KOSMOS
PROBLEMY NAUK BIOLOGICZNYCH

Tom 58 2009
Numer 3–4 (284–285)
Strony 485–498

Polskie Towarzystwo Przyrodników im. Kopernika

KRZYSZTOF SPALIK[1], MARCIN PIWCZYŃSKI[2]

[1]Zakład Systematyki i Geografii Roślin
Instytut Botaniki
Uniwersytet Warszawski
Aleje Ujazdowskie 4, 00-478 Warszawa
[2]Zakład Taksonomii i Geografii Roślin
Instytut Ekologii i Ochrony Środowiska
Uniwersytet Mikołaja Kopernika
Gagarina 9, 87-100 Toruń
E-mail: spalik@biol.uw.edu.pl
piwczyn@umk.pl

REKONSTRUKCJA FILOGENEZY I WNIOSKOWANIE FILOGENETYCZNE W BADANIACH EWOLUCYJNYCH

DARWIN, HAECKEL I FILOGENEZA

W lipcu 1837 r. Darwin naszkicował w swoim notatniku schematyczny graf relacji pokrewieństwa między gatunkami, obrazujący koncepcję drzewa życia. Ta idea, ubrana w daleko doskonalszą formę graficzną, pojawiła się także 22 lata później w jego rewolucyjnym dziele „O powstawaniu gatunków", ale wciąż jako koncept, a nie konkretne drzewo filogenetyczne, przedstawiające zależności ewolucyjne między gatunkami. Nie ma w tym nic dziwnego — Darwin zajmował się wyjaśnianiem mechanizmów ewolucji, nie zaś odtwarzaniem jej przebiegu. Pierwsze drzewo filogenetyczne pojawiło się w „Generelle Morphologie der Organismen" Ernsta Heackla w 1866 r. i tę datę można przyjąć jako oficjalny początek filogenetyki — gałęzi biologii zajmującej się rekonstrukcją filogenezy organizmów.

Drzewo filogenetyczne Haeckla było zapisem poglądów tego wybitnego uczonego na pochodzenie organizmów i podsumowaniem ówczesnego stanu wiedzy. Jednocześnie było hipotezą naukową, podlegającą weryfikacji w toku dalszych badań. Jeśli popatrzymy na zapisane na nim zależności filogenetyczne, to okaże się, że niewiele z nich zostało poprawnie odtworzonych, a obecny obraz drzewa życia jest zasadniczo odmienny. Jednak drzewo filogenetyczne nadal pozostaje jednocześnie podsumowaniem obecnego stanu wiedzy oraz hipotezą badawczą. W filogenetyce, podobnie jak w każdej nauce przyrodniczej, nie ma prawd absolutnych, a teorie i hipotezy uznajemy za prawdziwe, jeśli nikomu, mimo usilnych prób, nie udało się ich obalić. Warto pamiętać o tym zakresie niepewności, jaki towarzyszy wszystkim prawdom naukowym, a zwłaszcza dotyczącym odtwarzania przeszłości.

W tym artykule chcielibyśmy skupić się na metodyce badań filogenetycznych, a także pokazać, w jaki sposób otrzymane filogenezy służą wnioskowaniu ewolucyjnemu. Chcemy pokazać, że analizy filogenetyczne bazujące na danych molekularnych, aczkolwiek obarczone, jak w wypadku wszystkich nauk historycznych, nieuniknioną niepewnością, wsparte są na solidnych podstawach naukowych, a wypływające z nich wnioski nie są gorszej jakości od wniosków z badań eksperymentalnych.

