Mange fluer i én smekk

– DNA-meta-stekkoding identifiserer mange organismer samtidig

ENDRE WILLASSEN

En kolossal utvikling av bio- og informasjonsteknologi med såkalt «High Throughput Sequencing» er i ferd med å sette tydelige spor i forskningen på biologisk mangfold. Mulighetene for å identifisere organismer med DNA-sekvenser blir blant annet forsøkt utnyttet med metodikk som går under navnet metastrekkoding. Her gir jeg et lite innblikk i teknologien og viser med eksempler noen av de metodiske aspektene som definerer muligheter og begrensninger i metastrekkoding.

DNA-teknologi revolusjonerer biologien

Oppdagelsen av strukturen til DNA førte raskt til utvikling av en mengde instrumenter og kjemikalier som benyttes til studier av gener og genomer. Et særlig betydningsfullt metodisk gjennombrudd var *polymerase-kjedereaksjonen* (PCR) (Figur 1), som Gary Mullis fikk Nobelpris for, men som noen biokjemikere mener den senere bergensprofessoren Kjell Kleppe var den egentlige oppfinneren av. Etter covid-pandemien er det ikke lenger mange i vårt land som ikke har hørt om PCR. Med denne teknologien kan noen få picagram DNA, altså milliarddeler av et gram eller mer, kopieres og oppkonsentreres til de mengder som er nødvendig for videre analyser av forskjellig slag. Kombinasjonen av PCR og såkalt *Sanger-sekvensering* gjorde, i løpet av et par tiår, DNA-sekvens-data til «konsumvare» i en rekke biologiske forskningsdisipliner. DNA-strekkoding ble en internasjonal dugnad for biodiversitetskunnskap like etter årtusenskiftet.

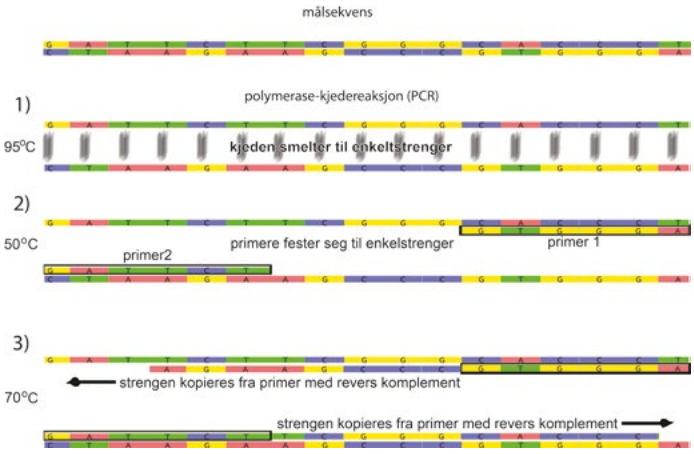


Fig. 1 | En PCR-syklus består av tre trinn: 1) oppvarming som smelter dobbelstrengen, 2) avkjøling med feste av primere, og 3) forlengelse av primerne ved at en polymerase bygger en komplementær kjede. Etter én syklus vil vi derfor ha $2^2 = 4$ kopier av sekvensen mellom de to primer-stedene. I en serie slike sykluser mangfoldiggjøres DNA eksponentielt. En PCR-reaksjon med 40 sykluser kan dermed produsere $2^{41} = 2\ 199\ 023\ 255\ 552$ kopier av molekylet. I dette enkle eksemplet er det sekvensen CCCGA og reverskomplement TCGGG som amplifiseres mellom to primerne.

Strekkoding og metastrekkoding

Den underliggende idéen ved DNA-strekkoding er at en organisme skal kunne bestemmes til art, dersom den først blir registrert med en unik DNA-sekvens i en database. Slik kan DNA-sekvensen fungere som et identitetsmerke, på tilsvarende vis som de strekkodene vi skanner når vi betaler varer i butikken. Å knytte slike strekkoder spesielt til flercellede livsformer var hovedmålet ved etableringen av samarbeidet i IBOL, International Barcode of Life. Siden IBOL, og det norske NORBOL, ble etablert i 2003, har teknologien for DNA-analyser gjennomgått store forandringer, og en av de

tydeligste effektene av såkalt *nestegenerasjons-sekvensering* (NGS) er de enorme kapasitetsøkningene for framstilling av sekvenser. Det kalles HTS (high Throughput Sequencing) i faglig sjargong fordi millioner av sekvenser kan produseres parallelt i fysisk adskilte «flytceller». Slike flytceller finnes på spesielle brikker («chips») med tusenvis av mikroskopiske brønner der reaksjonene mellom DNA og kjemikalier finner sted. Teknologien åpner for *metastrekkoding* («*metabarcoding*»), der en kan framskaffe DNA-sekvenser fra mange organismer samtidig, og dermed få et bilde av sammensetningen av ulike livsformer i en gitt miljøprøve. I relativ småskala eksperimenter kan en for eksempel analysere dietten til en dyreart fra mageinnhold og bidra til å utrede næringsnettverk i et økosystem. Men mulighetene for storskala undersøkelser av artsmangfold i prøver fra luft, jord eller vann har ført til sterk oppblomstring av anvendelser som også befatter seg med større økosystemer. Metastrekkoding brukes nå i økende grad til undersøkelser av biomangfold i ulike livsmiljø. Metodikken promoteres også som et nytt redskap for å vurdere miljøtilstanden i et økosystem og for å overvåke endringer som måtte skyldes uønsket påvirkning.

Sekvenseringsteknikker

I Sanger-sekvensering bruker en PCR i en prosess der polymerasen kan erstatte et komplementært *dNTP* med et komplementært, såkalt *ddNTP*. Hvis det skjer, vil kjeden slutte å vokse, fordi ddNTP ikke kan binde seg til et nytt nukleotid nedstrøms. Etter en slik sekvenseringsreaksjon med PCR vil derfor prøven bestå av amplifiserte fragmenter med ulike lengder. De korteste består av primer og én ddNTP. De lengste vil dekke hele sekvenslengden med tilknyttede dNTP og ha en «lysbærende» ddNTP i enden av kjeden. Ved elektroforese i et kapillærrør vil de ulike sekvenslengdene bli sortert slik at de korteste først vil passere en laserdetektor i sekvenseringsmaskinen, som registrerer nukleotidtypen i enden av kjeden, og dermed registrerer rekkefølgen av nukleotider i sekvensen. I hovedsak er det Sanger-sekvensering som er brukt for å lage de mange sekvensene fra ulike organismer som i dag finnes offentlig tilgjengelig i ulike gen-databaser, og som brukes som referansesekvenser for å identifisere ukjente sekvenser.

De nye teknologiene, som omtales som nestegenerasjons-sekvensering (NGS) baserer seg på andre, og litt ulike tekniske løsninger (Figur 2). Noen av disse egner seg bedre for lange DNA-sekvenser, andre for relativt korte. Det er også store forskjeller i priser for bruk og anskaffelse av instrumenter. For noen av sekvenserings-plattformene må en forvente

Teknologi	amplifisering	sekvensering	enzymer	signal	sekvenserte molekyler i én operasjon	lengder bp	Tilgjengelig år
Sanger	PCR i løsning	elektroforese av dideoksy-kjeder	med polymerase	optisk	-	400-900	1977
Roche 454	PCR på kuler	enkelt nukleotid	med polymerase	optisk	70 tusen-1 million	300-400	2007
SOLID	PCR på kuler	syklisk ligering	med ligase og endonuklease	optisk		75	2007
Ion Torrent	PCR på kuler	enkelt nukleotid	med polymerase	elektrisk H^+(pH)	3-80 millioner	200-400	2011
Illumina HiSeq	Bro-PCR på fast flate	syklisk reversible stopp	med polymerase	optisk	1.2-1.5 milliarder	75-150	2014
Oxford Nanopore	_	nanopore	-	elektrisk		72-260 tusen	2014
PacBio	_	enkelt molekyl sanntid SMRT	med polymerase	optisk	23-62 tusen	15-20 tusen	2011
Sequel	_	enkelt molekyl sanntid SMRT	med polymerase	optisk	4 millioner	30-100 tusen	2018

Fig. 2 | Noen sekvenseringsplattformer (som også kan forkomme i flere utgaver).

dager i produksjonstid. Andre leverer resultater etter 2–3 timer.

