekidači kodiraju jedan bit informacija ovisno o prisutnosti ili odsutnosti petlje (1, odnosno 0) koja se detektira gel-elektroforezom. Petlja će nastati kada se jednolančani privjesak hibridizira s vanjskim lancem kako je to prikazano na slici 6a. Vanjski lanac u ovom radu naziva se podatkovnim lancem, a privjesci su adrese. Nanoprekidač koji su napravili sastoji se od jednolančane okosnice (virusni genom M13) i kratkih komplementarnih oligonukleotida. Među tim komplementarnim oligonukleotidima okosnice, 12 lanaca je ravnomjerno odvojeno na okosnici i mogu se zamijeniti oligonukleotidima na adresnom mjestu. Komplementarne oligonukleotide ukosnice nazivaju se poštanskim mjestima, a njihovom različitosti postigli su različite veličine petlji. Oligonukleotidi na adresnom mjestu sadrže jednolančane privjeske koji se mogu vezati na određene podatkovne lance. Ovaj dizajn nanoprekidača ima šest adresa (razmaknutih ~600 nukleotida međusobno) koje mogu tvoriti pet jedinstvenih petlji pomoću pet različitih podatkovnih lanaca. Prvo mjesto adrese uobičajeno je među svih pet petlji tako da bilo koja molekula nanoprekidača može tvoriti samo jednu petlju koja predstavlja jedan bit. Kao rezultat toga, obuhvaća se 5-bitni memorijski sustav u kojem je svaki bit neovisno adresiran i razlučen na agaroznom gelu vidljivo na slici 6b (Chandrasekaran i sur. 2017).

Ovaj sustav omogućuje i pisanje i brisanje podataka u nekoliko ciklusa. Različiti bitovi imaju različite brzine pisanja koje ovise o veličine petlje i o vezanju. Broj mogućih ciklusa zapravo ovisi o koncentraciji podatkovnih lanaca, odnosno više ciklusa zahtjeva veće koncentracije. Ovdje je gornja granica u rasponu od 10 do 100 ciklusa. Tako su pokazali višebitnu obradu podataka kodiranjem riječi „Hello World“, zatim brisanje tih riječi i zatim pisanje „Good Bye“ na istom sustavu vidljivo na slici 6c (Chandrasekaran i sur. 2017).

Logičke odgovore ovog sustava testirali su na način da s projektirali jednu petlju koju mogu aktivirati dva ulazna lanca vidljivo na slici 6d. Za operaciju „OR“, programirali su dva adresna mjesta od kojih svako ima regije komplementarne dijelu ulaznog lanca A i regiju komplementarnu dijelu ulaznog lanca B. Dodatak bilo kojeg lanca rezultirao je u stvaranju petlje hibridizacijom ulaznog lanca A ili B na adresama. Za operaciju „AND“, sustav su programirali da sadrži samo jedno mjesto adrese koje je komplementarno dijelu ulaznog lanca A. Jedna polovica ulaznog lanca B dizajnirana je da se veže izravno na okosnicu, a druga polovica do ulaznog lanca A. Petlja se aktivira samo u prisutnosti oba ulaza, koja se drže zajedno hibridizacijom komplementarnih regija na A i B (Chandrasekaran i sur. 2017).

Iako je ova razina pohrane podataka mala u usporedbi s arhivskim DNA sustavima, mogla bi se pokazati korisnom za određene primjene, poput označavanja proizvoda. Pohranjene informacije mogu se oporaviti nakon sušenja kao što je to prikazano na slici 6c, što upućuje na to da bi proizvodi mogli biti diskretno označeni podacima poput datuma isteka ili proizvodnje, univerzalnim kodom proizvoda ili drugim identifikacijskim podacima. Takve se informacije mogu, na primjer, ugraditi u papir ili integrirati u pojedinačne tablete lijekova. Očitavanje gela prilično je jednostavno i jeftino u usporedbi s drugim metodama očitanja koje znaju biti korištene. Pokazali su razlučivost petlje nanoprekidača u samo 10 minuta, a takvo se očitanje moglo provesti izvan laboratorija pomoću trenutno dostupnih gelova bez pufera. Jedna prednost očitanja s jednom molekulom je mogućnost čitanja više bitova po molekuli, a ne jednog bita po molekuli koju smo koristili. Ovaj memorijski sustav također je kompatibilan sa shemom šifriranja. U toj shemi šifriranja, podatkovni lanci djeluju kao ključ za dešifriranje pripremljenih nanoprekidača. Bez pristupa fizičkoj smjesi koja sadrži ključ za dešifriranje, podaci se ne mogu dohvatiti. Budući da se nizovi podataka mogu proizvoljno dizajnirati i držati u tajnosti, šifrirani podaci ostaju sigurni čak i kad se javno distribuiraju (Chandrasekaran i sur. 2017).