DWA PODEJŚCIA DO KONSTRUOWANIA DRZEWA

Istnieje zasadnicza różnica metodologiczna między drzewem Haeckla a współczesnymi przedstawieniami filogenezy. To pierwsze było po prostu wyrazem poglądów badacza, wspartych wprawdzie rzetelną wiedzą i wynikającą z niej intuicją, ale nie powstało ono w wyniku żadnych określonych procedur. Współczesne drzewa są natomiast usyskiwane za pomocą określonych algorytmów obliczeniowych. Wyniki są zobiektywizowane i powtarzalne, a tym samym weryfikowalne. Ten przełom w filogenetyce został dokonany dzięki rozwojowi metod numerycznych oraz wynalezieniu komputerów, a przyniósł go nurt taksonomii zwany fenetyką, której „ojcami-założycielami" byli P. H. A. Sneath i R. R. Sokal. Jest paradoksem, że fenetyka jednocześnie odrzuciła biologiczny sens odtwarzania drzewa życia, a skoncentrowała się na konstruowaniu zależności wszechstronnego podobieństwa między organizmami, zakładając, że drzewo filogenetyczne wyjdzie mimochodem. Nie czyniła ona żadnych założeń o przydatności cech, traktując je równocennie. Odmienne podejście prezentowała kladystyka, której prekursorem był Willi Hennig. Ostro krytykowała ona podejście fenetyczne wskazując, że o pochodzeniu od wspólnego przodka świadczą jedynie wspólne unikatowe cechy pochodne ewolucyjnie, czyli synapomorfie, nie zaś cechy homoplastyczne: pierwotne ewolucyjnie, odziedziczone po odległym przodku (symplezjomorfie) albo powstałe niezależnie (parallelizmy). Rozróżnia się synapomorfie od homoplazji posługując się zasadą parsymonii (oszczędności), czyli wybierając spośród wszystkich możliwych drzew takie, które wyjaśnia różnorodność cech na liściach drzewa za pomocą najmniejszej liczby zmian na gałęziach, minimalizując tym samym konflikty cech. Spór między fenetyką a kladystyką był niezwykle gwałtowny — dziś emocje opadły, a w efekcie status obywatelstwa we współczesnej filogenetyce zyskały sobie koncepcje z obu nurtów. Nikt dzisiaj nie kwestionuje, że nadrzędnym problemem badawczym jest rekonstrukcja filogenezy, ale ten cel jest osiągany również za pomocą metod bazujących na podobieństwie.

REWOLUCJA MOLEKULARNA W FILOGENETYCE

Biologia molekularna, a przede wszystkim rozwój metod łańcuchowej reakcji polimerazy (PCR) oraz sekwencjonowania DNA, zrewolucjonizowała odtwarzanie drzewa rodowego organizmów. Dane z sekwencji okazały się daleko lepszymi znacznikami dla rekonstrukcji filogenezy niż tradycyjne cechy morfologiczne, anatomiczne czy biochemiczne. Składa się na to kilka powodów. Przede wszystkim, dane z sekwencji są genetyczne — przedstawiają nam od razu zapis informacji w DNA (lub RNA), podczas gdy dane z budowy organizmów mówią nam o tym zapisie pośrednio. Co gorsza, fenotyp organizmu jest wypadkową informacji genetycznej oraz jego interakcji ze środowiskiem zewnętrznym, a określona cecha morfologiczna może być determinowana przez jeden albo przez wiele loci. Wnioskowanie o podłożu genetycznym określonej cechy morfologicznej na podstawie jej zmienności jest zatem obarczone dużym błędem. Co więcej, aby taką cechę wykorzystać w analizie komputerowej, musimy jej zmienność zakodować, czyli przedstawić w formie liczb lub znaków, a sposób tego kodowania jest z konieczności arbitralny — natomiast dane z sekwencji nie wymagają kodowania, ponieważ są już zapisane jako ciąg znaków.

Sekwencje DNA dają nam też niezwykłą możliwość porównywania ze sobą bardzo odległych ewolucyjnie organizmów. Przykładowo — trudno na podstawie morfologii czy anatomii szacować odległość ewolucyjną między człowiekiem a bakterią *Escherichia coli*, ich budowa jest bowiem zbyt odmienna i trudno wskazać jakiekolwiek porównywalne cechy. Mają one jednak wiele podobnych genów, np. loci, w których są zapisane sekwencje rybosomalnego DNA. Dzięki takim genom możliwe jest stworzenie kompletnego drzewa życia.