De mest brukte NGS-metodene sekvenserer relativt korte segmenter og krever et par steg med forbehandling av DNA-prøvene før de kan sekvenseres. Dette omtales gjerne som å lage *bibliotek*. Teknikkene som sekvenserer korte fragmenter, kan kreve at lange *templater* på forhånd kuttes til kortere biter, rundt 200–1000 basepar (bp). Det kan gjøres mekanisk, med ultralyd, eller kjemisk med *endonukleaser*. Deretter reparerer en endene på de kuttede fragmentene, slik at de to DNA-trådene blir like lange og tilføyer såkalte *adaptersekvenser*, enten med ligase eller ved PCR og såkalte *fusjonsprimere*. Hensikten med disse er at de skal feste DNA-templatet til komplementære forankringspunkter i flytcellen. Disse punktene er som korte staver på bittesmå kuler (Roche 454 og IonTorrent) eller på en fast overflate (Illumina). Disse er setet for PCR og sekvensering med disse maskinene.

Dersom en først lager *amplikoner* med tradisjonell PCR, er fragmentering av DNA ikke nødvendig. Da kan avstanden mellom primerne være tilstrekkelig kort til at vi får fragmenter som sekvenseringsteknologien kan håndtere. Det er disse amplikonene som skal brukes til artsidentifikasjon. Festepunktene for adaptersekvensene befinner seg, sammen med kjemikalier, i den såkalte *flytcellen*, der reaksjonene

finner sted. I tillegg til adaptere kan en også tilknytte *indekssekvenser* til templatet. Disse kalles noen ganger «strekkoder», til forveksling med de strekkoder vi sekvenserer for å identifisere arter. Indekssekvenser er korte, men unike sekvenser som gjør det mulig å blande prøver fra ulike kilder inn i samme sekvenseringsreaksjon. Slik prosedyre med blandede prøver kalles *multipleksing*. Senere kan en, med databehandling, sortere de digitale sekvensene ut til deres respektive prøver, basert på indeksmerkene. Det kalles *demultipleksing*. På denne måten kan en utnytte den enorme prosesseringskapasiteten som fines i NGS-teknologi.
De fleste sekvenserings-plattformer bruker polymeraser og sekvenserer ved DNA-syntese. I SOLID-systemet benytter man derimot endonuklease og ligase til vekselvis kutting og sammenføying av nukleotider. Ulike prinsipper benyttes også for å registrere at et nukleotid binder seg til et annet. Metodene i bruk på Roche454 og IonTorrent har det til felles at de registrerer et signal når ett enkelt komplementært nukleotid aksepteres i en DNA-kjede som vokser ved syntese. Det krever at bare én av de fire dNTP-typene kan tilsettes reaksjonen av gangen. Med IonTorrent registreres bindingen av et nukleotid til templatet ved et fall i pH, fordi reaksjonen frigjør et H^+-ion. Dersom sekvensen har flere like baser etter hverandre, har denne teknikken en mulig feilkilde, at antallet like baser telles feil. Ved 454-sekvensering utløses en kjedereaksjon av enzymer når nukleotider bygges inn i kjeden. Her inngår også *luciferase*, stoffet som gir lysglimt hos «ildfluer» og mange selvlysende organismer. På grunn av de lysproduserende reaksjonene kalles dette pyrosekvensering.

Den mest anvendte teknologien nå er Illumina-sekvensering. Her amplifiseres templater med såkalt bro-PCR, og under sekvenseringen vil fluoriserende dNTP gi lys med bølgelengde og intensitet som signaliserer nukleotidtype. Illumina produserer svært korte sekvenser, og derfor må de skjøtes sammen digitalt til større lengder etterpå. Potensialet for feilkoblinger av slike sekvenser er størst når det ikke finnes en allerede eksisterende, lang sekvens som kan veilede sammenstillingen. En kan fordoble indekssekvensene slik at en med større sikkerhet kan sammenstille korte sekvenser til en lengre *konsensussekvens.* Dette er en strategi som kan passe til å lage strekkoder, men den er ikke er ideell for metastrekkoding.

Nanopore-sekvensering er basert på nanoteknologi, og er nokså forskjellig fra teknikkene omtalt ovenfor. En nanopore er et hull i en silika-membran. Det er rundt en milliontedels millimeter i diameter. Rundt åpningen ligger metall (jern)

som kan lede strøm. Gjennom dette hullet kan vi tre en enkelstreng DNA, omtrent som om vi trer en sytråd gjennom øyet på en nål. Transporten av DNA gjennom poren drives av elektrisk spenning, ettersom DNA også er elektrisk ladet. Spenningsendringer når ulike nukleotider passerer nanoporen, blir registrert og konvertert til sekvensavlesninger. Nanopor-sekvensering kan produsere lange sekvenser, og det finnes nå maskiner i lommeformat (MinION) som kan kobles til en laptop og sekvensere opptil 420 baser i sekundet. PacBio og Sequel har mange fortrinn fordi de produserer lange sekvenser med stor nøyaktighet ved såkalt SMRT-sekvensering (Single Molecule Real Time). Teknologien er viktig for genomforskning og klinisk grunnforskning, og har også blitt brukt i metastrekkoding.

Arbeidsflyt for metastrekkoding

Prøvetaking for metastrekkoding vil være ulikt utformet alt etter hva hensikten med undersøkelsene er. Prøvene kan være filtrert vann, jordprøver, bunnsedimenter, avskrap fra et fast underlag, feller for små dyr osv. (Figur 3). For å frigjøre DNA fra resten av materialet i prøven brukes ulike

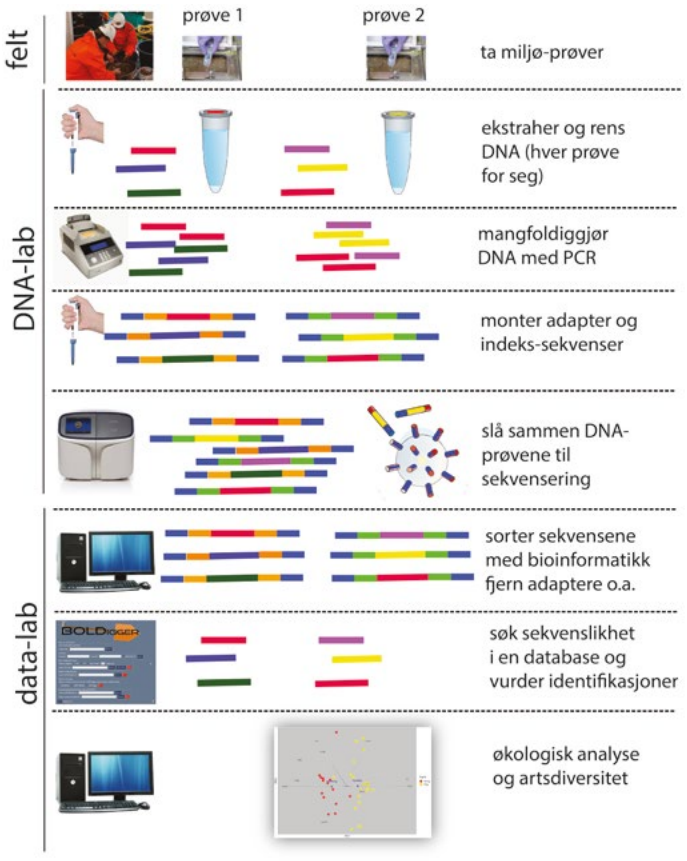


Fig. 3 | Forenklet oversikt over arbeidsflyt for metastrekkoding av prøver fra marine sedimenter.

