Metody „omiczne" w badaniach roślin

STRESZCZENIE

Przydatność komputerów w badaniach biologicznych została zauważona już w latach 60. XX wieku. Późniejszy rozwój techniki pozwolił na opracowanie wysoko wydajnych metod badawczych w biologii. Dzięki nim można uzyskać bardzo dużą ilość danych w krótkim czasie, nawet kilku dni. Z tym postępem wiąże się jednak konieczność opracowywania metod zarządzania danymi, kontroli ich jakości i sposobów ich analizy. W konsekwencji współczesna biologia w dużej mierze jest zależna od informatyki, a na styku tych dziedzin powstała bioinformatyka. Celem artykułu jest omówienie głównych współczesnych wielkoskalowych metod badań roślin i powiązanych z tymi metodami narzędzi bioinformatycznych.

dr Maciej Jończyk✉

Zakład Ekofizjologii Molekularnej Roślin, Instytut Biologii Eksperymentalnej i Biotechnologii Roślin, Wydział Biologii, Uniwersytet Warszawski

https://doi.org/10.18388/pb.2017_609

✉autor korespondujący: m.jonczyk@uw.edu.pl

Słowa kluczowe: rośliny, bioinformatyka, genomika, transkryptomika, fenomika, CRISPR-Cas9

Podziękowania: Praca powstała w wyniku realizacji projektu badawczego o nr 2020/39/B/NZ9/00801 finansowanego ze środków Narodowego Centrum Nauki

WPROWADZENIE

Za matkę bioinformatyki uważa się Margaret Dayhoff. Jej najlepiej znanym osiągnięciem było opracowanie macierzy podstawień aminokwasów (ang. *Point Accepted Mutation*, PAM) na podstawie dostępnych wówczas (1978 rok) informacji o rodzinach białek. Macierz PAM pozwala porównywać sekwencje białek i znajduje zastosowane nawet dziś [1]. Jednak już wcześniej (1962 rok) Dayhoff rozpoznała potencjał komputerów w badaniu sekwencji biologicznych i wspólnie z fizykiem Robertem S. Ledley'em opracowała COMPROTEIN – program służący do składania sekwencji peptydów z analiz sekwencjonowania białek metodą degradacji [1,2]. Był to pierwszy program bioinformatyczny, choć sam termin „bioinformatyka" powstał później, w roku 1970. Jego autorami jest dwoje holenderskich naukowców, Paulien Hogeweg i Ben Hesper. Zdefiniowali oni „bioinformatykę" jako naukę o procesach informatycznych w systemach biotycznych (ang. *the study of informatic processes in biotic systems*) [3].

Na potwierdzenie użyteczności nowej dziedziny nie trzeba było długo czekać, gdyż w 1977 roku Frederick Sanger wynalazł metodę sekwencjonowania DNA [4]. Stworzyło to potrzebę opracowania programów komputerowych służących do analizy wyników i zarządzania danymi. Bioinformatyka otrzymała silny impuls do rozwoju w latach dziewięćdziesiątych, w trakcie projektu sekwencjonowania genomu ludzkiego [5]. Było to pierwsze przedsięwzięcie generujące ogromną ilość danych biologicznych, do analizy których nie wystarczały dotychczasowe metody. Jednocześnie, dzięki rozwojowi techniki, zarówno możliwości komputerów, jak i ich dostępność szybko wzrastały. Rozwój nauki i techniki, a szczególnie miniaturyzacji, w późniejszych latach doprowadził do powstania czułych metod działających na małych ilościach prób. Jednocześnie dzięki automatyzacji stało się możliwe analizowanie wielu prób jednocześnie – nastała era „omiczna". Przyrostek „-omika" pochodzi od greckiego „-ωμα" (-ōma) – w polskim tłumaczeniu „całość" lub „zbiór" – i oznacza badania całościowe, systemowe, obejmujące wszystkie elementy danej kategorii biologicznej. Przykładami są: genomika, transkryptomika i proteomika. Genomika koncentruje się na strukturze, funkcjonowaniu i ewolucji genomu (całości materiału genetycznego organizmu). Transkryptomika obejmuje wszystko, co dotyczy cząsteczek RNA, m.in. ich: sekwencję, strukturę, funkcję, lokalizację, zmienność i poziom ekspresji. Proteomika zajmuje się całościowo białkami, m.in. ich: ekspresją, lokalizacją, modyfikacjami, interakcjami i zaangażowaniem w szlaki metaboliczne.

Bioinformatyka jest dziedziną interdyscyplinarną, a jej głównym zadaniem jest wydobycie użytecznej informacji z wielkich zbiorów danych i ułatwienie jej interpretacji. Dzięki tej nauce możliwe jest również połączenie danych z wielu eksperymentów „omicznych", spojrzenie na konkretne zagadnienie badawcze z różnych stron i otrzymanie informacji przekraczającej tę uzyskaną z osobnej analizy eksperymentów. Przykładowo, łącząc dane transkryptomiczne i proteomiczne, można sprawdzić, czy wykrywany transkrypt jest przepisywany na

Tabela 1. Podsumowanie tematów poruszonych w artykule.

	Obszar zainteresowania	Główne metody
Genomika	Poznanie sekwencji genomu, opis genów i sekwencji regulatorowych (adnotacja genomu)	Sekwencjonowanie DNA
Transkryptomika	Ekspresja genów w różnych próbach. Identyfikacja nieopisanych wcześniej transkryptów i izoform	RNA-seq
Proteomika	Skład jakościowy i ilościowy białek w różnych próbach Sekwencjonowanie białek. Identyfikacja modyfikacji potranslacyjnych	Spektrometria mas
Metabolomika	Ocena ilościowa określonych metabolitów w różnych próbach Skład jakościowy i ilościowy metabolitów w różnych próbach	Spektrometria mas, spektroskopia magnetycznego rezonansu jądrowego
Modyfikacje genetyczne	Uzyskiwanie roślin o określonych cechach, np. wydajnie produkujących metabolity wtórne	CRISPR-Cas9
Fenomika	Nieniszcząca ocena wzrostu, rozwoju i kondycji roślin	Obrazowanie, m.in. w świetle widzialnym, RGB, fluorescencyjne, 3D
Analiza funkcjonalna danych	Identyfikacja procesów biologicznych zmienianych w badanych warunkach	Analiza m.in. nadreprezentacji funkcjonalnej, sieci powiązań między białkami, ścieżek metabolicznych
Sztuczna inteligencja	Usprawnienie analizy wielkich zbiorów danych. Integracja danych z różnych metod „omicznych”	m.in. uczenie maszynowe
Zarządzanie danymi	Deponowanie danych badawczych w repozytoriach Udostępnianie danych badawczych i ich ponowne wykorzystanie	Zastosowanie zasad FAIR

białko, a więc czy jego funkcja rzeczywiście jest realizowana w komórce. Z drugiej strony pozwala to badać procesy degradacji mRNA. W niniejszym artykule opisano główne metody „omiczne” pod kątem ich wykorzystania w badaniach roślin oraz poruszono kwestie analizy danych z takich badań (Tab. 1).

GENOMIKA: SEKWENCJONOWANIE GENOMU *DE NOVO*, OKREŚLANIE FUNKCJI GENÓW

Sekwencjonowanie *de novo* (od nowa) polega na opracowaniu sekwencji genomu organizmu bez użycia materiału porównawczego. Takim materiałem mógłby być genom innej odmiany danego gatunku i takie porównanie stosuje się w tzw. resekwencjonowaniu. Dalsza część rozdziału koncentruje się ma sekwencjonowaniu *de novo*. W latach 90. XX wieku zostały zsekwencjonowane genomy pierwszych organizmów modelowych (wzorcowych), m.in. drożdży (*Saccharomyces cerevisiae*) i nicienia *Caenorhabditis elegans* [6]. W podobnym czasie (1996 rok) powstał *The Arabidopsis Genome Initiative* (AGI) – projekt poświęcony sekwencjonowaniu genomu rzodkiewnika (*Arabidopsis thaliana*), będącego rośliną modelową w badaniach roślin. Wynikiem tego przedsięwzięcia było opublikowanie w grudniu 2000 roku pierwszego genomu roślinnego [7], rok przed publikacją pierwszej wersji genomu ludzkiego [8].

Projekty te były możliwe dzięki zastosowaniu metody Sangera (terminacji łańcucha), zaliczanej dziś do metod sekwencjonowania pierwszej generacji. Metoda Sangera pozwala sekwencjonować długie cząsteczki (700-900 nukleotydów) z dużą dokładnością, jest jednak droga i pracochłonna, co więcej poznanie sekwencji jednej cząsteczki wymaga aż czterech reakcji syntezy DNA. Nawet unowocześniona i zautomatyzowana wersja metody nie jest wysokoprzepustowa, niemniej jej duża dokładność sprawia, że znajduje zastosowanie do dziś, np. przy poznawaniu sekwencji produktów łańcuchowej reakcji polimerazy (ang. *Polymerase Chain Reaction*, PCR). Właśnie wynalezienie PCR przez Kary Banks Mullis'a [9] otworzyło drogę do opracowania metod masowego sekwencjonowania. Pozwalają one w jednym przebiegu odczytać miliony, a nawet miliardy sekwencji [4]. Przełom ten został odzwierciedlony w nazwie: sekwencjonowanie nowej generacji (ang. *next-generation sequencing*, NGS). Dzisiaj te metody określane są mianem metod drugiej generacji. Pierwszym systemem masowego sekwencjonowania był 454 Genome Sequencer dostępny od 2005 roku. W obszarze tym była duża konkurencja i w kolejnych latach na rynek zostały wprowadzone m.in. instrumenty SOLiD, Ion Torrent i Solexa. W 2007 roku Solexa wykupiona została przez firmę Illumina, której urządzenia są obecnie najpopularniejsze. Choć pozwalają one odczytać tylko krótkie sekwencje (obecnie 50–300 nukleotydów w przypadku Illuminy), umożliwiły uzupełnienie braków w sekwencjach genomów. Kolejnym przełomem było opracowanie metod uzyskiwania długich odczytów (ang. *reads*) – nawet 10–20 tys. nukleotydów – i to z pojedynczych cząsteczek. Są to systemy PacBio HiFi i Oxford Nanopore, określane sekwencjonowaniem trzeciej generacji [4]. Długie odczyty mają wyższy poziom błędu niż krótkie, ale pozwalają poznać fragmenty genomu, których złożenie za pomocą metod drugiej generacji jest nieosiągalne, np. regiony bogate w sekwencje powtarzalne. Z tego powodu w obecnych czasach sekwencjonowanie *de novo* genomów jest dwustopniowe. W pierwszym etapie sekwencje są składane w chromosomy dzięki długim odczytom z metod trzeciej generacji. Następnie błędy są niwelowane dzięki krótkim odczytom z metod drugiej generacji [4].