Slika 6. Memorijski sustav temeljen na DNA nanoprekidaču. (a) DNA nanoprekidač dugi je dupleks s jednolančanim nastavcima koji su komplementarni s vanjskim DNA lanacem. Hibridizacija podatkovnog lanca s tim jednolančanim nastavcima rezultira stvaranjem petlje. (b) Dizajn 5-bitnog memorijskog sustava s više adresa (A0–A5) koje su djelomično komplementarne nizovima podataka (D1–D5). Vezanje različitih nizova podataka na adresama dati će petlje različitih veličina. (c) Višebitna obradu podataka kodiranjem riječi „Hello World“, zatim brisanje tih riječi i zatim pisanje „Good Bye“ na isto mjesto. (d) Operacija „OR“, tablica istine i rezultati nalazi se na gornjoj slici. Nanoprekidač sadrži dvije adrese od kojih svaka ima regije komplementarne dijelu ulaznog lanca A (narančasta) i dijelu ulaznog lanca B (svijetlo plava). Dodavanje bilo kojeg unosa uzrokuje petlju. Operacija „AND“, tablica istinitosti i rezultati su na slici dole. Nanoprekidač sadrži jedno adresno mjesto (narančasto) koje je komplementarno dijelu ulaznog lanca A. Jedna polovica ulaznog lanca B (svijetlo plava) veže se na jednolančano područje na okosnici (označeno strelicom), a druga polovica za ulaz u lanac A. Petlja se aktivira samo u prisutnosti oba ulaza. (preuzeto i prilagođeno Chandrasekaran i sur. 2017)

9. Zaključak

Do danas su dosegnuta velika postignuća korištenjem DNA kao medija za pohranu podataka i *in vitro* i *in vivo*. No, ova je metoda još daleko od praktične primjene i postoje izazovi za njen daljnji razvoj. Iako se veličina pohranjenih podataka kroz godine povećava, najveća prijavljena veličina se još uvijek ne može usporediti s medijima za pohranu podataka poput tvrdih diskova, zbog čega je potrebno razviti nove sheme kodiranja. Osim toga, procesi pisanja i čitanja relativno su složeni za postojeće pristupe, pa jedan ciklus pohrane traje dugo i može koštati. Uspoređujući pohranu podataka u DNA *in vitro* i *in vivo*, trenutno je praktičniji *in vitro* način skladištenja s obzirom na cijenu, stabilnost i povjerljivost. Bez obzira na to, prednosti *in vivo* pristupa se još moraju istražiti. Kako se područje sintetske biologije sve više razvija, *in vivo* skladištenje podataka bi moglo dati odgovore na trajne nedostatke *in vitro* metoda skladištenja. DNA pohrana podataka obećava alternativni format diska ili sličnih medija, koji trenutno dostižu maksimum svog kapaciteta skladištenja podataka. Budući da je DNA konzerviranija i vremenski vječna, odličan je izbor za dugoročnu pohranu podataka. Osim toga, DNA steganografija i kriptografija još su jedan značajan aspekt za proučavanje. Svakako će izgradnja sučelja između *in vitro* i *in vivo* sustava proširiti primjenu pohrane informacija u DNA i poboljšati inteligenciju sintetičkih bioloških sustava.