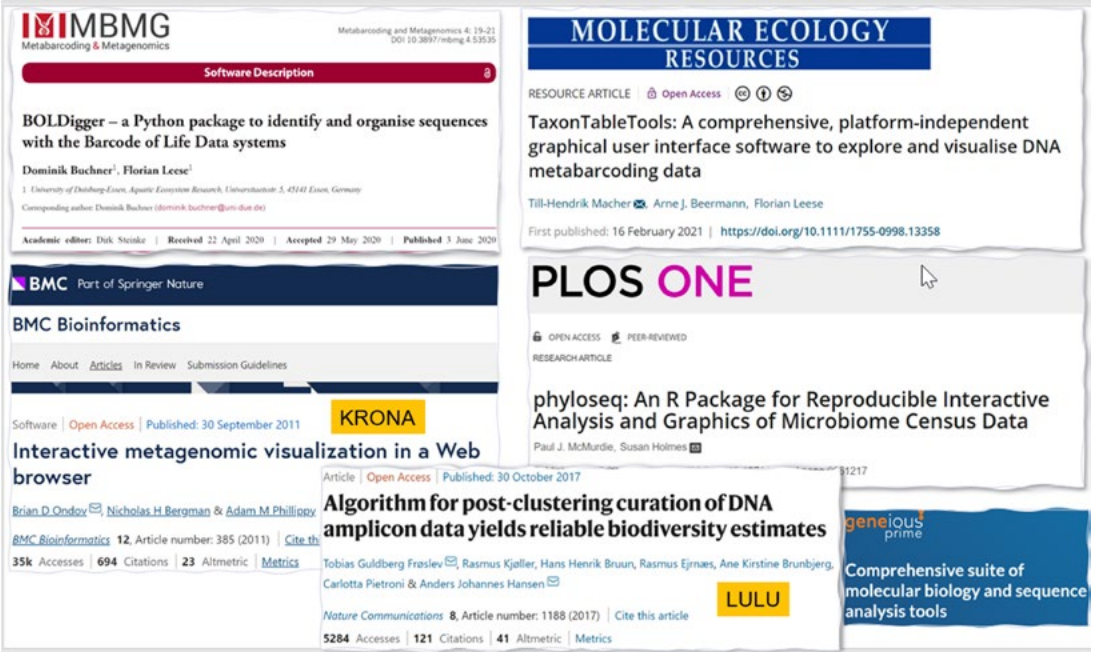


Fig. 4 | Til analyser basert på metastrekkoder kreves diverse bioinformatiske redskaper. Her er et lite utvalg dataprogrammer med grafisk brukergrensesnitt til hjelp for dem som er mindre komfortable med komplisert kommandospråk.

kjemiske metoder for å ekstrahere og rense DNA. De fleste metastrekkodingsprosjektene benytter PCR for å oppkonsentrere målsekvensene før sekvensering. I dette trinnet er det et kritisk punkt, fordi en ideelt må ha primere som passer til alle de ukjente strekkodesekvensene i prøven. Ettersom det ikke finnes universelle primere som passer på alle organismer, benytter en såkalt degenererte primere. Dette er blandinger av flere ulike primere som til sammen antas å passe til alle variantene av mangfoldet i prøven. For å sekvensere forbereder en prøvene til bibliotek med de nødvendige adaptere og indekssekvenser. Selve sekvenseringsprosessen kan ta et par timer eller flere døgn, alt etter teknologi og prøveomfang. Når sekvensene fra prøvene foreligger (Figur 5), brukes bioinformatikk for å fjerne adaptere. En må også sortere sekvensene etter indeksmerkene som holder sammenslåtte prøver fra hverandre. Fordi det kan oppstå feil, både i amplifiseringen og i sekvenseringen, bør en også forsøke å fjerne slike, så sant det finnes muligheter for å oppdage dem. En kan velge å behandle sekvensene på ulike vis ved å gruppere dem med ulike kriterier. For det første kan en betrakte hver av sekvensene som er helt like hverandre som enheter. Disse kalles da *amplikonsekvensvarianter* (ASV), og med informatikk kan en telle hvor mange sekvenser eller «reads» som forekommer av disse ASV i prøvene.

En annen type enhet er de som kalles OTU (Operational

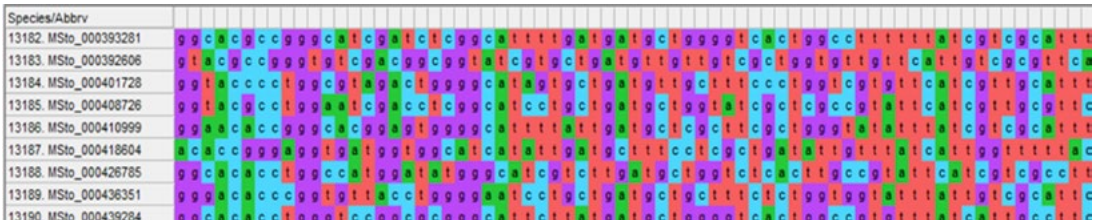
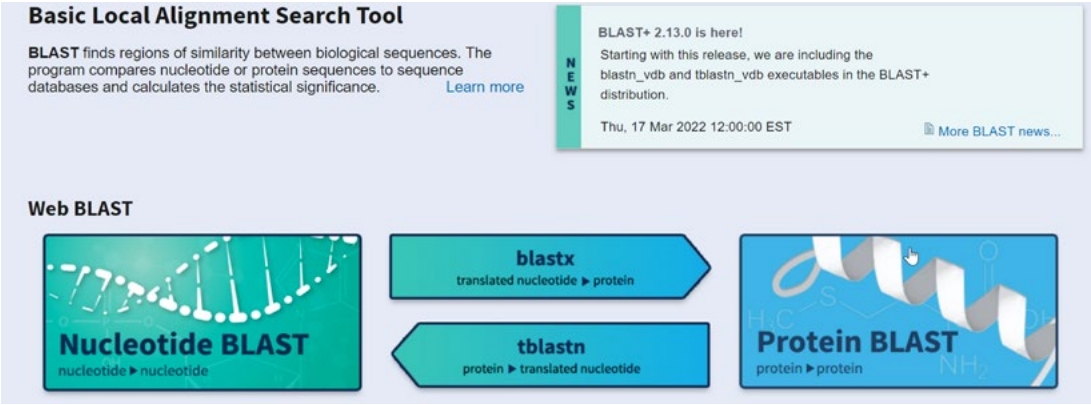


Fig. 5 | Et lite utsnitt av over 14 000 sekvenser fra et metastrekkodet prøvesett. Når adaptere og indekssekvenser er fjernet med bioinformatikk, kan en søke etter like sekvenser i en DNA-database.

Taxonomic Unit). På norsk kan vi kalle dem *operasjonelle taksonomiske enheter*. Slike OTU kan være definert ved at en grupperer sammen alle sekvenser som, for eksempel, er 97 % like. Dette tallet er en konvensjon som har vært praktisert av mikrobiologer med sekvenser fra det såkalte 16S RNA-genet hos bakterier, og som, med visse forbehold, også brukes i metastrekkoding av mitokondrier av eukaryote organismer. En tilsvarende type enhet er de såkalte BIN (Barcode Index Number). I BOLD-databasen vil en se at sekvenser er organisert i klynger i et «Barcode Index Number System». Ettersom slike OTU ofte betraktes som mer eller mindre sammenlignbare med de biologiske grupperingene vi kaller arter, brukes de gjerne som «valuta» ved kvantifiseringer av artsrikdom og biologisk mangfold. Noen forskere hevder at de ulike estimater for biologisk mangfold som baserer seg på uidentifiserte ASV eller OTU, kan være tilstrekkelig for å vurdere miljøstatus og miljøendringer. De viser gjerne til at det er kostbart og ressurskrevende å identifisere arter med tradisjonelle metoder, og at kunnskapsgrunnlaget ennå er for svakt til at man kan identifisere mangfoldet i en miljøprøve, basert på det som finnes av kjente og navngitte arter. Slike tilnærminger kan vi kalle *taksonfrie* og er klart inspirert av mikrobiologisk filosofi, der artsavgrensninger er basert på 97 % likhet. Zoologer og andre forskere vil helst forstå biomangfoldet i økosystemer med grunnlag i artsnavn og de kunnskaper om enkeltarter som er samlet gjennom lengre forskning. Da er det bra om, i det minste, noen av slike ASV eller OTU kan knyttes til kjente organismer med taksonnavn. For, som den kjente entomologen Daniel Janzen ofte har sagt: Navnet er nøkkelen til publisert kunnskap om en art.