Dla odtwarzania filogenezy nie bez znaczenia jest sposób, w jaki utrwaliły się analizowane zmiany cech (mutacje). Procesy prowadzące do rozpowszechnienia się mutacji możemy podzielić na dwa rodzaje: deterministyczne oraz stochastyczne (losowe). Procesem deterministycznym jest dobór naturalny — mutacje korzystne zwiększają swój udział w puli genowej, natomiast niekorzystne są z niej eliminowane (patrz rozdział ŁOMNICKIE-

GO *Dobór naturalny* w tym zeszycie KOSMOSU). Dobór naturalny jest architektem ewolucji, odpowiedzialnym za różnorodność organizmów. Paradoksalnie jednak, cechy utrwalone wskutek działania doboru mogą być zawodne w odtwarzaniu przebiegu ewolucji, silny nacisk selekcyjny sprzyja bowiem zmianom homoplastycznym — konwergencjom. W filogenetyce bardziej przydatne są cechy, które utrwaliły się przypadkowo, jest bowiem mało prawdopodobne, że taka sama przypadkowa zmiana utrwali się ponowne. Gdzie takich cech szukać? Fenotyp organizmu podlega silnej presji środowiska, a zatem zdecydowana większość cech fenotypowych musiała przejść przez sito doboru. Inaczej jest na poziomie genetycznym. Kiedy poznano sekwencje genów, zauważono dużą liczbę mutacji milczących, czyli niezmieniających sekwencji kodowanego białka; jeszcze więcej mutacji stwierdzono w sekwencjach niekodujących, np. w przestrzeniach międzygenowych albo w intronach. Spostrzeżenia te zaowocowały sformułowaniem neutralnej teorii ewolucji, której autorem był japoński badacz Motoo Kimura (patrz też artykuł ŁOMNICKIEGO *Dryf genetyczny* w tym zeszycie KOSMOSU). Zakłada ona, że większość substytucji (mutacji punktowych) jest neutralna lub prawie neutralna dla organizmu oraz że ich utrwalenie w populacji jest procesem przypadkowym. Ponieważ procesy powstawania i utrwalania mutacji są stochastyczne, to różnice między sekwencjami tego samego odcinka DNA u różnych organizmów są funkcją czasu, jaki upłynął od rozejścia się prowadzących do nich linii filogenetycznych. Umożliwia to nie tylko samo oszacowanie filogenezy, ale także — przy spełnieniu dodatkowych warunków — na opisanie tej filogenezy za pomocą skali czasu (patrz artykuł JERZMANOWSKIEGO w tym zeszycie KOSMOSU).