10. Literatura

Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. *New biotechnology*, *25*(4), pp.195-203.

Baudhuin, L.M., Lagerstedt, S.A., Klee, E.W., Fadra, N., Oglesbee, D. and Ferber, M.J., 2015. Confirming variants in next-generation sequencing panel testing by Sanger sequencing. *The Journal of Molecular Diagnostics*, *17*(4), pp.456-461.

Berg, J.M., Tymoczko, J.L. and Stryer, L., 2002. Biochemistry.

Blawat, M., Gaedke, K., Huetter, I., Chen, X.M., Turczyk, B., Inverso, S., Pruitt, B.W. and Church, G.M., 2016. Forward error correction for DNA data storage. *Procedia Computer Science*, *80*, pp.1011-1022.

Bornholt, J., Lopez, R., Carmean, D.M., Ceze, L., Seelig, G. and Strauss, K., 2016, March. A DNA-based archival storage system. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems* (pp. 637-649).

Byers, J.W., Luby, M. and Mitzenmacher, M., 2002. A digital fountain approach to asynchronous reliable multicast. *IEEE Journal on Selected areas in Communications*, *20*(8), pp.1528-1540.

Ceze, L., Nivala, J. and Strauss, K., 2019. Molecular digital data storage using DNA. *Nature Reviews Genetics*, *20*(8), pp.456-466.

Chen, K., Kong, J., Zhu, J., Ermann, N., Predki, P. and Keyser, U.F., 2018. Digital data storage using DNA nanostructures and solid-state nanopores. *Nano letters*, *19*(2), pp.1210-1215.

Chandrasekaran, A. R., Levchenko, O., Patel, D. S., MacIsaac, M., & Halvorsen, K. (2017). Addressable configurations of DNA nanostructures for rewritable memory. *Nucleic acids research*, *45*(19), pp. 11459–11465.

Church, G.M., Gao, Y. and Kosuri, S., 2012. Next-generation digital information storage in DNA. *Science*, *337*(6102), pp.1628-1628.

Clelland, C.T., Risca, V. and Bancroft, C., 1999. Hiding messages in DNA microdots. *Nature*, *399*(6736), pp.533-534.

Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B. and Birney, E., 2013. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, *494*(7435), pp.77-80.

Hao, Y., Li, Q., Fan, C. and Wang, F., 2021. Data Storage Based on DNA. *Small Structures*, *2*(2), p.2000046.

Hochstetter, A., 2020. Lab-on-a-Chip Technologies for the Single Cell Level: Separation, Analysis, and Diagnostics. *Micromachines*, *11*(5), p.468.

Hogrefe, R. and BioTechnologies, T., 2014. A short history of oligonucleotide synthesis. *Trilink Biotechnologies*, p.6.

Hong, F., Zhang, F., Liu, Y. and Yan, H., 2017. DNA origami: scaffolds for creating higher order structures. *Chemical reviews*, *117*(20), pp.12584-12640.

Huffman, D.A., 1952. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, *40*(9), pp.1098-1101.

Kang, H., Lin, T., Xu, X., Jia, Q.S., Lakerveld, R. and Wei, B., 2021. DNA dynamics and computation based on toehold-free strand displacement. *Nature Communications*, *12*(1), pp.1-9.

Kosuri, S. and Church, G.M., 2014. Large-scale de novo DNA synthesis: technologies and applications. *Nature methods*, *11*(5), pp.499-507.

Lee, H., Popodi, E., Tang, H. and Foster, P.L., 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences*, *109*(41), pp.E2774-E2783.

Lee, H., Wiegand, D.J., Griswold, K., Punthambaker, S., Chun, H., Kohman, R.E. and Church, G.M., 2020. Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage. *Nature communications*, *11*(1), pp.1-9.

Lin, B., Hui, J. and Mao, H., 2021. Nanopore Technology and Its Applications in Gene Sequencing. *Biosensors*, *11*(7), p.214.