Identifikasjon med databaser

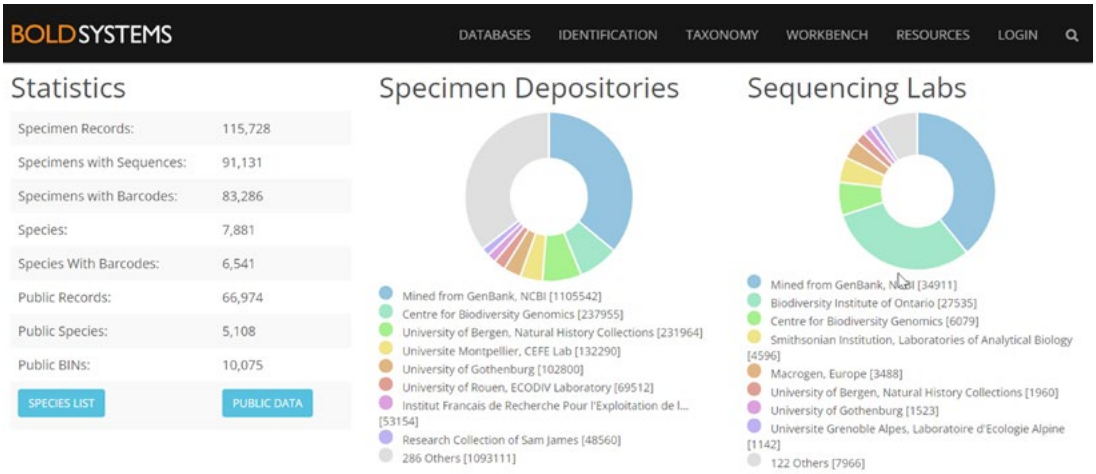
Den mest kjente databasen for sekvenser kalles GenBank, og er en funksjonelt integrert sammenslutning av tre institusjoner, en amerikansk, en europeisk, og en japansk. GenBank brukes rutinemessig for å sjekke resultater av målrettet sekvensering, særlig med såkalte BLAST-søk, der søkesekvensen forsøksvis sammenstilles med andre sekvenser for best mulig resultat med hensyn til likhet (Figur 6).



Enkelte forskere har hevdet at GenBank, med sekvenser av rundt 400 000 arter, er en god nok kilde til informasjon for metastrekkoding. Men kritiske røster viser til at databasen langt fra dekker det reelle biomangfold på mange millioner arter. Dessuten er det kjent at en betydelig andel av de navngitte sekvensene i GenBank er feilidentifisert. Boldsystems er en database som produseres nettopp for å brukes til identifikasjon med strekkoder (Figur 7). Strekkodene er hovedsakelig rundt 650 bp lange sekvenser fra det genet som kalles *cytokrom oksidase* del 1 (CO1). Ved etableringen av IBOL ble dette valgt fordi det ble ansett som godt egnet for å skille mellom arter. Fordelen med denne databasen er at prøvene er godt dokumentert, og at materialet som sekvensene stammer fra, i hovedsak er ivaretatt i vitenskapelige samlinger. De fleste NGS-plattformene produserer kortere sekvenser enn 650 bp, men noen undersøkelser indikerer at et kortere segment på litt mer enn 300 baser også vil fungere i iden-

Fig. 6 | Den amerikanske «genbanken» NCBI er en viktig referanse for gensekvenser med vitenskapelige taksonnavn. Ved ulike typer såkalt BLAST-søk, kan en forsøksvis få identifisert ukjente sekvenser. Databasene inneholder mange ulike gensekvenser fra mange organismer, men er likevel svært mangelfulle i representasjonen av det biologiske mangfoldet.

Fig. 7 | Databasen Boldsystems har i juni 2022 nesten 15 millioner prøver med DNA-strekkoder for 337 258 arter. Skjermbildet viser at Universitetsmuseet i Bergen har bidratt med 231 964 prøver av identifiserte leddormer (juni 2022). Imidlertid har mange av disse ikke gitt ønskede resultater med standard Sanger-sekvensering og mangler ennå strekkoder.



tifikasjon. Andre markører, som Cytb,12S, 16S, 18S, matK, rbcL, psbA-trnH, ITS, benyttes også som strekkoder i mange studier. Med multipleksing kan en sekvensere målrettet mot flere genmarkører samtidig. Egne databaser, med andre organismegrupper enn dyr, kan dessuten noen gang egne seg bedre for identifikasjon av slike grupper som sopp, bakterier, eller diatomeer. Noen databaser er ment tjene forskning i avgrensede økosystemer, som for eksempel MetaZooGene, for marint plankton.

Hva menes med DNA-likhet?

For å sammenligne DNA fra to prøver må en først forsøke å stille dem sammen på en slik måte at basene i de to sekvensene er mest mulig overensstemmende i rekkefølge og posisjon (Figur 8). Den underliggende idéen er at posisjoner med like baser er av samme type, fordi de har samme opphav og er såkalt *homologe*. I slike parvise samstillinger justeres posisjonene i den ene sekvensen opp mot den andre, slik at de to sekvensene blir maksimalt like. Det finnes forskjellige mål på likhet, og de følger logisk av forskjellige beregninger av ulikhet, eller «genetisk distanse». Det enkleste uttrykket for forskjell kalles *rå p-distanse.* Den kommer frem når en teller antallet posisjoner med forskjellige basepar og deretter deler antall forskjeller på sekvenslengden, altså det totale antallet posisjoner i sammenligningen. Uttrykt i prosent vil derfor to sekvenser med en p-distanse på 0,1 være 10 % ulike, eller dermed også 90 % like. Under gitte betingelser kan en forstå genetisk likhet som en indikasjon på felles identitet eller gruppetilhørighet, og dette er det grunnleggende prinsippet for DNA-strekkoding. Dersom en ukjent sekvens viser stort samsvar med en sekvens med kjent opphav, er det sannsynlig at de, for eksempel, tilhører samme art. Vi skjønner at utfallet av slike sammenstillinger er svært sentrale i en metodologi som sikter mot å identifisere taksa med ukjente sekvenser. Til dette hører også et spørsmål om *grader* av likhet og grenseverdier mellom taksa. Vil, for eksempel, en organisme med en sekvens som er 97 % lik sekvensen til arten *Aus beus* i en

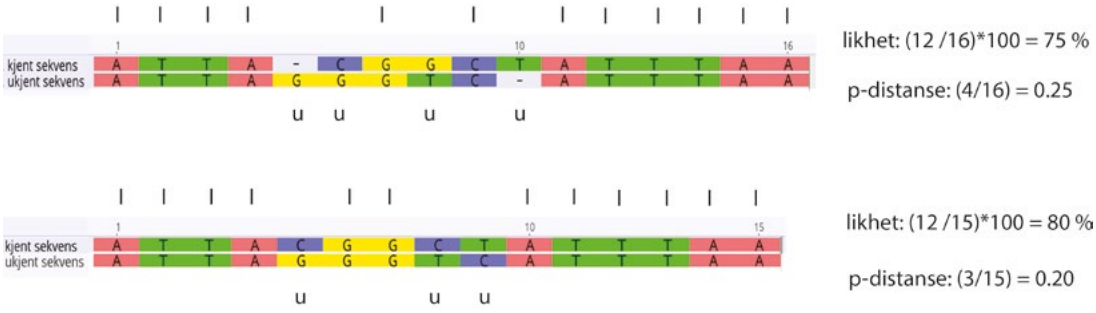


Fig. 8 | Eksempler på parvis sammenstilling av to DNA-sekvenser. Øverst er de to sekvensene 75 % like fordi 12 av de 16 posisjonene er like (1). Nederst er posisjonene justert slik at likheten mellom de to er 80 %. Såkalt p-distanse er antallet ulikheter (u) dividert på lengden (antallet posisjoner) av sammenstillingen. Den nederste sammenstillingen vil anses som bedre fordi p-distansen er mindre enn i den øverste.

database, kunne regnes som et individ av *Aus beus*? I så fall har vi jo identifisert et individ av *Aus beus*. Det fortoner seg, med et vitenskapsteoretisk perspektiv, som deduktiv metode:

- *Alle* Aus beus *har sekvenser som er 97 % eller mer like (premiss).*
- *Vår sekvens er 97 % lik en sekvens fra* Aus beus *(observasjon).*
- *Derfor må vår sekvens være fra et individ av* Aus beus *(konklusjon).*

Vi ser av Figur 8 at konklusjonen, altså utfallet av identifikasjonen, kan bli annerledes, dersom sammenstillingen av sekvensene er «uryddig» og ikke gir oss den faktisk minste mulige p-distansen. Vi må derfor forsikre oss om at vi har en algoritme som raskt kan sammenstille vår ukjente sekvens, én etter én, med alle andre kandidater og finne den som har absolutt best samsvar med vår søkesekvens. Men vi bør heller ikke glemme at konklusjonen hviler på et sett *premisser*. For det første må vi ta det for gitt at sekvensen vi sammenligner med, referansesekvensen, er riktig identifisert. Dette er, av ulike grunner, ikke alltid tilfellet.