Tabela 2. Informacje o wybranych roślinach, których genomy zostały zsekwencjonowane [67] (https://www.plabipd.de/pubplant_main.html). Wielkość genomu podana jest w milionach par zasad (ang. *megabases*, Mb).

Gatunek	Rok	Wielkość genomu (Mb)	Wykorzystanie
Arabidopsis thaliana (rzodkiewnik pospolity)	2000	135	Pierwszy zsekwencjonowany genom rośliny. Model dla roślin dwuliściennych.
Oryza sativa (ryż)	2002	500	Drugi zsekwencjonowany genom rośliny. Roślina ważna ekonomicznie. Model dla zbóż i ogólnie roślin jednoliściennych.
Populus trichocarpa (topola kalifornijska)	2006	500	Drzewo liściaste o szybkim wzroście, ważne w przemyśle drzewnym. Model w badaniach drzew, m.in. wzrostu wtórnego i drewnienia.
Physcomitrium patens	2007	472	Roślina modelowa dla mszaków. Ważna w badaniu ewolucji i adaptacji roślin lądowych, fizjologii i biologii rozwoju.
Lotus japonicus (komonica)	2008	472	Model dla roślin z rodziny bobowatych, m.in. grochu, fasoli, soi. Ważna w badaniu symbiotycznego wiązania azotu.
Zea mays (kukurydza)	2009	2400	Model dla roślin o fotosyntezie C4, a także w genetyce, fizjologii i biologii rozwoju.
Solanum tuberosum (ziemniak)	2011	844	Model dla roślin z rodziny psiankowatych. Ważnej ekonomicznie grupy roślin obejmującej również pomidora, paprykę i bakłażana.
Picea abies (świerk pospolity)	2013	20000	Pierwszy zsekwencjonowany genom drzewiastej rośliny nagonasiennej. Ważny w badaniu ewolucji drzew iglastych, w tym rozwoju szyszek.
Quercus robur (dąb szypułkowy)	2018	740	Drzewo istotne ekonomicznie i kulturowo. Ważne w badaniu długowieczności drzew i ich odporności na stres biotyczny, w tym owady i patogeny.
Triticum aestivum (pszenica zwyczajna)	2018	17000	Jedna z najważniejszych roślin uprawnych. Bardzo skomplikowany genom powstały w wyniku hybrydyzacji kilku gatunków traw. Roślina ważna w badaniu ewolucji genomów.

Poznanie sekwencji genomu to wielkie osiągnięcie, ale aby wykorzystać tę informację trzeba opracować tzw. adnotację genomu. Jest to proces dwustopniowy. Najpierw wykonuje się adnotację strukturalną. Jej celem jest określenie, które miejsca genomu zawierają geny i jakie sekwencje regulatorowe im towarzyszą. Drugim krokiem jest adnotacja funkcjonalna – określenie, jaką funkcję spełnia produkt genu (białko albo niekodujące RNA). Bez tego zsekwencjonowany genom to po prostu bardzo długi ciąg liter. Adnotacja genomu nie jest łatwa, ponieważ genomy roślinne zawierają dużo przestrzeni międzygenowych i sekwencji powtarzalnych. Rośliny są eukariontami, a więc ich geny oprócz sekwencji kodujących (egzonów) posiadają sekwencje niekodujące (introny).

Adnotacja strukturalna ma za zadanie odnaleźć regiony kodujące geny, czyli otwarte ramki odczytu (ang. *open reading frame*, ORF). Są to ciągłe sekwencje ograniczone kodonami start i stop. Adnotacja ta jest wykonywana na trzy główne sposoby [10].

Sekwencja genomu jest analizowana pod kątem zawartości („treści"). Wykorzystuje się właściwości ORF, podzielność przez 3 (ze względu na trójkowy kod genetyczny) i stosunkowo wysoką zawartość zasad purynowych. W identyfikacji ORF pomaga fakt, że niektóre aminokwasy mogą być kodowane przez więcej niż jeden kodon, a w sekwencjach kodujących określone kodony są używane częściej niż inne.

Wyszukiwane są sekwencje regulacyjne, związane m.in. z inicjacją i regulacją transkrypcji albo translacji, wiązaniem rybosomów oraz wycinaniem intronów (ang. *splicing*) [11].

Poszukuje się w bazach danych sekwencji homologicznych do wytypowanych sekwencji kodujących. W porównaniach sekwencji najczęściej wykorzystuje się różne warianty programu BLAST (ang. *Basic Local Alignment Search Tool*) [12]. Najprostsza opcja to porównanie sekwencji nukleotydowej do bazy sekwencji nukleotydowych. Jednak ponieważ sekwencja aminokwasowa jest bardziej konserwowana niż nukleotydowa, lepsze wyniki dają porównania, w których badane sekwencje są przepisywane na aminokwasy i porównywane do bazy aminokwasowej.

Wynikiem adnotacji strukturalnej jest lista potencjalnych genów wraz z ich położeniem w genomie, a także sekwencje transkryptów i ich przepisanie na sekwencje aminokwasowe, które wykorzystuje się w adnotacji funkcjonalnej. Najlepszym sposobem wykonania adnotacji funkcjonalnej byłaby weryfikacja eksperymentalna działania każdego badanego genu, ale jest ona czasochłonna, skomplikowana i kosztowna. Z tego względu, aby przyspieszyć badania, używa się analiz bioinformatycznych.

Porównuje się sekwencje aminokwasowe z bazami danych zwykle za pomocą programu BLAST lub bardziej czułego PSI-BLAST. W przypadku znalezienia sekwencji o wysokim podobieństwie wnioskuje się o homologii, a więc o podobnej funkcji. Ograniczeniem jest fakt, że żadna z baz nie zawiera pełnego zbioru sekwencji i ich opisów. Wśród baz danych wyróżnia się Swiss-Prot (https://www.expasy.org/resources/uniprotkb-swiss-prot/) [13]. Zawiera ona sekwencje białkowe opisane i nadzorowane przez ekspertów. Jednak ponieważ jest to czasochłonne i wymagające, baza ta jest znacznie mniejsza niż inne.

W sekwencjach aminokwasowych wyszukuje się charakterystyczne sekwencje, m.in. motywy, domeny (np. baza InterPro, https://www.ebi.ac.uk/interpro/) [14] i sygnały lokalizacji (np. program SignalP, https://services.healthtech.dtu.dk/services/SignalP-6.0/) [15].

Analizuje się kontekst genomowy – współwystępowanie w genomie zestawu genów może świadczyć o ich powiązaniu funkcjonalnym.

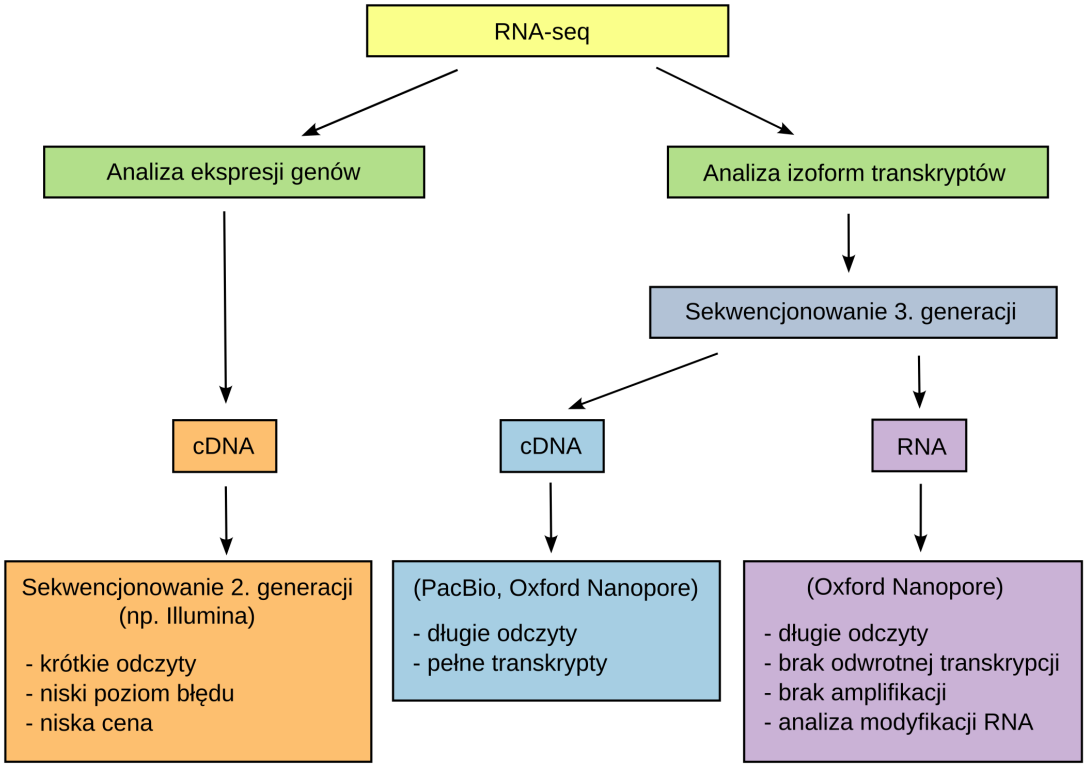
Przy opisywaniu genomu używa się danych eksperymentalnych, dostępnych dla badanego organizmu. Wykorzystuje się również dane dotyczące spokrewnionych organizmów, przy czym decydującą rolę odgrywają tu narzędzia bioinformatyczne. W celu opisania funkcji genów rzodkiewnika zostało założone *The Arabidopsis Information Resource* (TAIR) [16]. Informacje o wybranych zsekwencjonowanych genomach roślin modelowych i ważnych ekonomicznie zebrane są w tabeli 2.