FILOGENETYKA MOLEKULARNA A TRADYCYJNA TAKSONOMIA

„Inwazja" metod molekularnych do taksonomii oraz tradycyjnej filogenetyki bazującej na cechach morfologicznych nie odbyła się bez oporów. Wnioski płynące z badań molekularnych były rewolucyjne, obalały bowiem wiele głęboko zakorzenionych poglądów na relacje pokrewieństwa między organizmami. Niekiedy tradycyjnym taksonomom trudno było się pogodzić z tymi wnioskami, a także z tym, że badania molekularne w krótkim czasie dały odpowiedź na pytania, nad którymi oni biedzili się przez całe życie. Nieufność do wyników badań molekularnych pogłębiały błędne oznaczenia gatunków w niektórych analizach (biolodzy molekularni nie zadali sobie trudu zweryfikowania użytego do badań materiału) oraz niestabilność kladów (gałęzi drzewa) spowodowana niedostatecznym próbkowaniem taksonomicznym (liczba taksonów) i genetycznym (reprezentatywność i długość sekwencji). Ponadto, dało się zauważyć pewną nonszalancję taksonomów molekularnych, połączoną z naiwną wiarą, że drzewo molekularne jest odpowiedzią na wszystkie pytania. Wkrótce jednak okazało się, że drzewo molekularne jest nie tyle końcem, co początkiem badań — trzeba bowiem je zinterpretować i sprawdzić, czy istotnie odpowiada na jakiekolwiek pytania ewolucyjne. Dziś już oba nurty — molekularny i morfologiczny — zgodnie koegzystują w taksonomii i biologii ewolucyjnej, korzystając wzajemnie z uzupełniających się kompetencji. Do absolutnych wyjątków należy kwestionowanie wyników badań molekularnych, jak to ostatnio uczynili GREHAN i SCHWARTZ (2009), postulując na podstawie zaledwie kilkudziesięciu cech morfologicznych, a wbrew badaniom molekularnym, że najbliższym krewnym człowieka jest orangutan, a nie szympans. Ich krytyka filogenetyki molekularnej jest naiwna i świadczy o podstawowych brakach w wiedzy — odrzucają oni bowiem wyniki analiz molekularnych twierdząc, że podobieństwo molekularne nie świadczy o pokrewieństwie, ustalenie homologii jest wątpliwe, a morfologia jest bardziej stabilna ewolucyjnie. Absolutne zdumienie budzi fakt, że artykuł ten został opublikowany w bardzo prestiżowym czasopiśmie, jakim jest Journal of Biogeography. Jednak towarzyszący mu komentarz od redakcji świadczy, że głównym powodem publikacji była raczej „polityczna poprawność" — oddanie głosu zanikającej mniejszości — a sami redaktorzy mają świadomość, iż dla każdego biologa molekularnego albo taksonoma lub antropologa choćby nieco obeznanego z filogenetyką molekularną wnioski autorów są nonsensowne. Filogenetyka molekularna to jednak coś więcej niż prosta analiza podobieństwa molekularnego, co oczywiście nie znaczy, że wnioskowanie filogenetyczne na podstawie danych molekularnych jest zawsze bezbłędne i nieobarczone niepewnością. Warto sobie uświadomić źródła tej niepewności.

HOMOLOGIA SEKWENCJI I SYGNAŁ FILOGENETYCZNY

Porównując te same sekwencje DNA otrzymane od osobników z różnych populacji lub z różnych gatunków możemy oczekiwać, że bliżej spokrewnione będą osobniki (gatunki), które różnią się mniejszą liczbą mutacji. Czy zatem wnioskowanie o pokrewieństwach między organizmami jest prostym zabiegiem polegającym na porównaniu sekwencji i obliczeniu liczby różniących je podstawień? Sytuacja nie jest tak prosta, a droga do odtworzenia filogenezy jest pełna pułapek. Po pierwsze, sekwencje wybrane do analizy powinny być homologiczne, czyli pochodzące od wspólnego przodka. Homologia na poziomie sekwencji ma jednak dwojakie oblicze. Sekwencje ortologiczne zajmują ten sam locus i ewoluują niezależnie od czasu rozejścia się linii filogenetycznych, czyli od specjacji. To one niosą sygnał filogenetyczny — zapis historii ewolucyjnej danej linii ewolucyjnej. W trakcie ewolucji regularnie występują jednak także duplikacje loci (patrz artykuł JERZMANOWSKIEGO w tym zeszycie KOSMOSU), w wyniku czego powstają sekwencje paralogiczne. Pomieszanie sekwencji ortologicznych i paralogicznych uniemożliwia odtworzenie prawidłowej filogenezy, ponieważ sekwencje paralogiczne ewoluują niezależnie od momentu duplikacji locus, a nie od rozejścia się linii filogenetycznych.