For det andre er problemet med graden av likhet fundamentalt, ikke bare for identifikasjon av taksa, men også for hele forståelsen av biologisk mangfold i en prøve. Hvor like må to DNA-sekvenser egentlig være for at de skal kunne sies å være fra samme art? Hva om vårt premiss er strengere, slik at vår ukjente sekvens egentlig bør være 100 % lik referansesekvensen før vi kan akseptere at vi har identifisert *Aus beus*? Begge disse forholdene ved premisset i vår deduktive slutningsrekke er problemer som *taksonomisk* forskning befatter seg med. Og det er akkurat i slike spørsmål at nye DNA-sekvensdata griper inn i tradisjonell forståelse av arter, der artskjennetegn for det meste er basert på bygningsmessige trekk (morfologi).

Å telle artsmangfold – arter, OTU, eller ASV?

Framtredende pionerer for strekkoding av dyr mente at grensen mellom to arter går ved en forskjell på 2 % (riktignok med et noe annet distansemål, kalt K2P). Det førte til publiserte overskrifter som «..*Ti arter i én: DNA-strekkoding avslører kryptiske arter* ...». Denne ideen om en 2 % terskelverdi for artsavgrensing er et lån fra soppforsker-miljøet, der slike distanser mellom sekvenser fra en såkalt «gene spacer-region» blir brukt som kriterium for å skille mellom arter. Her møter vi tilsynelatende et eksempel på induktiv metode, ettersom «fakta» fra soppgener aksepteres som et allment prinsipp for alle organismer. Siden den gang har mange undersøkelser, ikke minst av evertebrater, vist at den

genetiske distansen mellom individer av det som tradisjonelt har vært oppfattet som én art, kan være opptil 20 % eller mer. Slike oppdagelser betyr at vi muligens har mer enn én art med samme navn. Dette kan kreve taksonomisk utredning. Har vi å gjøre med kryptiske arter? Men poenget er at det neppe finnes noen fast terskelverdi for avstanden mellom arter. Noen arter vil være nokså variable på CO1-genet, og ulike arter vil være preget slik av ulike evolusjonshistorier. Dessuten vil ulike gener befinne seg på et spektrum fra hypervariable til helt konserverte i sammenligninger. Derfor er genetiske avstander på ulike skalaer fra gen til gen. Genet 16S synes, for eksempel, som mer konservert enn CO1. Spørsmålet er også om lavere grad av likhet kan gi oss en identifikasjon på et høyere taksonomisk nivå. Finnes det, for eksempel, tilsvarende grenseverdier for slekt, familie, eller orden? Svaret på dette er klart nei, selv om enkelte forskere hevder dette er mulig i et snevert utvalg av én dyregruppe. Men i et bredt utvalg av organismer, slik som i en uavgrenset database, vil strekkodene miste presisjon etter hvert som likheten i CO1 faller ned mot 80 %. Ved slike verdier kan en ape, en midd, en sommerfugl og en flue alle ha samme avstand fra en søkesekvens som du på forhånd, med sikkerhet, vet er produsert fra en «tanglus». Her ligger et metodisk dilemma for metastrekkoding. Hvor sikker er identifikasjonen, dersom vår søkesekvens ikke er 100 % sammenfallende med treffsekvensen i referansedatabasen? For problemet er at det det finnes genetisk variasjon mellom individer i en art, og i et proteinkodende gen som CO1, kan det være ulikheter i hvert tredje nukleotid, uten at det får stor betydning for cellefunksjonen. Dette er fordi disse variantene likevel koder for det samme proteinproduktet. Om vi bare skulle akseptere identifikasjoner som er 100 % like, bør vi derfor sekvensere alle genetiske varianter av en art for referansedatabasen. Ellers ville vi kanskje ikke oppdage en art, fordi den bare finnes med andre varianter i prøven. Det ville, uten tvil, være en uoverkommelig oppgave.

I Figur 9 ser vi et typisk resultat fra et søk med én OTU. Det beste treffet er 98,11 % likt. Innenfor en grenseverdi på 95 % finnes også tre arter fra tre ulike slekter. Skal vi nøye oss med å konkludere at det er en dinoflagellat, klasse Dinophyceae? Kan vi også driste oss til å si at den tilhører familien Gymniodiniaceae, eller til og med foreslå en slekt?

En av utfordringene med denne metoden er altså at ulike gener muterer og endrer seg med ulik evolusjonshastighet. Det samme genet kan ha forskjellig evolusjonshastighet i ulike arter og artsgrupper. Å finne rett strekkodemarkør for et gitt taksonomisk presisjonsnivå kan være vrient. Fordi det

You searched for	Phylum	Class	Order	Family	Genus	Species	Similarity	Status	Process ID
>JI01_001443910	Pyrrophycophyta	Dinophyceae					98.11	Private	
	Pyrrophycophyta	Dinophyceae	Gymnodiniales	Gymnodiniaceae	Karlodinium	venificum	96.21	Published	DACOI005-09
	Pyrrophycophyta	Dinophyceae	Gymnodiniales	Gymnodiniaceae	Gymnodinium	catenatum	96.21	Published	DINO757-07
	Pyrrophycophyta	Dinophyceae	Gymnodiniales	Gymnodiniaceae	Amphidinium	asymmetricu	95.96	Private	
	Pyrrophycophyta	Dinophyceae	Peridiniales	Peridiniaceae	Peridinium		94.2	Published	DINO403-07
	Pyrrophycophyta	Dinophyceae	Peridiniales	Peridiniaceae	Peridinium		90.78	Published	DINO300-07
	Pyrrophycophyta	Dinophyceae	Gonyaulacales	Goniodomataceae	Alexandrium	tamarense	90.43	Private	
	Pyrrophycophyta	Dinophyceae	Gonyaulacales	Goniodomataceae	Alexandrium	minutum	90.43	Published	DINO105-06
	Pyrrophycophyta	Dinophyceae	Gonyaulacales	Gonyaulacaceae	Lingulodinium	polyedrum	90.43	Published	DINO518-07
	Pyrrophycophyta	Dinophyceae	Gonyaulacales	Gonyaulacaceae	Gonyaulax	spinifera	90.1	Published	DINO329-07
	Pyrrophycophyta	Dinophyceae	Gonyaulacales	Pyrophacaceae	Fragillidium	subglobossun	89.9	Private	
	Pyrrophycophyta	Dinophyceae	Gonyaulacales	Goniodomataceae	Alexandrium	Catenella	89.64	Private	
	Pyrrophycophyta	Dinophyceae	Gymnodiniales	Gymnodiniaceae	Gymnodinium	catenatum	89.11	Published	DINO725-07
	Pyrrophycophyta	Dinophyceae	Prorocentrales	Prorocentraceae	Prorocentrum	micans	89.11	Published	DINO851-07
	Pyrrophycophyta	Dinophyceae	Gonyaulacales		Azadinium	dalianense	89.1	Published	JRPAA6992-1
	Pyrrophycophyta	Dinophyceae	Gonyaulacales		Azadinium	cf. poporum	89.1	Published	JRPAA4362-1
	Pyrrophycophyta	Dinophyceae	Gonyaulacales	Gonyaulacaceae	Lingulodinium	polyedrum	88.96	Private	
	Pyrrophycophyta	Dinophyceae	Gonyaulacales	Goniodomataceae	Alexandrium	catenella	88.78	Published	JRPAA3148-1

Fig. 9 | I dette reelle eksemplet ser vi resultatet av et søk med én av mange sekvenser, en OTU, fra metastrekkoding mot databasen Boldsystems.org. Det beste treffet er 98,11 % lik, men er en ikke nærmere identifisert sekvens som heller ikke er publisert. Tre andre sekvenser er også mer enn 95 % like. To arter fra ulike slekter er dessuten 96,21 % like søkesekvensen.

ennå ikke finnes enhetlig forståelse av samsvar mellom arter i linnésk taksonomi og de enhetene som bygger på DNA, slike som OTU og ASV, kan det komme til store avvik mellom disse metodene for å registrere og estimere biodiversitet.