TRANSKRYPTOMIKA: ANALIZA EKSPRESJI GENÓW

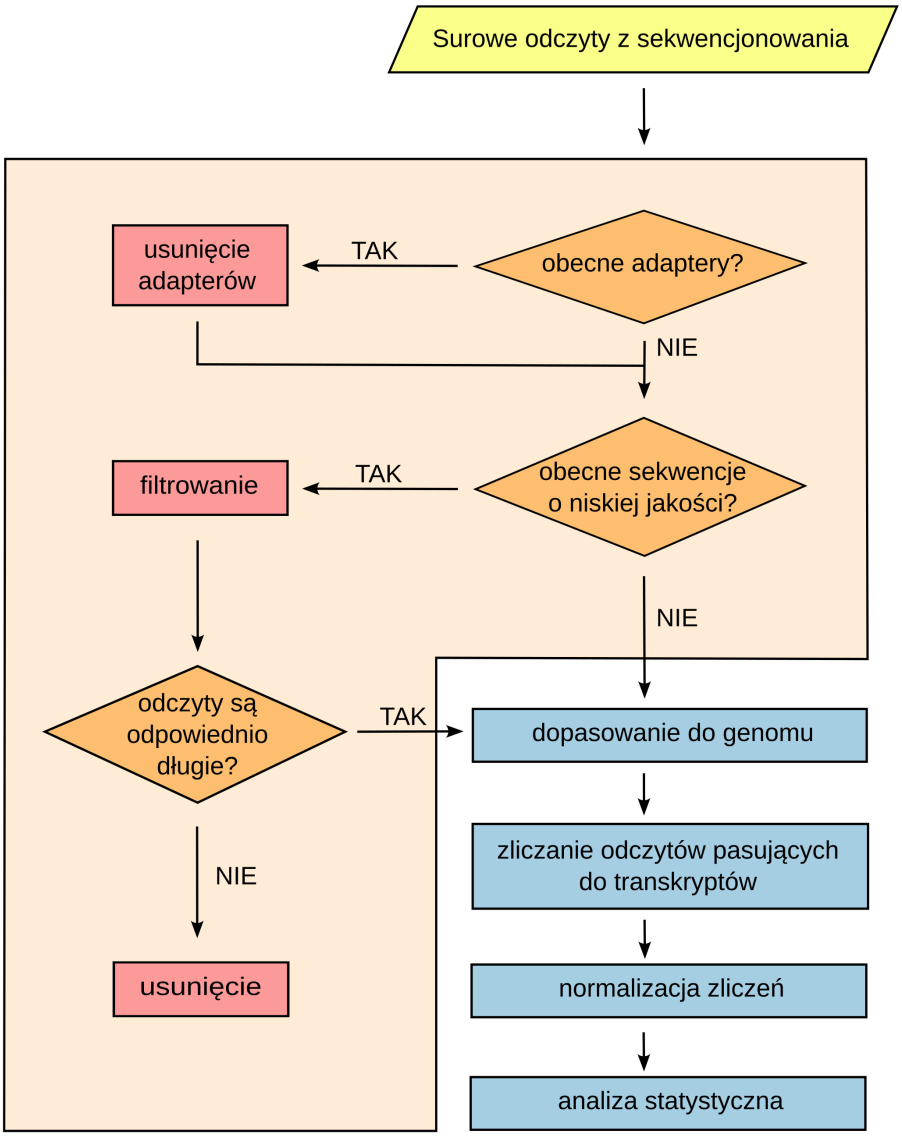
Sekwencjonowanie DNA pozwala nie tylko składać genomy, lecz także badać ekspresję genów. Metoda ta nazywa się RNA-seq, ale sekwencjonuje się w niej DNA zsyntetyzowane na matrycy mRNA. Konieczność przepisywania RNA na DNA wynika z małej stabilności RNA oraz z wymagań samej techniki sekwencjonowania, m.in. detekcji dołączanych nukleotydów podczas namnażania DNA na matrycy DNA. W przeważającej liczbie przypadków do badania ekspresji genów używa się metod drugiej generacji [4] ze względu na niski poziom błędu i stosunkowo niską cenę, najczęściej Illuminy. Krótkie odczyty uzyskiwane z platform drugiej generacji utrudniają jednak analizę izoform transkryptów. W takiej analizie przewagę mają metody trzeciej generacji, zapewniające długie odczyty. Co więcej, system Oxford Nanopore umożliwia bezpośrednie sekwencjonowanie cząsteczek RNA. Pozwala to uniknąć błędów potencjalnie wprowadzanych na etapie odwrotnej transkrypcji, takich jak nierówna amplifikacja różnych cząsteczek czy zanieczyszczenia adapterami lub primerami [17]. Możliwości wyboru metody sekwencjonowania przedstawione są na rycinie 1.

Dział biologii zajmujący się badaniem ekspresji genów to transkryptomika. Pozwala ona określać skład ilościowy i jakościowy prób zarówno pod względem RNA kodujących, jak i RNA niekodujących. Celem analizy ilościowej jest określenie, jak silnie są eksprymowane poszczególne geny i ich izoformy. Celem analizy jakościowej zaś jest ustalenie, które geny i które z ich izoform są eksprymowane. Analiza jakościowa pozwala identyfikować nieznane dotąd transkrypty lub izoformy transkryptów, co pomaga w adnotacji genomu. Transkryptomika pozwala badać zmiany transkryptomu m.in: 1) w odpowiedzi na określone warunki (np. stres, nawożenie); 2) w zależności od stadium rozwojowego i wieku; 3) w różnych tkankach i komórkach. Badania na poziomie transkryptomu możliwe są od czasu opracowania metod sekwencjonowania drugiej generacji. Pierwsze badanie z wykorzystaniem tej metody dotyczyło ludzkich komórek nowotworu prostaty i zostało wykonane za pomocą 454 Genome Sequencer w 2005 roku [18].

Chociaż transkryptomika jest techniką przełomową, w badaniach roślin przez długi czas była ograniczona do ana-



Rycina 1. Możliwości wyboru metody sekwencjonowania w zależności od celu badań.



Rycina 2. Etapy filtrowania i analizy danych RNA-seq.

liz na poziomie fragmentu organu rośliny, np. liść, korzeń – a więc materiału zawierającego różne tkanki (ang. *bulk RNA sequencing*). W konsekwencji ekspresja z różnych tkanek i grup komórek jest uśredniana, a różnice między nimi niemożliwe do rozgraniczenia. Badania transkryptomiczne na poziomie tkanek są możliwe dzięki wycinaniu (ang. *microdissection*) fragmentów tkanki pod mikroskopem. Można to robić ręcznie lub za pomocą lasera (ang. *laser capture microdissection*, LCM) [19].

Przełomem było wynalezienie technik analizy ekspresji genów osobno w każdej komórce (ang. *single-cell RNA-seq*, scRNA-seq) [20]. W przypadku roślin pojedyncze komórki otrzymywane są dzięki enzymatycznemu trawieniu ścian komórkowych (np. cellulazą lub pektynazą). Kolejnym etapem jest oczyszczanie komórek dzięki ich sortowaniu ze względu na wielkość lub obecność fluorescencyjnego znacznika. Możliwe jest to dzięki metodom takim jak mikroprzepływy lub sortowanie aktywowane fluorescencją (ang. *fluorescence-activated cell sorting*, FACS). W następnym kroku RNA z pojedynczych komórek jest izolowane i przepisywane na cDNA, które po namnożeniu jest sekwencjonowane [21]. scRNA-seq pozwoliło m.in. na opracowanie przestrzenno-czasowego atlasu ekspresji genów w korzeniach *Arabidopsis thaliana* [22].

Surowe dane z eksperymentów RNA-seq to krótkie sekwencje z danymi jakości, czyli oceną prawdopodobieństwa, że nukleotyd został odczytany poprawnie. Pierwszym krokiem analizy bioinformatycznej jest kontrola jakościowa. Odczyty o niskiej jakości są usuwane, a z pozostałych odcinane są sztuczne sekwencje dołączane w trakcie procedury. Odczyty po filtrowaniu są dopasowywane do sekwencji genomu organizmu źródłowego (ang. *alignment*), a następnie określa się, ile z nich pasuje do genów. Wynikiem są tzw. zliczenia (ang. *counts*). Etapy filtrowania i analizy danych RNA-seq przedstawione są na rycinie 2.

Krytycznym etapem analizy danych RNA-seq jest normalizacja danych, tak aby próby były porównywalne. Oprócz zastosowanych w badaniu czynników eksperymentalnych na wyniki sekwencjonowania mają wpływ: całkowita liczba

odczytów z próby (głębokość sekwencjonowania), długość transkryptów i skład próby. Uwzględnienie głębokości sekwencjonowania jest konieczne, gdyż próba zsekwencjonowana z podwojoną głębokością (dwukrotnie większą liczbą odczytów) będzie miała średnio dwa razy więcej zmapowanych odczytów dla każdego genu. Dla przykładu, chcąc ocenić ekspresję genu A w próbie RNA sekwencjonujemy ją z głębokością 10 milionów odczytów. W rezultacie 500 odczytów mapuje się do genu A. Przy sekwencjonowaniu tej samej próby z głębokością 20 milionów odczytów około 1000 z nich będzie się mapowało do genu A. Przykładem normalizacji są „zliczenia na milion" (ang. *counts per million*, CPM). Polega ona na pomnożeniu liczby odczytów zmapowanych do genu razy milion i podzieleniu przez sumę zmapowanych odczytów w próbie [23]. Normalizacja CPM pozwala porównywać ekspresję tego samego genu w różnych próbach. Niektóre metody normalizacji uwzględniają również długość transkryptów. Jest to ważne, ponieważ transkrypt dwa razy dłuższy od innego będzie miał średnio dwa razy więcej odczytów. Normalizacja tego rodzaju pozwala na porównania poziomu ekspresji różnych transkryptów w próbie, a jej przykładami są „transkrypty na milion" (ang. *transcripts per million*, TPM) i „odczyty na kilobazę transkryptu na milion zmapowanych odczytów" (ang. *reads per kilobase of transcript per million reads mapped*, RPKM). Podobną do RPKM miarą jest FPKM (ang. *fragments per kilobase of transcript per million fragments mapped*). FPKM jest używana do danych, w których cząsteczki sekwencjonowane były z dwóch stron, w trybie sparowanych odczytów (ang. *paired end*, PE) [24].