Wybór sekwencji ortologicznych nie gwarantuje jednak, że informacja o ich historii ewolucyjnej jest niezaburzona. Procesami, które powodują, że sekwencje są do siebie bardziej podobne, niżby to wynikało z czasu, który upłynął od ich rozejścia się, są:

— mutacje wsteczne (rewersje), czyli powrót do nukleotydu występującego w sekwencji u wspólnego przodka;

— wielokrotne podstawienia, czyli kilkukrotne zamiany nukleotydów w tym samym miejscu, wskutek czego obserwujemy mniej podstawień, niż ich w rzeczywistości było;

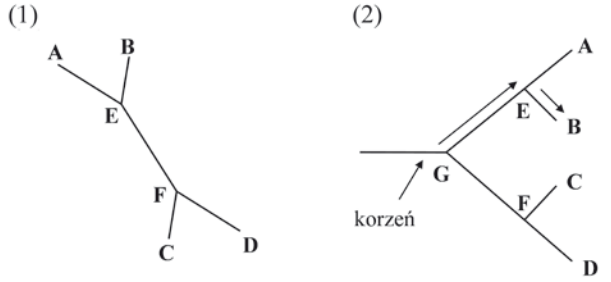
— podstawienia równoległe, czyli niezależne podstawienia w tej samej pozycji przez ten sam nukleotyd w obu porównywanych sekwencjach.

Wszystkie te procesy zaburzają liniową zależność między czasem rozejścia się organizmów a liczbą obserwowanych mutacji oraz zacierają sygnał filogenetyczny, czyli mutacje synapomorficzne, dzięki którym można zidentyfikować pokrewieństwo gatunków.

Bardzo istotnym problemem jest także zidentyfikowanie homologicznych pozycji w sekwencji, czyli dokonanie ich przyrównania. Nie zawsze jest to zadanie łatwe, ponieważ w trakcie ewolucji zachodzą nie tylko podstawienia nukleotydów, ale także ich insercje (wstawienia) i delecje (usunięcia). W wypadku sekwencji kodujących białka insercje i delecje są zazwyczaj usuwane przez dobór oczyszczający, albowiem wstawienie bądź usunięcie jednego lub dwóch nukleotydów zmienia odczyt, wskutek czego białko przestaje być funkcjonalne. Jedynie wstawienia trzech (albo wielokrotności trzech) nukleotydów mają szansę na przejście przez sito doboru. Natomiast w sekwencjach niekodujących, np. w intronach lub przestrzeniach międzygenowych, delecje i insercje zdarzają się często. Proces przyrównywania sekwencji jest kluczowy do właściwego oszacowania pokrewieństw między organizmami żywymi i obecnie istnieje wiele algorytmów umożliwiających dokonanie takiego przyrównania.

UKORZENIANIE DRZEWA

Przyjrzyjmy się strukturze drzewa filogenetycznego jako zapisowi relacji pokrewieństwa ewolucyjnego między organizmami. Drzewo to zbudowane jest z węzłów — zewnętrznych i wewnętrznych — i łączących je gałęzi (Ryc. 1). W drzewie w pełni rozwiązanym każdy węzeł wewnętrzny połączony jest z innymi węzłami za pomocą trzech gałęzi, zaś do węzłów zewnętrznych prowadzi tylko jedna. W drzewie nie w pełni rozwiązanym gałęzi wychodzących z jednego węzła może być więcej, czyli występują politomie. Węzły zewnętrzne nazywamy inaczej liśćmi; każdy z nich odpowiada badanemu organizmowi. Natomiast węzły wewnętrzne można przypisać hipotetycznym wspólnym przodkom określonych konarów (kladów) drzewa. Drzewo zrekonstruowane metodami filogenetycznymi jest zazwyczaj drzewem niezakorzenionym, a więc takim, w którym nieznany jest kierunek ewolucji. Innymi słowy, nie wiemy, która z tych trzech gałęzi wchodzi do dane-



Rycina 1. Struktura drzewa niezakorzenionego (1) i zakorzenionego (2).