Manglende sekvenser

Dersom søkesekvensen ikke allerede finnes i en database, vil identifikasjonen feile, men i noen tilfeller kan det være fristende å vurdere treff som har mindre likhet enn 97 % fordi mange arter later til å ha mer variasjon. I et sett med over 14 000 OTU framstilt fra marine bunnprøver ga søk i Boldsystems treff for bare 103 OTU, dersom vi også tok med de som var mer enn 90 % like. Det skyldes nok delvis at marine evertebrater, til tross for stor innsats ved Universitetsmuseet og andre steder, ikke har god dekning med strekkoder av det faktiske artsmangfoldet. Weigand og kolleger (2019) fant at bare 22 % av de 16 962 artene som var registrert i europeisk marin fauna, var representert med en strekkode i Bold-databasen. Ved sammenligning av artene i det såkalte AMBI-systemet, som benyttes til vurdering og overvåkning av økologisk status i marint miljø, fant disse forfatterne at 47,6 % av de 3012 artene i systemet hadde strekkode. I Figur 10 ser vi et eksempel på et søkeresultat som skyldes manglende sekvenser. Sommerfugler, en landsnegl, en grønnalge, en minerflue og andre insekter er blant de beste treffene i referansedatabasen for en prøve fra om lag 300 m dyp i havet. Det viser bare at lignende sekvenser ikke finnes i databasen, og understreker hvor viktig det er at databasene har fyldig representasjon av sekvenser fra de organismegruppene en kan vente å finne i det miljøet en vil undersøke med metastrekkoding. Dessuten virker det usannsynlig at disse landlevende dyreartene faktisk finnes der prøvene ble tatt, på 300 meters dyp i kaldt hav. Men som vi skal se i neste avsnitt,

You searched	Phylum	Class	Order	Family	Genus	Species	Similarity	Status	Process ID
>JI01_000765	Arthropoda	Insecta	Lepidoptera	Nymphalidae	Napeogenes	sodalis	86.75	Published	GBLN2052-09
	Mollusca	Gastropoda	Stylommatophora	Limacidae	Limax		85.88	Published	GBMLG16637-13
	Arthropoda	Insecta	Diptera	Agromyzidae	Phytobia	vanduzeei	83.68	Private	
	Chlorophyta	Trebouxiophycea	Chlorellales	Chlorellaceae	Nannochloris	sp.	82.65	Published	JRPAA8455-15
	Arthropoda	Insecta	Lepidoptera	Bucculatricidae	Bucculatrix	cordiaella	82.63	Private	
	Arthropoda	Insecta	Lepidoptera	Geometridae	Desmoclystia	oniria	82.38	Published	PNGTY061-12
	Arthropoda	Insecta	Diptera	Ceratopogonidae			82.3	Private	
	Arthropoda	Insecta	Ephemeroptera	Baetidae	Cloeon	dipterum	82.3	Published	GBA23142-15
	Arthropoda	Insecta	Lepidoptera	Noctuidae	Lacinipolia	buscki	82.16	Private	
	Arthropoda	Collembola	Symphypleona	Sminthuridae			82.14	Early-Release	
	Arthropoda	Insecta	Diptera	Ephydridae			82.14	Published	BBDCQ704-10
	Arthropoda	Insecta	Diptera	Mycetophilidae	Syntemna	relicta	82.14	Published	SSJAF184-13
	Arthropoda	Insecta	Diptera	Mycetophilidae	Syntemna	relicta	82.14	Published	SSJAF1523-13
	Arthropoda	Insecta	Diptera	Ceratopogonidae			82.14	Private	
	Arthropoda	Insecta	Diptera				82.14	Private	
	Arthropoda	Insecta	Ephemeroptera	Baetidae	Cloeon	dipterum	82.12	Private	
	Arthropoda	Insecta	Ephemeroptera	Baetidae	Cloeon		82.06	Published	STUBA032-12
	Arthropoda	Insecta	Diptera	Calliphoridae	Auchmeromyia	bequaerti	82.01	Published	ASILO524-17
	Arthropoda	Insecta	Ephemeroptera	Baetidae	Cloeon	dipterum	81.97	Published	GBA22461-15
	Arthropoda	Insecta	Ephemeroptera	Baetidae	Cloeon	dipterum	81.97	Published	GBA23137-15

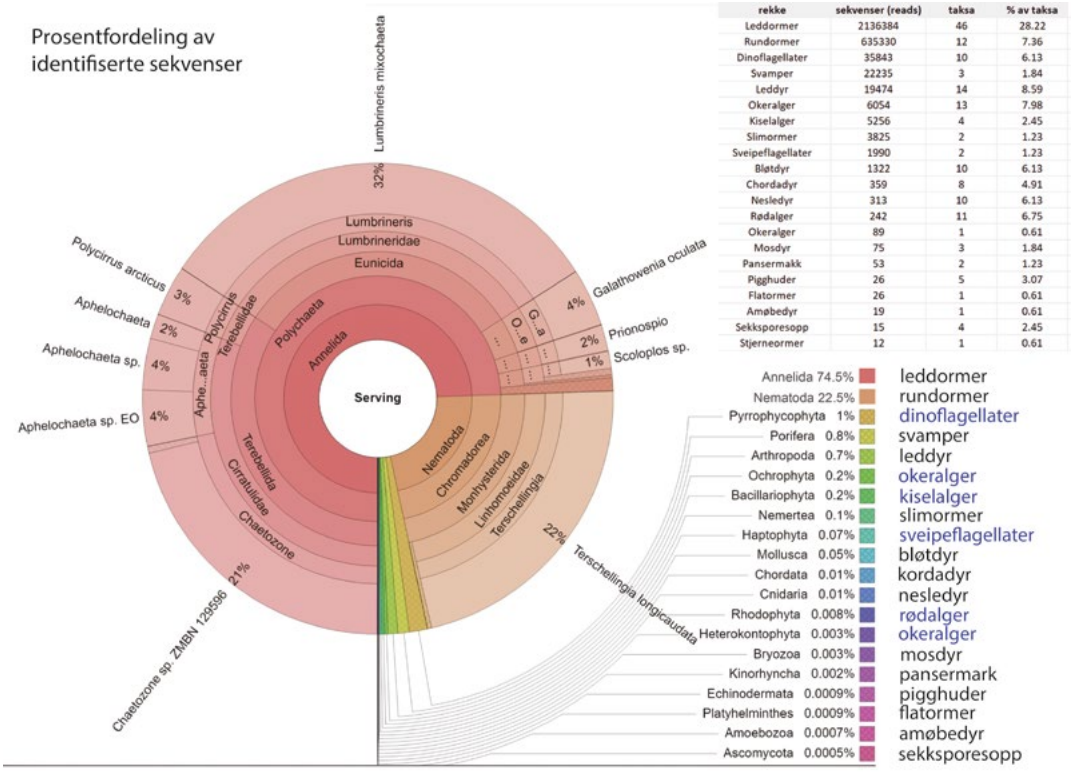
Fig. 10 | De 20 beste samsvarene med én av 14 000 ulike miljø-DNA-sekvenser er 86,75 til 81,97 % like søkesekvensen. Ulike insektgrupper, en spretthale, en snegl og en grønnalge er blant disse. Dette er et eksempel på at søkealgoritmen ikke finner en helt lik sekvens i databasen. Det viser også at likhetsverdier bedre enn 82 % her har liten verdi, ettersom både dagsommerfugl, en landsnegl, eller en grønnalge er svært usannsynlige identifikasjoner.

kan metastrekkoding by på overraskelser med «organismer på feil sted».

Levende eller døde organismer?

I Figur 11 ser vi en sammenstilling av identifiserte resultater fra bunnprøver tatt i to Svalbardfjorder. Vi ser at børstemark, leddormer, dominerer antallet avleste sekvenser, og at denne gruppen dessuten er den med flest arter i prøvene. Ellers er det sekvenser av andre bunndyrgrupper som vi ville vente å finne i slikt miljø. Det gjelder svamper, krepsdyr, bløtdyr, mosdyr, slimormer, stjerneormer, sekkedyr (i kordadyr), pigghuder og flatormer. I disse resultatene ser vi også hvorfor metastrekkoding kan være et utmerket supplement til tradisjonelle undersøkelser av bunndyrforekomster. Her kan vi få påvist organismer som er så små at de vanligvis ikke oppdages med det blotte øye, slike som pansermark. Det er en relativt dårlig kjent gruppe i den såkalte meiofauna, som er dyr som er mindre enn 1 mm lange. Vi kan også, gitt forutsetninger jeg har omtalt ovenfor, få identifikasjoner på organismer som svært få personer har kompetanse eller kapasitet til å identifisere, for eksempel hydroider og rundormer.