Przywołane wyżej metody normalizacji są wrażliwe na skład próby, np. na obecność niewielkiej liczby silnie eksprymowanych genów i na różnice w całkowitej liczbie eksprymowanych genów w różnych próbach. W konsekwencji żadna z wyżej wymienionych metod normalizacji nie jest zalecana przed analizą statystyczną różnic w ekspresji genów. Odpowiednie metody zostały opracowane w ramach pakietów do analizy danych RNA-seq. Są to „średnia obcięta wartości M" (ang. *trimmed mean of M values*), zastosowana w pakiecie edgeR, oraz „mediana stosunków" (ang. *median of ratios*) z pakietu DESeq2 [24].

PROTEOMIKA: POZNAWANIE SEKWENCJI BIAŁEK, BADANIE EKSPRESJI BIAŁEK

Ekspresja genów jest badana najczęściej na poziomie transkryptu, nie ma on jednak prostego przełożenia na proteom, czyli zawartość białek w komórkach. W konsekwencji, chcąc ocenić skład jakościowy i ilościowy białek w próbach, trzeba to robić bezpośrednio. Badanie białek jest trudniejsze niż badanie kwasów nukleinowych. Do budowy DNA są używane jedynie cztery rodzaje nukleotydów, podczas gdy do budowy białek u roślin jest wykorzystywane 20 rodzajów aminokwasów kanonicznych (standardowych) i jeden rzadki aminokwas niestandardowy – selenocysteina. W odróżnieniu od sekwencjonowania DNA nie można wykorzystać tu procesu istniejącego w naturze – białka syntetyzowane są inaczej niż kwasy nukleinowe. Co więcej, białka mogą składać się z kilku łańcuchów peptydowych, a więc ich struktura nie jest liniowa jak w przypadku DNA. Wiele białek ulega również modyfikacjom potranslacyjnym, co ma kluczowe znaczenie dla ich funkcji, ale dodatkowo utrudnia poznanie ich sekwencji. Pomimo tych wszystkich komplikacji metoda sekwencjonowania białek powstała już w 1949 roku dzięki Pehr Edman'owi i jest znana jako „metoda degradacji" (ang. *Edman degradation*) [2]. Ma ona jednak dużo ograniczeń, pozwala na analizę peptydu o długości maksymalnie 50 aminokwasów i jest mało wydajna.

Metoda Edman'a była niezastąpiona w badaniach białek aż do lat 90. XX wieku, gdy została wyparta przez spektrometrię mas (ang. *mass spectrometry*, MS) [25, 26]. MS pozwala: określać skład jakościowy próby z wykorzystaniem baz danych (zarówno białkowych, jak i RNA), porównywać próby pod względem jakościowym i ilościowym, a także poznawać sekwencję nieznanych białek (sekwencjonowanie *de novo*). Są dwa główne warianty procedury, *top-down* i *bottom-up*. W wariancie *top-down* analizie poddaje się oczyszczone białko o pełnej długości. Celem jest tu identyfikacja białka, jego izoform oraz ilościowe oznaczanie modyfikacji potranslacyjnych. W wariancie *bottom-up*, któremu będzie poświęcona dalsza część rozdziału, analizie poddaje się mieszaninę białek pofragmentowanych na peptydy [27]. Pierwszym etapem jest trawienie białek przez enzym – proteazę, np. trypsynę. Otrzymane peptydy są rozdzielane metodą chromatografii cieczowej (ang. *liquid chromatography*, LC). Spektrometr jonizuje peptydy, a następnie rozdziela je na podstawie stosunku ich masy i ładunku – ten etap określa się jako MS1 (pierwszą spektrometrię mas). Często procedura jest dwustopniowa (MS/MS). Wtedy wybierana jest określona liczba wierzchołków widma, odpowiadające im peptydy są fragmentowane, a powstałe jony poddawane analizie – ten etap to MS2 (druga spektrometria mas). Unikalny wzór rozdziału jonów pozwala na określenie sekwencji peptydu z wykorzystaniem baz danych. Alternatywnie można użyć programu porównującego otrzymane widmo do bazy wszystkich możliwych do uzyskania z danego peptydu. Aby zmniejszyć liczbę wyników fałszywie pozytywnych, oprócz bazy danych białek istniejących (ang. *target*) używa się bazy białek syntetycznych (ang. *decoy*), w której sekwencje białek są odwrócone albo przetasowane. Porównanie proporcji dopasowań do sekwencji *target* i *decoy* pozwala oszacować liczbę niepoprawnych identyfikacji wśród dopasowań do sekwencji *target* i ustalić wiarygodne kryteria filtrowania wyników. Jeśli celem eksperymentu jest porównanie białek z różnych prób, to są one specyficznie znakowane [28]. Metoda *bottom-up* umożliwia także sekwencjonowanie *de novo*. W tym celu porównuje się uzyskane dane do wzorcowych danych dla aminokwasów. Analiza taka może być przeprowadzona bez użycia informacji z baz danych, ale często jest o nie uzupełniana, aby poprawić jakość wyników [27].

Dane ze spektrometrii mas to widma MS1 i MS2. Na ich podstawie określa się jakiemu jonowi, peptydowi, a dalej białku odpowiadają uzyskane dane. Wynikiem jest tabela zawierająca zestawienie białek wraz z ich zawartością w poszczególnych próbach. Dane są sprawdzane pod względem jakości, a te niespełniające wymogów analizy są usuwane. W przypadku proteomiki istotnym problemem są wartości brakujące, które mogą wystąpić z dwóch przyczyn. Pierwsza z nich jest natury biologicznej. Jeśli białko nie jest wyrażane w danej próbie albo jest wyrażane na bardzo niskim

poziomie, to nie jest wykrywane. Druga przyczyna jest techniczna. Spektrometr nie jest w stanie przeanalizować wszystkich białek w próbie, poza tym jonizacja pewnych peptydów jest mało wydajna i niektóre z nich nie zostają zidentyfikowane. Dane dla białek ze zbyt wysoką liczbą brakujących wartości można usunąć. Alternatywnie dane są uzupełniane – w ich miejsce wstawiana jest najbardziej prawdopodobna wartość. Jeśli celem analizy jest porównanie ilości poszczególnych białek w różnych próbach, to przeprowadza się testy za pomocą odpowiednich modeli statystycznych [28].

METABOLOMIKA: BADANIE METABOLITÓW ROŚLIN

Efektem ekspresji genów i syntezy funkcjonalnych białek enzymatycznych jest powstawanie związków organicznych, takich jak metabolity pierwotne (np. cukry, aminokwasy, lipidy) i wtórne (np. alkaloidy, flawonoidy, terpenoidy). Metabolity pierwotne są niezbędne do życia rośliny, podczas gdy metabolity wtórne pełnią specjalne funkcje, np. w ochronie przed stresem lub interakcjach z innymi organizmami. Należy jednak podkreślić, że podział na metabolity pierwotne i wtórne nie jest ostry. Kompleksową analizą i interpretacją składu, zmienności i funkcji metabolitów zajmuje się metabolomika. Metabolity definiowane są tu jako związki o masie do 2000 Daltonów, co odpowiada masie peptydu o długości około 18 aminokwasów. Do najczęściej stosowanych metod analizy w metabolomice należą: spektrometria mas (MS) i spektroskopia magnetycznego rezonansu jądrowego (spektroskopia NMR, ang. *nuclear magnetic resonance spectroscopy*).

Podobnie jak w badaniach proteomicznych, MS połączona jest najczęściej z techniką chromatograficzną, pozwalającą wstępnie rozdzielić cząsteczki. Można użyć chromatografii cieczowej (LC) albo gazowej (ang. *gas chromatography*, GC). Ponieważ całość syntetyzowanych w roślinie związków (metabolom) przedstawia ogromne zróżnicowanie, konieczne jest dobranie metody analizy do jej celu. Celem metabolomiki nieukierunkowanej (ang. *non-targeted metabolomics*) jest wykrycie jak największej liczby związków bez wcześniejszego założenia, jakie związki są obecne. Dzięki temu można również wykryć nieznane metabolity. W takim przypadku podstawową metodą jest MS. Celem metabolomiki ukierunkowanej (ang. *targeted metabolomics*) jest wykrycie związków dobrze scharakteryzowanych i ich ocena ilościowa. W tym przypadku metoda dopasowana jest do charakteru chemicznego cząsteczki [29].

Drugą metodą popularną w analizach metabolomicznych jest spektroskopia NMR. W porównaniu z MS zapewnia wysoką precyzję i powtarzalność, ale niską czułość, a więc wymaga większej ilości próby. Za to przygotowanie prób jest łatwiejsze, a instrument zapewnia bardzo dobrą identyfikację cząsteczek. Obie metody mają wady i zalety, zatem nie można wskazać jednoznacznie lepszej, a wybór zależy od celu analizy. NMR nie jest odpowiednia np. do ukierunkowanej analizy metabolitów o niskim stężeniu, takich jak hormony roślinne (fitohormony). Pozwala za to zidentyfikować w jednej analizie związki zróżnicowane chemicznie, np. pod względem polarności [30].

Wstępna analiza surowych danych z analiz metabolomicznych (widm) obejmuje usunięcie szumów i identyfikację wierzchołków widma, odpowiadających metabolitom. Tak samo jak w przypadku proteomiki problem stanowią wartości brakujące i w metabolomice rozwiązywany jest on tak samo. Dane są również normalizowane, tak aby różne próby były porównywalne. Normalizacja może być przeprowadzana na różne sposoby, na przykład przez podzielenie przez całkowitą intensywność sygnału lub z użyciem standardów wewnętrznych. Kolejnym krokiem jest adnotacja szczytów widma, czyli ich przypisanie do konkretnych metabolitów. Adnotacja opiera się na porównaniu ze standardami lub z bazami danych (np. https://metaspace2020.org/, http://gmd.mpimp-golm.mpg.de/) [31,32].