Oba drzewa mają tę samą topologię. A, B, C, D oznaczają liście, czyli węzły zewnętrzne drzewa, zaś E, F i G — węzły wewnętrzne, przy czym drzewo (2) jest ukorzenione w węźle G. Strzałki przy drzewie zakorzenionym wskazują kierunek ewolucji, w przeciwieństwie do drzewa niezakorzenionego, w którym kierunek ten jest nieznany.

go węzła, a które z niego wychodzą (Ryc. 1). Ukorzenienie polega na dodaniu dodatkowego węzła na jednej z gałęzi, tożsamego ze wspólnym przodkiem wszystkich badanych organizmów. Innymi słowy, łamiemy tę gałąź na dwie oraz do miejsca złamania (węzła) dołączamy korzeń drzewa. Po ukorzenieniu drzewa możemy zauważyć, że zmienia się status gałęzi w węzłach. W drzewie niezakorzenionym wszystkie trzy gałęzie zbiegające się w węźle wewnętrznym są równocenne, natomiast w drzewie zakorzenionym jedna z nich jest gałęzią wchodzącą, a dwie wychodzącymi. Grupy wywodzące się z jednego węzła nazywamy siostrzanymi. Warto zauważyć, że bez ukorzenienia drzewa nie możemy wyciągać żadnych sensownych wniosków o ewolucji badanej grupy, np. o monofiletyzmie określonych taksonów albo o kierunku zmian morfologicznych.

Najlepszym sposobem ukorzenienia drzewa jest uwzględnienie w analizie filogenetycznej nie tylko badanej grupy, ale także jej najbliższych krewnych, czyli tzw. grupy zewnętrznej. Odszukujemy na uzyskanym drzewie wspólny wewnętrzny węzeł dla badanej grupy i wspólny wewnętrzny węzeł dla grupy zewnętrznej, a następnie przełamujemy łączącą je gałąź. Przykładowo, wiemy z innych badań, że grupą siostrzaną roślin okrytozalążkowych są nagozalążkowe. Rekonstruując filogenezę okrytozalążkowych, wybieramy więc sekwencje jednego lub kilku nagozalążkowych, np. sosny, sagowca, miłorzębu lub welwiczji, jako przedstawicieli grupy zewnętrznej. Następnie na drzewie odszukujemy gałąź łączącą okryto- i nagozalążkowe i przełamujemy ją, dodając korzeń. Dzięki takiemu zabiegowi jesteśmy w stanie stwierdzić, w jakiej kolejności oddzielały się poszczególne linie rodowe okrytozalążkowych i która grupa współczesnych gatunków jest filogenetycznie najstarsza. Ukorzenianie drzewa za pomocą grupy zewnętrznej jest standardową procedurą w badaniach filogenetycznych. Warto jednak zauważyć, że błędne wybranie grupy zewnętrznej, a tym samym błędne zakorzenienie, sprawia, że błędnie odczytujemy na nim kierunek ewolucji.