Lin, C., Jungmann, R., Leifer, A.M., Li, C., Levner, D., Church, G.M., Shih, W.M. and Yin, P., 2012. Submicrometre geometrically encoded fluorescent barcodes self-assembled from DNA. *Nature chemistry*, *4*(10), pp.832-839.

Ma, S., Tang, N. and Tian, J., 2012. DNA synthesis, assembly and applications in synthetic biology. *Current opinion in chemical biology*, *16*(3-4), pp.260-267.

Mikheyev, A.S. and Tin, M.M., 2014. A first look at the Oxford Nanopore MinION sequencer. *Molecular ecology resources*, *14*(6), pp.1097-1102.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H., 1986, January. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 51, pp. 263-273). Cold Spring Harbor Laboratory Press.

Nguyen, H.H., Park, J., Hwang, S., Kwon, O.S., Lee, C.S., Shin, Y.B., Ha, T.H. and Kim, M., 2018. On-chip fluorescence switching system for constructing a rewritable random access data storage device. *Scientific reports*, *8*(1), pp.1-11.

Organick, L., Ang, S.D., Chen, Y.J., Lopez, R., Yekhanin, S., Makarychev, K., Racz, M.Z., Kamath, G., Gopalan, P., Nguyen, B. and Takahashi, C.N., 2018. Random access in large-scale DNA data storage. *Nature biotechnology*, *36*(3), pp.242-248.

Panda, D., Molla, K.A., Baig, M.J., Swain, A., Behera, D. and Dash, M., 2018. DNA as a digital information storage device: hope or hype?. *3 Biotech*, *8*(5), pp.1-9.

Ping, Z., Ma, D., Huang, X., Chen, S., Liu, L., Guo, F., Zhu, S.J. and Shen, Y., 2019. Carbon-based archiving: current progress and future prospects of DNA-based data storage. *GigaScience*, *8*(6), p.giz075.

Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. *Nature methods*, *5*(1), pp.16-18.

Shipman, S.L., Nivala, J., Macklis, J.D. and Church, G.M., 2017. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature*, *547*(7663), pp.345-349.

Zhirnov, V., Zadegan, R.M., Sandhu, G.S., Church, G.M. and Hughes, W.L., 2016. Nucleic acid memory. *Nature materials*, *15*(4), pp.366-370.

Yim, S.S., McBee, R.M., Song, A.M., Huang, Y., Sheth, R.U. and Wang, H.H., 2021. Robust direct digital-to-biological data storage in living cells. *Nature Chemical Biology*, *17*(3), pp.246-253.

11. Sažetak

Posljednjih godina suočeni smo s eksponencijalnim rastom količine podataka što uzrokuje potražnju novih medija za pohranu istih. DNA molekula pruža alternativni način pohrane digitalnih podataka zbog svoje gustoće, trajnosti i kompaktnosti. Kroz prošlo desetljeće postigao se veliki napredak iskorištavanjem umjetno dizajniranih DNA komponenata za pohranu podataka. Ovdje se objašnjavaju postupci pohrane podataka koji se sačinjavaju od kodiranja, pisanja, pohrane, dohvatanja i čitanja te dekodiranja. Osim klasičnog kodiranja sekvenca nukleinskih kiselina, spominje se strategija pohrane podataka pomoću DNA nanostruktura. Osim toga, napretkom tehnologije omogućena je i *in vivo* pohrana podataka upotrebljavajući CRISPR-Cas sustav. Prikazani su izazovi i mogućnosti za razvoj i primjenu pohrane podataka temeljenih na DNA molekuli.

12. Summary

We've recently encountered an exponential surge in the number of data that requires the acquisition of new storage media. Because of its density, durability, and compactness, the DNA molecule is a viable alternative for storing digital data. Using artificially engineered DNA materials to store data has made significant progress over the last decade.The procedures for storing data made up of encoding, writing, storage, retrieving, reading, and decoding are explained here. Besides the common coding od nucleic acid sequences, a strategy for dana storage is mentioned using DNA nanostructures. In addition, the advancement of technology also enabled *in vivo* dana storage using the CRISPR-Cas system. The challenges and opportunities for developing and using data storage based on DNA molecules are discussed.