Men resultatene viser også at flere av de identifiserte DNA-sekvensene her er transportert inn fra omkringliggende områder. Blant de klareste eksemplene på DNA-sekvenser fra organismer som ikke lever i disse habitatene, er de ulike algene. De fotosyntetiserer og har ikke levelige kår i mørket på 300 meter. Vi registrerte også landlevende organismer, som penicillin, hårmidd, støvmidd og fluer. Det antyder at organisk materiale tilføres ovenfra, og sammen med dette fant vi DNA fra seler, gås, teist og to fiskearter. Flere av evertebratene var også slike som lever pelagisk eller veksler



rekke	sekvenser (reads)	taksa	% av taksa
Leddormer	2136384	46	28.22
Rundormer	635330	12	7.36
Dinoflagellater	35843	10	6.13
Svamper	22235	3	1.84
Leddyr	19474	14	8.59
Okeralger	6054	13	7.98
Kiselalger	5256	4	2.45
Slimormer	3825	2	1.23
Sveipeflagellater	1990	2	1.23
Bløtdyr	1322	10	6.13
Chordadyr	359	8	4.91
Nesledyr	313	10	6.13
Rødalger	242	11	6.75
Okeralger	89	1	0.61
Mosdyr	75	3	1.84
Pansermakk	53	2	1.23
Pigghuder	26	5	3.07
Flatormer	26	1	0.61
Amøbedyr	19	1	0.61
Sekksporesopp	15	4	2.45
Stjerneormer	12	1	0.61

Fig. 11 | Eksempel på identifiserte organismer fra bunnsedimenter i fjorder på Svalbard. Størrelsen på sektorene i diagrammet er skalert etter det prosentvise antallet avleste sekvenser («reads»). I tabellen ser vi dessuten at leddormer og rundormer utgjør majoriteten av sekvensene, og at leddormene er mest artstrike. I disse prøvene ble det også registrert DNA fra organismer som ikke finnes i det lysfrie dypet. Det viser at miljø-DNA kan være transportert inn fra omkringliggende områder.

mellom frittsvømmende plankton og fastsittende bunndyr gjennom livshistorien.

Dette forteller oss at vi noen ganger ikke kan være sikre på hvorvidt prøvene viser et korrekt bilde av artsforekomstene i habitatet. Det er fordi vi har sekvensert miljø-DNA, som kan stamme fra flere ulike kilder. Det kan være utskilt fra organismer som faktisk lever på stedet, eller kanskje bare nettopp har besøkt det. Det kan også være fra det organiske «regnet», som produseres andre steder, og som forsyner dypet med næringsmidler til bunnlevende konsumenter. Om opprinnelsesstedet for DNA i dette regnet kan spores, kan vil lære mer om energiomsetningen i slike dyphavslokaliteter.

Mange eller få – store eller små?

Estimater på hvor mange arter som finnes i en prøve, eller et område, er viktige i biodiversitetsstudier. Dessuten vil en gjerne ha tall på individer eller biomasse. Økologer benytter ofte begge disse parameterne i numeriske indekser for biomangfold. Jeg har allerede antydet at artsdefinisjoner som baserer seg på terskelverdier på 2–3 %, kan åpne for langt høyere estimater av biodiversitet enn de som beregnes fra telling av tradisjonelle linneske arter. I selve sekvenseringspro-

sessen kan det også oppstå feil, og noen av metodene er særlig upresise når sekvensen har flere nukleotider av samme type på rad. Det kan føre til amplikoner som har enten flere eller færre nukleotider enn originalen. Og hvis en ikke er påpasselig, kan disse feilsekvenseringene komme ut av analysene som flere enn én OTU.

På den annen side vil miljøprøver, av praktiske og økonomiske årsaker, kanskje måtte dekke en mindre romlig skala enn tradisjonelle prøvetakingsprogrammer. Dette er tydelig i våre undersøkelser av bunnprøver ved Svalbard, der opptil 80 % av børstemark-artene som ble registrert med sortering og mikroskopering, ikke ble oppdaget med metastrekkoding fra de samme prøvene. Den hittil vanligste forklaringen på slike avvik er at de uoppdagede artene mangler i referansedatabasene. Imidlertid kunne vi vise at over 90 % av disse artene faktisk har strekkodesekvenser i databasen. Derfor må det skyldes andre forhold at de ikke ble oppdaget med DNA. Vi mener forskjellene i prøveskala er én viktig årsak til avvikene, ettersom dyrene lever spredt i bunnsedimentene og de små DNA-provene til sammen bare utgjør ca. 4 % av det volumet som ble undersøkt med mikroskopering.

Men det finnes en annen feilkilde, og den har å gjøre med selve PCR-prosessen. En vellykket PCR avhenger av at vi har primere som fester seg til templatet. Men dessverre er stedene der primerne skal feste seg, fremdeles så ulike fra art til art at det som en gang ble lansert som «universelle» primere, likevel ikke passer til alle organismer. Dette er nok også en viktig årsak til mange mislykkede forsøk på å lage strekkoder for noen organismer. Løsningen på problemet har vært å bruke såkalt degenererte primere. Det er en blanding av ulike

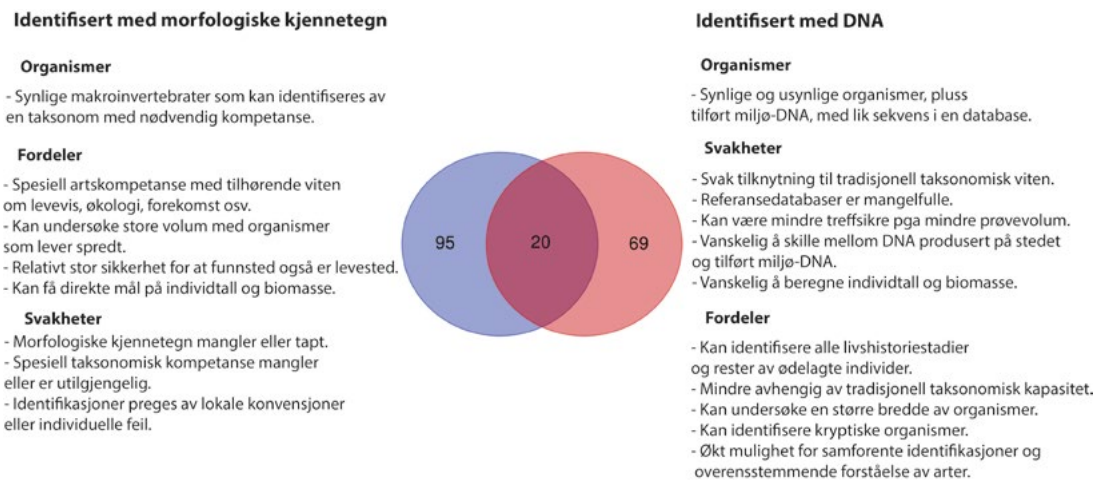


Fig. 12 | Enkel oppsummering av noen kontraster mellom to metoder: tradisjonell sortering og identifikasjon fra morfologiske kjennetegn og fra metastrekkoding. En sammenligning av arter marine bunndyr identifisert med miljø-DNA (rødt) og med tradisjonell mikroskopi (blått). Bare 20 av artene ble identifiserte med begge metodene. Relativt lite samsvar skyldes komplekse forhold som belyses nærmere i teksten.

primere som er ment å dekke variasjonen på primerstedet for templatet. Våre analyser har vist at slike degenererte primere som brukes i metastrekkoding, likevel ikke dekker den store variasjonen hos marine evertebrater. Det kan føre til at noen templater ikke amplifiseres og så å si forblir gjemt i observasjonsundersøkelsen.