MODYFIKACJE GENETYCZNE – ZMIANA SZLAKÓW METABOLICZNYCH

Oprócz cząsteczek niezbędnych do przeżycia i rozmnażania rośliny produkują dużo związków o specjalnych funkcjach, tzw. metabolitów wtórnych. Są one syntetyzowane w odpowiedzi na stres – mogą np. odstraszać roślinożerców, będąc dla nich trującymi lub niesmacznymi. Jednocześnie wiele metabolitów wtórnych roślin jest cennych dla człowieka i używane są m.in. jako leki, barwniki, aromaty lub składnik żywności [33]. W wyższych stężeniach metabolity wtórne, które są produkowane przez roślinę, mogą być toksyczne dla niej samej. Z tego względu obecne są w niej w małych ilościach, a co za tym idzie, aby uzyskać ilość istotną terapeutycznie, potrzeba dużej masy roślin – w przeważającej mierze uzyskiwanych z naturalnych populacji [34]. Jednocześnie synteza wielu metabolitów wtórnych jest skomplikowana i mało wydajna [35]. Sprawia to, że dostępność leków bazujących na roślinnych metabolitach wtórnych jest ograniczona i są one drogie. Z drugiej strony wysokie zapotrzebowanie na te substancje powoduje silną eksploatację naturalnie występujących populacji roślin, zagrażając ich istnieniu. Alternatywą jest hodowla nowych odmian wydajniej produkujących pożądane substancje. Można wybrać odmiany lub osobniki o pożądanych cechach i namnażać je w warunkach laboratoryjnych (*in vitro*). Wykorzystuje to unikalną zdolność komórek roślinnych do odróżnicowania, czyli zmiany programu rozwojowego i przekształcenia się w inne komórki.

Na fenotyp, czyli obserwowalne właściwości organizmu (m.in. kolor, wielkość, budowę) mają wpływ czynniki środowiskowe, genotyp i mechanizmy epigenetyczne. Czynniki środowiskowe, takie jak ilość i jakość światła, temperatura i skład pożywki, możemy dziś kontrolować w szerokim zakresie, choć w stosunkowo małej skali, wykorzystując komory do hodowli roślin lub szklarnie. Wyżej wymienione czynniki są badane pod kątem zwiększania produkcji metabolitów wtórnych [36]. Sprawdzany jest również wpływ hormonów i elicytorów [37,38]. Elicytory to substancje lub czynniki fizyczne powodujące ekspresję genów regulacyjnych, determinujących określone reakcje biochemiczne, a dalej odpowiedź morfologiczną i fizjologiczną. Kategoria ta jest bardzo szeroka i obejmuje m.in. ekstrakty z grzybni, metale ciężkie, promieniowanie UV [38]. W porównaniu z modyfikacjami genetycznymi, wymienione sposoby zwięk-

szania produkcji metabolitów wtórnych są tanie i łatwe, jednak zwykle dają gorszy wynik [37].

Zmiany na poziomie genotypu były tradycyjnie wprowadzane dzięki krzyżowaniu roślin blisko spokrewnionych i hodowli wsobnej. Metody mutagenezy znane są od drugiej dekady XX wieku [39], pierwsze z nich prowadziły do losowych zmian w przypadkowych miejscach genomu. Odkrycie enzymów restrykcyjnych w latach 70. XX wieku otworzyło drogę do wprowadzania do genomu określonych sekwencji [40]. Współczesne metody pozwalają na edycję genomu, czyli jego precyzyjne modyfikacje, a najpopularniejsza jest metoda CRISPR-Cas9 opracowana w 2012 roku. Pozwala ona zarówno wprowadzać krótkie losowe insercje lub delecje, jak i precyzyjnie wycinać albo wstawiać określoną sekwencję nawet na poziomie jednej pary nukleotydów [41]. CRISPR-Cas9 jest wariantem systemu CRISPR-Cas, bakteryjnego adaptacyjnego układu odpornościowego. Bakteria w locus zwanym CRISPR (ang. *clustered regularly interspaced short palindromic repeats*) posiada krótkie sekwencje przechwycone od wirusów. Podczas infekcji, DNA wirusa łączy się z komplementarnym RNA i jest degradowane dzięki ukierunkowanemu cięciu przez białko Cas [42]. W porównaniu z konkurencyjnymi technikami wynalezionymi w tamtym czasie, CRISPR-Cas9 nie wymaga projektowania nowego białka do każdej nowej modyfikacji [43]. Niemniej, aby zwiększyć ekspresję białka Cas9 można dopasować je do gatunku, który ma być modyfikowany. Chodzi o zmianę kodonów na te preferowane w genach o wysokiej ekspresji, a także o umieszczenie wydajnego promotora, a nieraz intronu, co również u niektórych gatunków zwiększa ekspresję genu [44]. Do każdej modyfikacji genetycznej trzeba zsyntetyzować specjalną cząsteczkę RNA (ang. *single guide RNA*, sgRNA). Zawiera ona fragment konieczny do działania Cas9 oraz 20-nukleotydową sekwencję komplementarną do fragmentu genomu, który ma być zmodyfikowany. Wymaga to znajomości sekwencji genomu i jego adnotacji, a także uwzględnienia specyficzności, tak aby zmiana wprowadzona była tylko w pożądanym miejscu [45]. Kolejnym krokiem jest wybór odpowiedniego wektora zapewniającego żądaną aktywność i dopasowanego do modyfikowanego organizmu. Pomaga w tym ogólnie dostępna baza Addgene (https://www.addgene.org/) [46]. Weryfikacja wyniku eksperymentu również wymaga narzędzi bioinformatycznych. DNA z potencjalnych mutantów sekwencjonuje się, najczęściej metodą Sangera lub drugiej generacji, np. Illumina. Techniki bazujące na CRISPR-Cas są rozwijane i opracowywane są inne metody edycji genów [43]. W tworzeniu każdej z nich niezbędne jest użycie narzędzi bioinformatycznych. Mimo wielkiego postępu wydajność modyfikacji genetycznych nie jest wysoka – do 50%, gdy używa się komórek pozbawionych ściany (protoplastów) [47]. Niemniej zaledwie jedna zmieniona komórka wystarczy do uzyskania całego organizmu, w którym wszystkie komórki mają dokładnie tę samą modyfikację.

Do zmiany ścieżki syntezy pożądanych związków metodą modyfikacji genetycznych konieczna jest znajomość zarówno sekwencji genomu badanej rośliny, jak i szlaku syntezy tych związków. Niektóre z tych szlaków są poznane, np. szlak kwasu szikimowego [48], m.in. dzięki badaniom proteomicznym i metabolomicznym [49]. Celem inżynierii szlaków metabolicznych może być ingerencja w wydajność określonych etapów, zmiana architektury szlaku lub usunięcie hamującego wpływu produktu końcowego [50]. Przykładowo, dzięki wprowadzeniu trzech genów do ścieżki syntezy tokochromanoli udało się zwiększyć syntezę witaminy E w roślinach tytoniu i pomidora. Również dzięki transformacji trzema genami udało się uzyskać rośliny ryżu o trzykrotnie wyższej zawartości witaminy B1 w ziarniakach w porównaniu do odmiany standardowej. Z kolei potencjał zwiększenia ilości beta-karotenu w ziarniakach był badany za pomocą transformacji roślin kukurydzy różnymi kombinacjami pięciu genów szlaku syntezy karotenoidów [51].

FENOMIKA: BADANIE FENOTYPU ROŚLIN

Opisane wcześniej nauki „omiczne" pozwalają uzyskać bardzo dużo danych na poziomie molekularnym, jednak celem wielu badań jest zmiana obserwowalnych cech rośliny (fenotypu), np. tempa wzrostu, morfologii, składu. Pomiar i analizę fenotypu na różnych poziomach organizacji, od organów po okrywę roślinną, określamy mianem fenotypowania [52]. Stosunkowo długo było ono ograniczone do pomiarów ręcznych, a co za tym idzie było długotrwałe, podatne na błędy i wykonywane w małej skali. Co więcej, ocena np. biomasy wymagała zabicia rośliny, a więc mogła być zrobiona tylko raz w sposób „inwazyjny" i nie pozwalało to na śledzenie zmian w czasie. Dopiero postęp po roku 2010 w takich dziedzinach, jak robotyka i obrazowanie pozwolił na rozwój automatycznego fenotypowania w dużej skali. Ogólnie takie systemy można podzielić na dwa rodzaje: „roślina do sensora" – rośliny są przemieszczane do urządzeń pomiarowych, a następnie przenoszone z powrotem; „sensor do rośliny" – maszyna zbliża się do rośliny i wykonuje pomiary. Do tego drugiego typu należą duże instalacje obejmujące konkretne pole, ale również maszyny o ruchu swobodnym, naziemne i latające, zarówno kierowane przez człowieka, jak i bezzałogowe (drony) [53]. Maszyny te mogą być używane w różnych miejscach: w fitotronie, w szklarni, na polu. W systemach „roślina do sensora" możliwy jest pomiar masy rośliny, natomiast w obu wymienionych systemach możliwe jest nieniszczące obrazowanie w różnych widmach światła. Rejestrowana może być np. fluorescencja chlorofilu lub światło widzialne w zakresie czerwieni, zieleni i błękitu (ang. *Red*, *Green*, *Blue*, RGB). Parametry te dają informację o kondycji rośliny. Często również dzięki kilku kamerom możliwe jest odtworzenie trójwymiarowego obrazu rośliny i ocena jej masy [52,53].