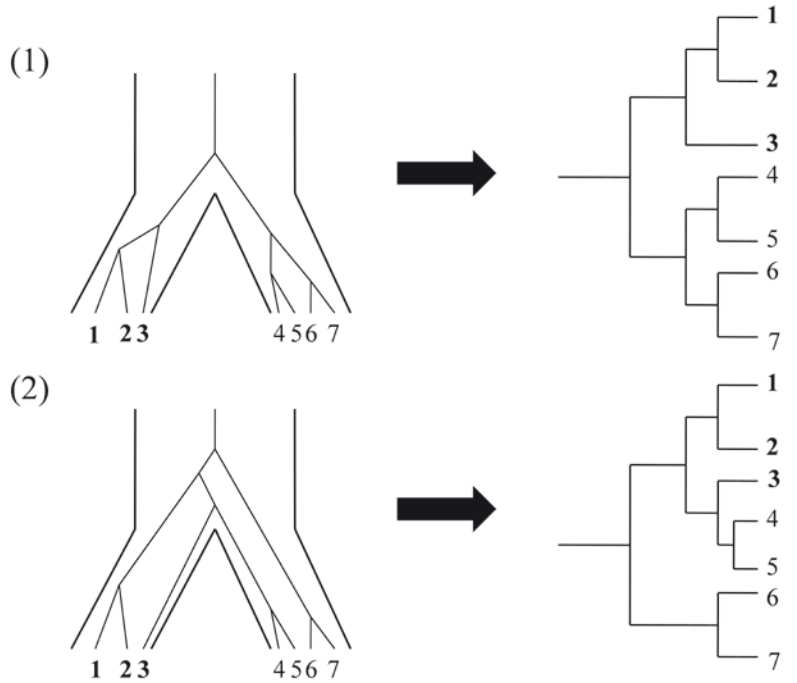
Wszystkie metody szacowania filogenezy dają możliwość obliczenia długości gałęzi łączących poszczególne węzły. Jeśli długość gałęzi jest proporcjonalna do liczby mutacji (obserwowanej lub oszacowanej), które zaszły między węzłami, to takie drzewo nazywamy filogramem. Natomiast jeśli długość gałęzi odpowiada czasowi względnemu lub absolutnemu, wtedy mówimy o chronogramie. Czasami interesuje nas tylko topologia drzewa (wzór rozgałęzień), a długość gałęzi jest nieistotna — takie drzewo nazywamy kladogramem.

DRZEWO GATUNKÓW I DRZEWO GENÓW

Warto zwrócić uwagę, że drzewo filogenetyczne odzwierciedla relację między przyrównanymi sekwencjami, dlatego też jest ono zawsze drzewem genów. Nie zawsze historia ewolucyjna genów odpowiada historii gatunków. Mechanizmów prowadzących do takich niezgodności jest kilka. Najważniejsze to:

— rekombinacja genów paralogicznych lub rekrutacja pseudogenów, wskutek czego powstały rekombinowany allel zawiera sygnał filogenetyczny z dwóch lub więcej loci;

— horyzontalny przepływ genów, czyli „przeskoczenie" materiału genetycznego z jednej linii filogenetycznej do drugiej; zjawisko to jest powszechne u bakterii, ale stosunkowo rzadkie wśród eukariotów, choć u roślin kwiatowych, szczególnie w genomie mitochondrialnym, znaleziono geny pocho-



Rycina 2. Konstrukcja genealogii genu dla dwóch gatunków.

W pierwszym przypadku (1) drzewo gatunków jest identyczne z drzewem genów, zaś w drugim (2) część alleli nie jest całkowicie posortowana. W tym przypadku drzewo genów nie jest zgodne z drzewem gatunków. Sytuacja ta występuje szczególnie u gatunków, u których specjacja zaszła stosunkowo niedawno, a liczba alleli danego genu przed rozejściem się była wysoka.

dzące od bakterii, mszaków lub innych roślin kwiatowych, zwłaszcza pasożytniczych;

— niepełne sortowanie linii genealogicznych po rozejściu się puli genowych; ponieważ proces rozdziału alleli w trakcie specjacji jest losowy, może się zdarzyć, że do jednej puli trafią dwa odległe genealogicznie allele, bliższe allelom z drugiej puli, a nie sobie nawzajem (Ryc. 2);

— silny dobór premiujący polimorfizm alleli w loci, którego najlepszym przykładem są allele genów głównego układu zgodności tkankowej; przykładowo, wszystkie naczelne odziedziczyły podobny polimorfizm alleli tego układu po wspólnym przodku, a tym samym w puli genowej człowieka znajdują się allele, które są bliżej spokrewnione z odpowiednimi allelami występującymi u szympansów niż z innymi allelami u człowieka; zauważmy, że efekt takiego doboru jest podobny, jak w wypadku niepełnego sortowania linii genealogicznych, inne są jednak przyczyny obu zjawisk — stochastyczne w wypadku sortowania linii genealogicznych i deterministyczne w wypadku selekcji faworyzującej polimorfizm;

— hybrydyzacja i introgresja[1], wskutek czego zależności międzygatunkowe opisywane są raczej za pomocą topologii sieci[2], a nie drzewa; zjawisko hybrydyzacji wydaje się stosunkowo częste u roślin, zwłaszcza okrytozalążkowych, wśród których spotykamy wiele allopoliploidów[3], powstałych właśnie wskutek hybrydyzacji.