Men det er andre forhold ved PCR-prosessen som ikke bare avgjør om arter oppdages eller ikke, men som også har betydning for antallet sekvenser av hver OTU eller ASV som framkommer av et sekvenseringsforsøk. For det første vil det i utgangspunktet være mer DNA fra noen organismer enn fra andre i en reaksjonsblanding med DNA fra mange ulike organismer. Vi vet ikke hvorfor. Kanskje er det fordi noen dyr er større, kanskje er det flere av dem i prøven, eller kanskje utskiller de rett og slett mer DNA med avskalling eller utskilte kroppsvæsker. For det andre vil primerne binde seg sterkere til noen templater enn til andre. Det kommer an på hvor godt primerne passer med templatsekvensen, og dessuten kompliserte termodynamiske forhold i reaksjonene. Det kan innebære at templater med svak primerbinding «stiller med handikap» og blir liggende etter mer «konkurransedyktige» templater når polymerasen er i aksjon (se figur 1). Derfor er det vanskelig å fastslå hvilken sammenheng det eventuelt finnes mellom antallet sekvensavlesninger («reads») og det som økologer tradisjonelt kaller abundans. I metastrekkoding brukes begrepet om antallet «reads». Økologer benytter mange ulike numeriske indekser for å vurdere artssammensetninger i et miljø. Vi ser at også de som inkluderer individtall, er utsatte for feilmålinger.

Avslutning

Det er antakelig langt igjen før det eventuelt er tilrådelig å erstatte tradisjonelle inventeringer av biomangfold med metastrekkoding. Men metastrekkoding har høy appell, blant annet med løfter om billigere undersøkelser og raske resultater. Derfor legges det for tiden ned en stor internasjonal forskningsinnsats på dette feltet, der mange forskergrupper, med økende bevissthet om fordeler og ulemper ved metodene, forsøker å forbedre metodologien. Et av de viktigste tiltakene for å forbedre metoden er å sørge for at databasene, som er fundamentet for slike identifikasjoner, er representative for den reelle biodiversiteten som finnes i området der en ønsker å anvende metastrekkoding. I denne prosessen gjøres stadig nye oppdagelser, som endrer forståelsen av mange taksonomiske grupper. Det er dette som, litt svulstig, har blitt kalt «den taksonomiske tilbakekoblings-løkken» («the taxonomic feedback loop»). Svulstig eller ikke, vi ser at metodene med metastrekkoding må hvile

på en god forståelse av hvilken sammenheng det er mellom DNA-sekvenser, og hvilke kategorier av identiteter de egentlig markerer. Er de naturlige enheter eller bare konstruksjoner fra menneskelig tenkning? Stikkordet er taksonomi.

Takk
Noen av eksemplene er tatt fra resultater fra et samarbeid mellom Universitetsmuseet og MAREANO-gruppen ved Havforskningsinstituttet. Takk til dem som bidro til dette arbeidet.

Ordforklaringer

Abundans: 1) antallet / mengden av et gitt takson i et område. 2) antallet sekvensavlesninger (reads) av en viss type (se ASV og OTU).

Adapter: et kort, dobbelstrenget DNA-fragment med kjent sekvens som bindes til endene av et DNA-templat ved hjelp av et enzym (en ligase) eller med PCR, slik at de skal fungere som et ankerpunkt til en primer, eller en annen fysisk binding.

Amplifisering: mengdeøkning av molekylene i en målsekvens (et templat) ved kopiering med PCR eller en annen klonemetode.

Amplikon: DNA-sekvens som er produsert med PCR eller annen amplifisering.

ASV (amplikonsekvensvariant): en unik variant av en sekvens som skiller seg fra andre sekvenser med én eller flere nukleotider. Se også OTU, som kan bestå av flere ASV.

Bibliotek: 1) et sett av DNA-prøver som er klargjort for sekvensering med adapter-/indekssekvenser. 2) en database med digitale DNA-sekvenser med tilknyttet taksonomi.

BIN: klynger av sekvenser som er rundt 97–98 % like og er samlet og navngitt under et «Barcode Index Number».

Bp: forkortelse for basepar eller nukleotidpar som angir lengden på en DNA-streng.

Degenerert primer: en blanding av primere der én eller flere av posisjonene har alternative baser som (forhåpentlig) tilsvarer variasjonen av mutasjoner i de målsekvensene som skal sekvenseres.

Demultipleksing: Se under indekssekvens.

dNTP (deoksyribonukleosid trifosfat): De fire byggesteinene som tilsettes en reaksjon for kopiering (amplifisering) av en DNA-kjede.

ddNTP (dideoksyribonukleotid): modifiserte versjoner av de fire DNA-nukleosidene som tilsettes en reaksjon for Sanger-sekvensering. Dersom et ddNTP binder seg komple-

mentært i stedet for et dNTP, vil det stoppe forlengelsen av DNA kjeden.
Endonuklease: et restriksjonsenzym som kan bryte sukker-fosfat-bindingen mellom nukleotider og slik klippe over en DNA-kjede.
Flytcelle: et sted med gjennomstrømming av kjemikalier og DNA-sekvenseringsreaksjoner.
Fusjonsprimer: en primer som brukes til å inkludere adaptersekvenser (se det) til templatene med PCR.
Genom: totalen av genetisk materiale i en organisme. I tillegg til kjerne-genom kan en celle også ha plasmider og organelle-genomer (mitokondrier, kloroplaster).
HTS (High Throughput Sequencing): betegnelse for ulike teknologier som har økt både volum og hastighet på sekvenseringsprosesser.
Indekssekvens: en unik sekvens (til forvirring også kalt strekkode), som festes til målsekvensene i en prøve slik at de kan blandes med andre unikt merkede sekvenser fra andre prøver (multipleksing) og senere sorteres fra hverandre (demultipleksing) med databehandling.
Konsensussekvens: en (digital) sekvens som er resultat av sammenstilling av flere sekvenser.
Ligase: et enzym som kan lime sammen nukleotider i en DNA-kjede.
Multipleksing: se under indekssekvens.
NGS (nestegenerasjons-sekvensering): fellesbetegnelse for nye sekvenseringsteknikker som fulgte etter Sanger-sekvensering.
OTU («Operational Taxonomic Unit»): en samling like sekvenser som noen ganger forstås som stedfortredere for arter. En likhet på 97–98 % legges ofte til grunn for gruppering i en OTU.
PCR: polymerasekjedereaksjon. En serie kjemiske reaksjoner i tre trinn som kontrolleres med temperaturregulering: 1) ca. 95 °C, som løsner bindingene mellom de parede DNA-strengene, 2) ca. 50 °C, der primere fester seg til det komplementære målstedet på en enkelstreng DNA, 3) ca. 70 °C, der frie nukleotider, ved hjelp av en polymerase, bindes komplementært til enkeltstreng DNA fra primerstedet.
p-distanse: et mål for ulikhet mellom to sammenstilte sekvenser der antallet baseulikheter er dividert på antallet posisjoner i sammenstillingen. Distansen kan også uttrykkes i prosent. Komplementet av ulikhet er likhet, slik at to sekvenser som er 3 %, ulike også er 97 % like.
Polymerase: et enzym som katalyserer syntese av en ny komplementær DNA-streng langs et enkelstreng DNA-templat.
Primer: et kort (ca. 20 bp eller mer) enkeltstrenget DNA-molekyl med kjent sekvens som tilsettes en PCR-reaksjonsblanding for å binde seg til templat-DNA og fungere som

startpunkt for kopiering av sekvensen den binder seg til. Primere inneholder de fire nukleotidene i DNA, men kan også være forsynt ned andre nukleotider, som f.eks. inosin. Inosin kan binde seg til mer eller mindre til alle nukleotidene i DNA og kan derfor brukes i en degenerert primer (se det).
Reads: sekvensavlesninger, dvs. antallet leste molekyler i en sekvensering.
Referansesekvens: en DNA-sekvens i en database som fortrinnsvis er identifisert taksonomisk slik at den også har et organismenavn.
Strekkode: 1) en kort serie av nukleotider som festes til et DNA-templat for å markere prøveidentitet (bruk heller indekssekvens), eller 2) en DNA-sekvens som antas å være en unik markør for en art (eller annen taksonomisk enhet).
Templat: DNA-fragment som skal amplifiseres og sekvenseres.
Takson (flertall taksa): en navngitt gruppe organismer med en gitt rang i et linnésk klassifikasjonssystem.
Taksonomi: 1) vitenskapelige studier som definerer, navngir og klassifiserer grupper av organismer. 2) et anvendt klassifikasjonssystem for organismer.