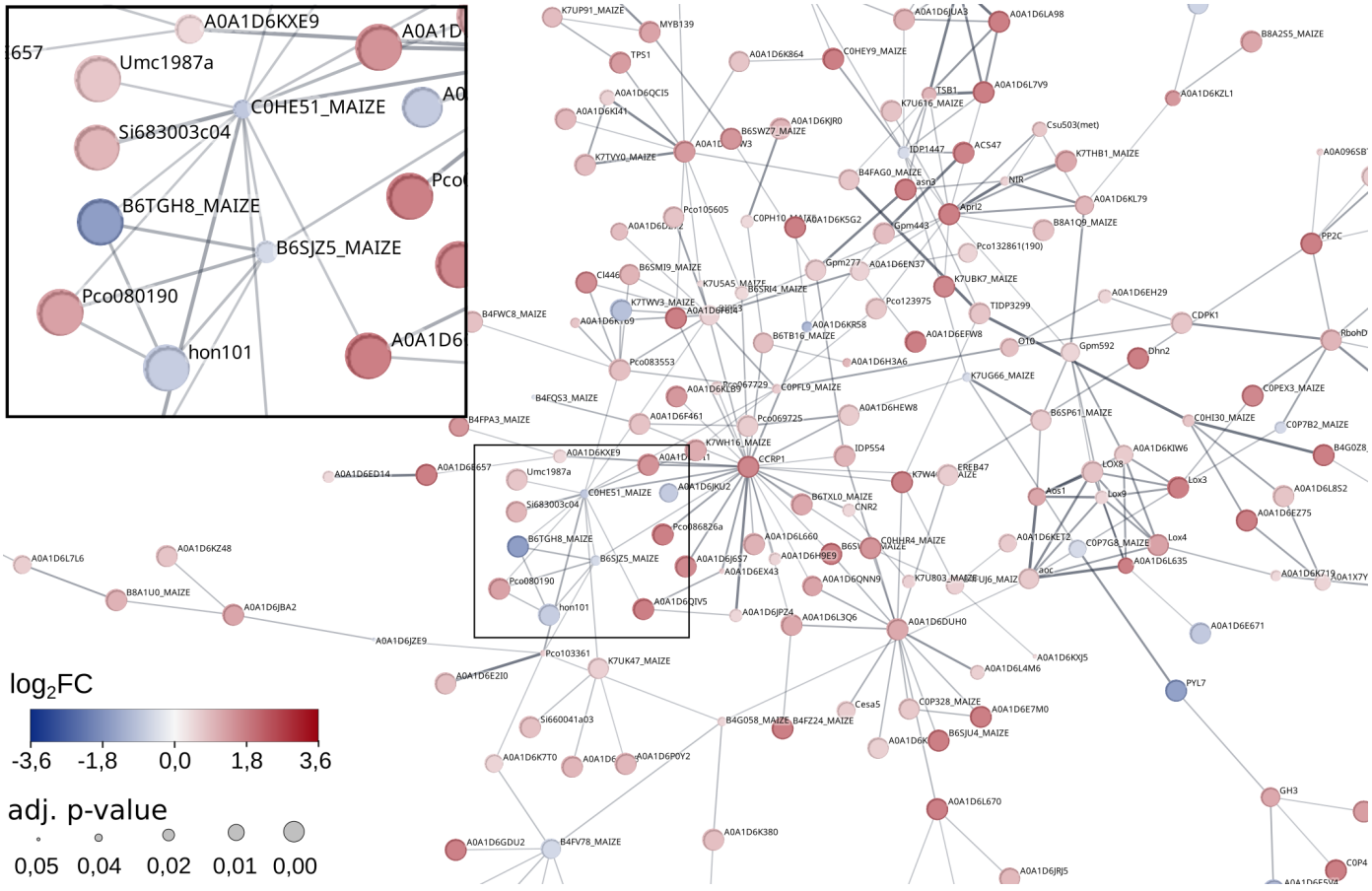
Badaniem fenotypu w kontekście genotypu i środowiska zajmuje się fenomika (ang. *phenomics*). Eksperymenty tego typu dają bardzo dużo danych, które mogą być analizowane na wiele sposobów. Można np. modelować rozwój roślin w czasie za pomocą krzywych wzrostu. Jeśli eksperyment przeprowadzany był na polu z jednoczesnym pomiarem parametrów pogodowych, to możliwe jest sprawdzenie, jak wpływają one na fenotyp roślin. Ciekawym typem analizy jest badanie asocjacyjne w skali genomu (ang. *Genome Wide Association Study*, GWAS). Polega ono na korelacji cech fenotypowych z wariantami genetycznymi i ma na celu znalezienie genów lub sekwencji regulatorowych silnie związanych z daną cechą organizmu [54].

ANALIZA FUNKCJONALNA DANYCH „OMICZNYCH”

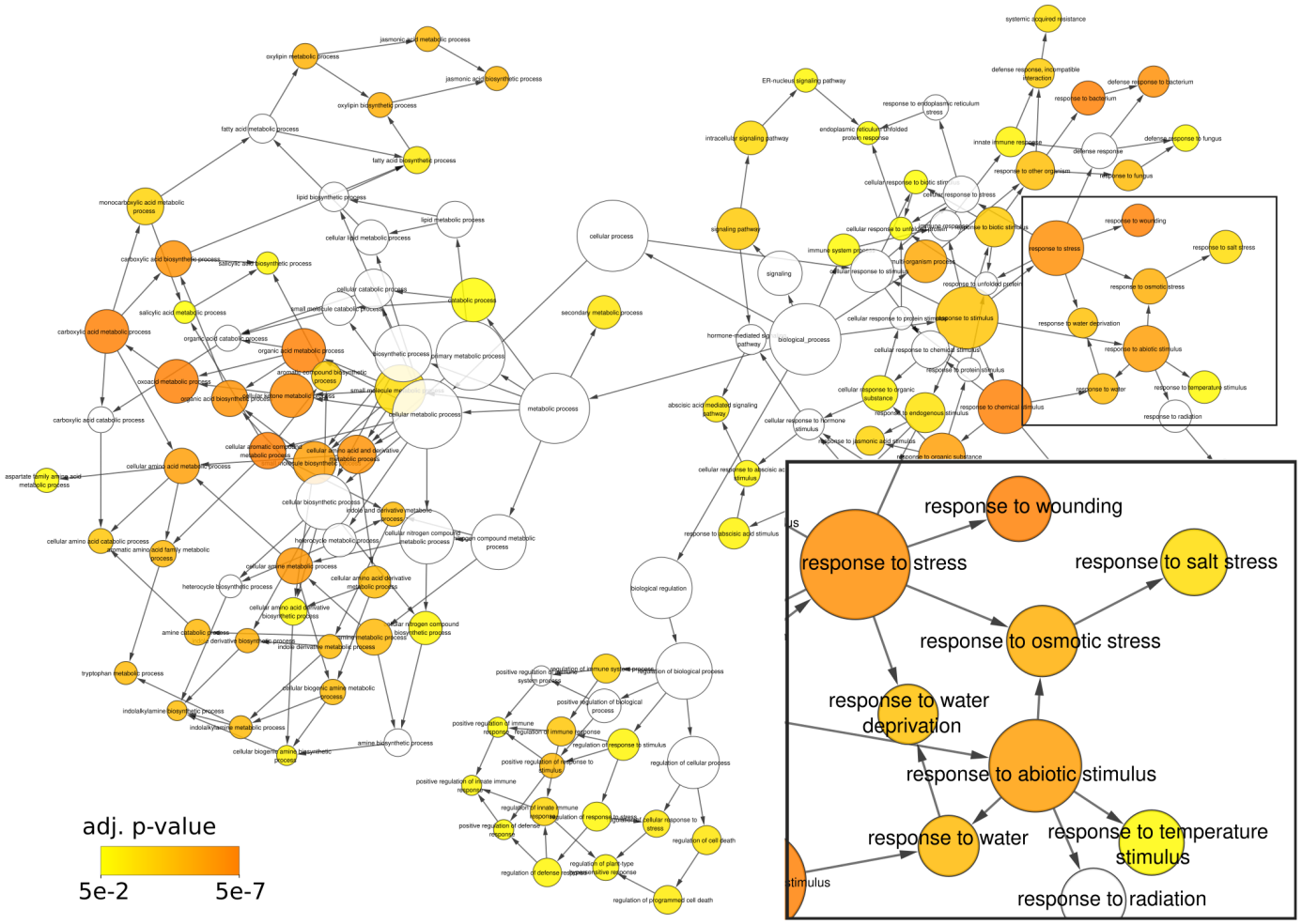
W rozdziałach o transkryptomice, proteomice i metabolomice opisane zostały ogólnie podstawowe metody analizy surowych danych. Metody te są specyficzne dla danej „omiki”. W dalszej zaawansowanej analizie wyników wykorzystuje się jednak podobne narzędzia. Dotyczy to szczególnie danych transkryptomicznych i proteomicznych. W przypadku porównywania wariantów eksperymentalnych stwierdzenie różnic między nimi to dopiero początek. W danych poszukuje się wzorców i wykonuje analizę funkcjonalną. Jedną z metod jest klastrowanie, które pozwala wydzielić grupy, np. genów albo białek o podobnej ekspresji. Z kolei korelacja wzorów ekspresji umożliwia zbudowanie sieci powiązań (ang. *coexpression network*). Dane transkryptomiczne i proteomiczne można również połączyć z dostępnymi w bazach sieciami powiązań między białkami. Bazy te zawierają dane o bezpośrednich interakcjach fizycznych zweryfikowanych laboratoryjnie, a niektóre również dane o interakcjach pośrednich (funkcjonalnych) oraz informacje przepisane od innych organizmów na podstawie homologii. Przykładem dużej bazy integrującej dane z różnych źródeł jest STRING (https://string-db.org/) [55]. Sieć powiązań między białkami można przedstawić graficznie, np. w programie Cytoscape (https://cytoscape.org/) [56], i ją analizować, m.in. wyszukiwać geny kluczowe o największej liczbie oddziaływań lub nałożyć na sieć dane, np. o ekspresji genów (Ryc. 3).

Wyniki dotyczące ekspresji genów i wyniki metabolomiczne można także analizować, wykorzystując bazy danych dedykowane szlakom metabolicznym. Największą taką bazą danych, dotyczącą roślin, jest *Plant Metabolic Network* (PMN, https://plantcyc.org/) [57]. Zawiera ona ścieżki metaboliczne dla 155 gatunków roślin i glonów, a także ujednoliconą, ogólną bazę dla roślin. W zakres PMN wchodzą zarówno dane eksperymentalne, jak i przewidywania obliczeniowe. Oprócz samych ścieżek PMN zawiera informacje o reakcjach, enzymach oraz odnośniki do literatury. Bazą o podobnym zakresie jest *Kyoto Encyclopedia of Genes and Genomes* (KEGG, https://www.genome.jp/kegg/) [58]. Zawiera ona informacje na temat różnych grup taksonomicznych. W kontekście roślin gromadzi dane na poziomie systemowym (ścieżki metaboliczne i klasyfikacje funkcjonalne, m.in. genów i cząsteczek chemicznych), genomicznym i metabolicznym. KEGG, podobnie jak PMN, pozwala na analizę danych.

Najczęściej wykonywaną analizą danych transkryptomicznych i proteomicznych jest nadreprezentacja kategorii funkcjonalnych. Punktem wyjścia jest lista genów, np. o zwiększonej ekspresji w danych warunkach, i adnotacja funkcjonalna genów badanego organizmu. Analiza nadreprezentacji pozwala stwierdzić, które kategorie funkcjonalne genów są reprezentowane w naszych zbiorach bardziej,



Rycina 3. Sieć oddziaływań białek kukurydzy sporządzona na podstawie genów o istotnie zmienionej ekspresji w stresie odwodnienia. Użyto danych z pracy [65] i bazy oddziaływań białek STRING w wersji 12.0 [55]. Kolor węzła odpowiada zlogarytmowanej zmianie ekspresji w stresie względem kontroli (ang. log_2 *fold change*, log_2FC). Błękit odpowiada represji, natomiast czerwień wyższej ekspresji w stresie. Wielkość węzłów odpowiada p-wartości dopasowanej na porównania wielokrotne (ang. *adj. p-value*) – im większy węzeł, tym bardziej istotny wynik. Grubość linii łączących węzły odpowiada punktacji oddziaływania w bazie STRING – im grubsza linia, tym wyższa punktacja (pewniejsze oddziaływanie). Fragment sieci jest powiększony.



Rycina 4. Nadreprezentacja kategorii funkcjonalnych Gene Ontology dla sieci z rysunku 3. Analiza obejmowała klasę Biological Process. Użyto danych ekspresji genów z pracy [65] i adnotacji GO z pracy [66]. Próg p-wartości dopasowanej na porównania wielokrotne (ang. *adj. p-value*) ustawiono na 0,001. Kolor węzłów odpowiada istotności statystycznej, natomiast ich wielkość odpowiada wielkości kategorii GO. Fragment sieci jest powiększony.

niż wynikałoby to z przypadku. W analizie tej najczęściej używa się ontologii genów (ang. *Gene Ontology*, GO) [59]. Składa się ona z dwóch elementów: hierarchicznego systemu kategorii funkcjonalnych wraz z ich relacjami oraz z adnotacji GO genów. W ramach GO kategorie zgrupowane są w trzy klasy: funkcji molekularnej (ang. *Molecular Function*), procesu biologicznego (ang. *Biological Process*) i lokalizacji komórkowej (ang. *Cellular Component*). Przykład wyniku analizy nadreprezentacji GO przedstawiono na rycinie 4. Analizę nadreprezentacji można również wykonać dla kategorii funkcjonalnych zawartych w KEGG.

SZTUCZNA INTELIGENCJA I UCZENIE MASZYNOWE, INTEGRACJA DANYCH „OMICZNYCH”

W ostatnich latach istotną rolę w badaniach „omicznych” spełnia sztuczna inteligencja (ang. *artificial intelligence*, AI). Pozwala ona uzyskać informacje, które mogą być niedostępne lub trudne do wychwycenia dla człowieka. Największe zastosowanie ma tu gałąź AI zwana uczeniem maszynowym (ang. *machine learning*, ML). Polega ono na wytrenowaniu modelu na dużym zestawie danych, weryfikacji modelu, a następnie wykorzystaniu go, np. do klasyfikacji (przypisywania obiektów do kategorii), klastrowania (grupowania) lub predykcji (przewidywania wyników dla nowych danych). Uczenie maszynowe znajduje zastosowanie w analizie danych metabolomicznych za pomocą sieciowania molekularnego (ang. *molecular networking*). Opiera się ono na tym, że w szlakach metabolicznych w kolejnych reakcjach powstają cząsteczki o podobnej budowie, dające podobne wzorce fragmentacji w widmach MS. Analiza sieciowania molekularnego pomaga zidentyfikować te reakcje, jak również wykryć nieznane dotąd cząsteczki [60]. ML spełnia podwójną rolę w fenotypowaniu. Z jednej strony pozwala optymalizować ruch wykonujących pomiary maszyn, np. uczyć je omijania przeszkód. Z drugiej strony ML pozwala analizować dane, i to na różnych poziomach, począwszy od rozpoznawania obiektów i wydzielania roślin z tła, aż do ich klasyfikacji ze względu na mierzone parametry, np. roślina pożółkła – chora, roślina zielona – zdrowa [61].

Bioinformatyka umożliwia połączenie danych z różnych dziedzin „omicznych” (badania „multiomiczne”) i spojrzenie na określone zagadnienie z różnych stron. Przykładowo integracja danych transkryptomicznych i metabolomicznych pozwala sprawdzić, z ekspresją których genów jest skorelowana ilość pożądanego produktu. Z wyników MS można stworzyć sieć, w której metabolity połączone są krawędziami, zgodnie z podobieństwem widm fragmenta-

cji. Na te dane można nałożyć wyniki ekspresji genów lub białek, co pomaga wytypować geny odpowiedzialne za syntezę badanych metabolitów [62]. Dane „multiomiczne" są bardzo złożone, a ich analiza stanowi wyzwanie. Z tego względu w coraz większym stopniu znajduje tu zastosowanie sztuczna inteligencja. Generatywna AI pozwala m.in. zautomatyzować proces tworzenia modeli sieci metabolicznych. Możliwe są również analizy par enzym-substrat, w czym pomaga model AlphaFold, przewidujący strukturę przestrzenną białek [60].

ZARZĄDZANIE DANYMI I ZASADY FAIR

Bioinformatyka jest kluczowym elementem badań wysokoprzepustowych. Jednak do przechowywania i udostępniania danych z tych badań używane są narzędzia szeroko rozumianej informatyki. Jako pierwsze w bazach danych były umieszczane wyniki transkryptomiczne i proteomiczne, a najstarsze bazy gromadzące takie dane to ArrayExpress i Gene Expression Omnibus (GEO) [63]. Umieszczanie danych z badań wielkoskalowych w repozytorium stało się wymogiem przy publikowaniu artykułów w czasopismach naukowych. Ma to zapewnić czytelnikom dostęp do danych, będących podstawą wyników opisanych w konkretnym artykule. Koncepcja przejrzystości badań naukowych w odniesieniu do danych badawczych została rozwinięta jako zasady FAIR. Kolejne litery skrótu oznaczają: *Findable* (możliwe do znalezienia), *Accessible* (łatwo dostępne), *Interoperable* (możliwe do połączenia z innymi danymi i działające w różnych systemach) i *Reusable* (możliwe do ponownego wykorzystania) [64]. Dwie pierwsze zasady oznaczają, że dane powinny mieć trwały identyfikator, by można było je znaleźć, i powinny być możliwe do pobrania za darmo bez ograniczeń prawnych. Zasada *Interoperable* odnosi się do ujednoliconych formatów, ułatwiających ponowne wykorzystanie danych i łączenie danych z różnych źródeł. Dzięki temu możliwe są analizy danych zarówno z wielu eksperymentów tego samego typu, np. transkryptomicznych, jak i analizy „multiomiczne". Podstawą ostatniej zasady (*Reusable*) są jasno określone warunki ponownego wykorzystania danych np. w badaniach naukowych i edukacji.

PODSUMOWANIE

W niniejszej pracy opisano główne metody „omiczne", stosowane w badaniach roślin i ogólnie organizmów żywych. Wejście biologii na drogę badań wysokoprzepustowych sprawiło, iż uzyskiwane są wielkie ilości danych i to na różnych poziomach organizacji świata żywego – od komórki do okrywy roślinnej. Możliwe stało się również połączenie danych z różnych poziomów funkcjonowania roślin, od genów do fenotypu – badania „multiomiczne". Jednocześnie wzrost liczby publicznie dostępnych zbiorów danych pozwala badaczom integrować je z danymi z własnych eksperymentów i dzięki temu uzyskiwać pewniejsze wyniki. Te wszystkie podejścia nie są możliwe bez komputerów o dużej mocy obliczeniowej i odpowiednich narzędzi bioinformatycznych. Obszary te rozwijają się bardzo dynamicznie, a w analizie wyników coraz większą rolę odgrywają sztuczna inteligencja (AI) i uczenie maszynowe (ML). AI i ML w szczególny sposób przydają się w badaniach fenomicznych, pozwalając zarówno zwiększyć automatyzację pomiarów, jak i ułatwić analizę olbrzymiej ilości danych pomiarowych. Należy się spodziewać, że przyrost ilości danych w biologii będzie coraz szybszy, a bioinformatyka będzie się rozwijała z coraz szerszym wykorzystaniem AI. W konsekwencji badania będą wykonywane z coraz większą efektywnością.

PIŚMIENNICTWO

1. Gauthier J, Vincent AT, Charette SJ, Derome N (2019) A brief history of bioinformatics. Brief Bioinform 20:1981–1996
2. Floyd BM, Marcotte EM (2022) Protein Sequencing, One Molecule at a Time. Annu Rev Biophys 51:181–200
3. Hogeweg P (2011) The Roots of Bioinformatics in Theoretical Biology. PLoS Comput Biol 7:e1002021
4. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH (2017) DNA sequencing at 40: past, present and future. Nature 550:345–353
5. Bartlett A, Lewis J, Williams ML (2016) Generations of interdisciplinarity in bioinformatics. New Genet Soc 35:186–209
6. Dunisławska A, Łachmańska J, Sławińska A, Siwek M (2017) Next generation sequencing in animal science – a review. Anim Sci Pap Rep 35:205–224
7. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815
8. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
9. Zhu H, Zhang H, Xu Y, Laššáková S, Korabečná M, Neužil P (2020) PCR Past, Present and Future. BioTechniques 69:317–325
10. Bączkowski K, Mackiewicz P, Kowalczuk M, Banaszak J, Cebrat S (2005) Od sekwencji do funkcji – poszukiwanie genów i ich adnotacje. Biotechnologia 3:22–44
11. Voichek Y, Hristova G, Mollá-Morales A, Weigel D, Nordborg M (2024) Widespread position-dependent transcriptional regulatory sequences in plants. Nat Genet 56:2238–2246
12. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y, Raytselis Y, Sayers EW, Tao T, Ye J, Zaretskaya I (2013) BLAST: a more efficient report with usability improvements. Nucleic Acids Res 41:W29–W33
13. The UniProt Consortium (2025) UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res 53:D609–D617
14. Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, Pinto BL, Orr A, Paysan-Lafosse T, Ponamareva I, Salazar GA, Bordin N, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Llinares-López F, Marchler-Bauer A, Meng-Papaxanthos L, Mi H, Natale DA, Orengo CA, Pandurangan AP, Piovesan D, Rivoire C, Sigrist CJA, Thanki N, Thibaud-Nissen F, Thomas PD, Tosatto SCE, Wu CH, Bateman A (2025) InterPro: the protein sequence classification resource in 2025. Nucleic Acids Res 53:D444–D456
15. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, Von Heijne G, Nielsen H (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol 40:1023–1025
16. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res 29:102–105
17. Monzó C, Liu T, Conesa A (2025) Transcriptomics in the era of long-read sequencing. Nat Rev Genet
18. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, Mardis ER, Sadar MD, Siddiqui AS, Marra MA, Jones SJ (2006) Analysis of the prostate

cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 7:246

19. Asano T, Masumura T, Kusano H, Kikuchi S, Kurita A, Shimada H, Kadowaki K (2002) Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: toward comprehensive analysis of the genes expressed in the rice phloem. The Plant Journal 32:401–408
20. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP, Sandberg R (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 30:777–782
21. Singh S, Praveen A, Dudha N, Sharma VK, Bhadrecha P (2024) Single-cell transcriptomics: a new frontier in plant biotechnology research. Plant Cell Rep 43, 294
22. Denyer T, Ma X, Klesen S, Scacchi E, Nieselt K, Timmermans MCP (2019) Spatiotemporal Developmental Trajectories in the Arabidopsis Root Revealed Using High-Throughput Single-Cell RNA Sequencing. Developmental Cell 48:840–852
23. Bushel PR, Ferguson SS, Ramaiahgari SC, Paules RS, Auerbach SS (2020) Comparison of Normalization Methods for Analysis of TempO-Seq Targeted RNA Sequencing Data. Front Genet 11:594
24. Lüleci HB, Uzuner D, Cesur MF, İlgün A, Düz E, Abdik E, Odongo R, Çakır T (2024) A benchmark of RNA-seq data normalization methods for transcriptome mapping on human genome-scale metabolic networks. Npj Syst Biol Appl 10:124
25. Sinha A, Mann M (2020) A beginner's guide to mass spectrometry-based proteomics. The Biochemist 42:64–69
26. Fochtman D, Wojakowska A, Marczak Ł (2024) Spektrometria mas z mobilnością jonów w badaniach multi-omicznych. Postępy Biochem 70:204–211
27. Vitorino R, Guedes S, Trindade F, Correia I, Moura G, Carvalho P, Santos MAS, Amado F (2020) *De novo* sequencing of proteins by mass spectrometry. Expert Rev Proteomics 17:595–607
28. Shuken SR (2023) An Introduction to Mass Spectrometry-Based Proteomics. J Proteome Res 22:2151–2171
29. Hao Y, Zhang Z, Luo E, Yang J, Wang S (2025) Plant metabolomics: applications and challenges in the era of multi-omics big data. aBIOTECH 6:116–132
30. Mascellani Bergo A, Leiss K, Havlik J (2024) Twenty Years of ^{1}H NMR Plant Metabolomics: A Way Forward toward Assessment of Plant Metabolites for Constitutive and Inducible Defenses to Biotic Stress. J Agric Food Chem 72:8332–8346
31. Wadie B, Stuart L, Rath CM, Drotleff B, Mamedov S, Alexandrov T (2024) METASPACE-ML: Context-specific metabolite annotation for imaging mass spectrometry using machine learning. Nat Commun 15:9110
32. Hummel J, Strehmel N, Bölling C, Schmidt S, Walther D, Kopka J (2013) Mass Spectral Search and Analysis Using the Golm Metabolome Database. W: Weckwerth W, Kahl G (red) The Handbook of Plant Metabolomics, 1st ed. Wiley, str. 321–343
33. Thirumurugan D, Cholarajan A, Raja SSS, Vijayakumar R (2018) An Introductory Chapter: Secondary Metabolites. W: Vijayakumar R, Raja SSS (red) Secondary Metabolites–Sources and Applications. IntechOpen
34. Cragg GM, Schepartz SA, Suffness M, Grever MR (1993) The Taxol Supply Crisis. New NCI Policies for Handling the Large-Scale Production of Novel Natural Product Anticancer and Anti-HIV Agents. J Nat Prod 56:1657–1668
35. Wu T, Kerbler SM, Fernie AR, Zhang Y (2021) Plant cell cultures as heterologous bio-factories for secondary metabolite production. Plant Commun 2:100235
36. Prashant SP, Bhawana M (2024) An update on biotechnological intervention mediated by plant tissue culture to boost secondary metabolite production in medicinal and aromatic plants. Physiol Plant 176:e14400
37. Selwal N, Rahayu F, Herwati A, Latifah E, Supriyono, Suhara C, Kade Suastika IB, Mahayu WM, Wani AK (2023) Enhancing secondary metabolite production in plants: Exploring traditional and modern strategies. J Agric Food Res 14:100702
38. Kaur H, Chahal S, Jha P, Lekhak MM, Shekhawat MS, Naidoo D, Arencibia AD, Ochatt SJ, Kumar V (2022) Harnessing plant biotechnology-based strategies for in vitro galanthamine (GAL) biosynthesis: a potent drug against Alzheimer's disease. Plant Cell Tissue Organ Cult PCTOC 149:81–103
39. DeMarini DM (2020) The mutagenesis moonshot: The propitious beginnings of the environmental mutagenesis and genomics society. Environ Mol Mutagen 61:8–24
40. Hamdan MF, Tan BC (2024) Genetic modification techniques in plant breeding: A comparative review of CRISPR/Cas and GM technologies. Hortic Plant J
41. Liao H, Wu J, VanDusen NJ, Li Y, Zheng Y (2024) CRISPR-Cas9-mediated homology-directed repair for precise gene editing. Mol Ther – Nucleic Acids 35:102344
42. Horvath P, Barrangou R (2010) CRISPR/Cas, the Immune System of Bacteria and Archaea. Science 327:167–170
43. Jiang C, Li Y, Wang R, Sun X, Zhang Y, Zhang Q (2024) Development and optimization of base editors and its application in crops. Biochem Biophys Res Commun 739:150942
44. Grützner R, Martin P, Horn C, Mortensen S, Cram EJ, Lee-Parsons CWT, Stuttmann J, Marillonnet S (2021) High-efficiency genome editing in plants mediated by a Cas9 gene containing multiple introns. Plant Commun 2:100135
45. Ajmal CPM, Keerthana S, Hemalatha N, Suhail KA (2025) Development of CRISPR-Cas9 Algorithm for Designing gRNAs for Polyploid Sugarcane. Sugar Tech
46. Kamens J (2014) Addgene: Making Materials Sharing "Science As Usual." PLoS Biol 12:e1001991
47. Yue J-J, Hong C-Y, Wei P, Tsai Y-C, Lin C-S (2020) How to start your monocot CRISPR/Cas project: plasmid design, efficiency detection, and offspring analysis. Rice 13, 9
48. Shende VV, Bauman KD, Moore BS (2024) The shikimate pathway: gateway to metabolic diversity. Nat Prod Rep 41:604–648
49. Yang D, Du X, Yang Z, Liang Z, Guo Z, Liu Y (2014) Transcriptomics, proteomics, and metabolomics to reveal mechanisms underlying plant secondary metabolism. Eng Life Sci 14:456–466
50. Mipeshwaree Devi A, Khedashwori Devi K, Premi Devi P, Lakshmipriyari Devi M, Das S (2023) Metabolic engineering of plant secondary metabolites: prospects and its technological challenges. Front Plant Sci 14:1171154
51. Gharat SA, Tamhane VA, Giri AP, Aharoni A (2025) Navigating the challenges of engineering composite specialized metabolite pathways in plants. The Plant Journal 121:e70100
52. Kaya C (2025) Optimizing Crop Production With Plant Phenomics Through High-Throughput Phenotyping and AI in Controlled Environments. Food Energy Secur 14:e70050
53. Ma Z, Rayhana R, Feng K, Liu Z, Xiao G, Ruan Y, Sangha JS (2022) A Review on Sensing Technologies for High-Throughput Plant Phenotyping. IEEE Open J Instrum Meas 1:1–21
54. Clauw P, Ellis TJ, Liu H-J, Sasaki E (2024) Beyond the Standard GWAS—A Guide for Plant Biologists. Plant Cell Physiol pcae079:1–13
55. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C (2023) The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 51:D638–D646
56. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang P-L, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2:2366–2382
57. Hawkins C, Xue B, Yasmin F, Wyatt G, Zerbe P, Rhee SY (2025) Plant Metabolic Network 16: expansion of underrepresented plant gro-

ups and experimentally supported enzyme data. Nucleic Acids Res 53:D1606–D1613

58. Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M (2025) KEGG: biological systems database as a model of the real world. Nucleic Acids Res 53:D672–D677
59. The Gene Ontology Consortium (2023) The Gene Ontology knowledgebase in 2023. Genetics 224:iyad031
60. Reinhardt JK, Craft D, Weng J-K (2025) Toward an integrated omics approach for plant biosynthetic pathway discovery in the age of AI. Trends Biochem Sci 50:311–321
61. Murphy KM, Ludwig E, Gutierrez J, Gehan MA (2024) Deep Learning in Image-Based Plant Phenotyping. Annu Rev Plant Biol 75:771–795
62. Zhu F, Wen W, Cheng Y, Alseekh S, Fernie AR (2023) Integrating multiomics data accelerates elucidation of plant primary and secondary metabolic pathways. aBIOTECH 4:47–56
63. Bono H (2020) All of gene expression (AOE): An integrated index for public gene expression databases. PLOS ONE 15:e0227076
64. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018
65. Ding Y, Virlouvet L, Liu N, Riethoven J-J, Fromm M, Avramova Z (2014) Dehydration stress memory genes of Zea mays; comparison with Arabidopsis thaliana. BMC Plant Biol 14:141
66. Wimalanathan K, Friedberg I, Andorf CM, Lawrence-Dill CJ (2018) Maize GO Annotation—Methods, Evaluation, and Review (maize-GAMER). Plant Direct 2:1–15
67. Schwacke R, Bolger ME, Usadel B (2025) PubPlant: a continuously updated online resource for sequenced and published plant genomes. BioRxiv 2025.03.12.642823

Omics methods in plant research

Maciej Jończyk✉

Department of Plant Molecular Ecophysiology, Institute of Experimental Plant Biology and Biotechnology, Faculty of Biology, University of Warsaw

✉corresponding author: m.jonczyk@uw.edu.pl

Keywords: plants, bioinformatics, genomics, transcriptomics, phenomics, CRISPR-Cas9

SUMMARY

The usefulness of computers in biological research has been identified already in the sixties of the twentieth century. Subsequent technical progress allowed development of high-throughput experimental methods in biology. Thanks to them, one can obtain a very large amount of data in a short time, even few days. However, with this progress came a necessity of development of data management, data quality control and analysis methods. Consequently, modern biology to a great extent rely on informatics and at the intersection of these fields emerged bioinformatics. The aim of this article is a review of modern high-throughput methods in plant science, and relevant bioinformatic tools.