ODTWARZANIE DRZEWA

Rekonstrukcja drzewa filogenetycznego jest złożonym zagadnieniem statystycznym i algorytmicznym. Istnieje wiele metod rekonstrukcji filogenezy, odwołujących się do różnych założeń statystycznych i biologicznych. Warto więc wykonywać analizę filogenetyczną za pomocą różnych narzędzi, a następnie porównać wyniki i szukać przyczyn ewentualnych rozbieżności między nimi.

Wyróżniamy cztery podstawowe metody rekonstrukcji filogenezy:

— największej parsymonii (ang. Maximum Parsimony, MP),

[1]Introgresja to krzyżowanie się mieszańca międzygatunkowego z jednym z gatunków rodzicielskich, wskutek czego dochodzi do przepływu genów z jednej puli genowej do drugiej.

[2]Sieć, w przeciwieństwie do drzewa, charakteryzuje się występowaniem tzw. cykli, czyli zamkniętych ścieżek łączących poszczególne węzły.

[3]Wiele gatunków roślin powstało poprzez hybrydyzację, a następnie poliploidyzację, która przywróciła homologie między chromosomami (patrz artykuł SZYMURY w tym zeszycie KOSMOSU).

— odległościowe (np. ang. Neighbour-Joining, NJ),
— największej wiarygodności (ang. Maximum Likelihood, ML)
— bayesowskie (ang. Bayesian Phylogenetics, BP).

Trzy ostatnie grupy metod bazują na modelach substytucji nukleotydów.

METODA NAJWIĘKSZEJ PARSYMONII

Metoda największej parsymonii jest jedną z najwcześniej zaproponowanych procedur rekonstrukcji filogenezy (CAMIN i SOKAL 1965). Polega ona na poszukiwaniu w przestrzeni wszystkich możliwych drzew takiego, które najoszczędniej tłumaczy obserwowaną zmienność cech na liściach drzewa. W tym celu odtwarza się stany poszczególnych cech w węzłach wewnętrznych drzewa, przyporządkowując jednocześnie zmiany stanów gałęziom, czyli mapując je na gałęziach. Przykładowo, jeśli w dwóch sekwencjach siostrzanych występuje nukleotyd A, to według kryterium parsymonii ich wspólny przodek ma także adeninę w tej pozycji, ponieważ taki układ nie wymaga żadnej zmiany na gałęziach. Gdyby była tam cytozyna (albo jakikolwiek inny nukleotyd), to musielibyśmy założyć, że na obu gałęziach nastąpiło podstawienie cytozyny przez adeninę. Suma wszystkich zmian dla każdego miejsca w przyrównanych sekwencjach buduje długość drzewa. Zgodnie z kryterium parsymonii, drzewo najkrótsze uważane jest za najlepsze.

Mimo swojej prostoty, metoda największej parsymonii w pewnych sytuacjach zawodzi. Wykazano, że w wypadkach silnie zróżnicowanego tempa ewolucji w poszczególnych gałęziach i intensywnej radiacji (krótkich odcinków czasu między rozgałęzieniami drzewa), metoda MP jest wrażliwa na homoplazje — interpretuje je jako synapomorfie. Takich fałszywych synapomorfii jest więcej na długich gałęziach (wykazujących szybsze tempo podstawiania nukleotydów), a zatem takie gałęzie są mylnie łączone. Zjawisko to nazwano „efektem przyciągania się długich gałęzi". Pomimo tej krytyki metoda MP pozostaje silnym narzędziem do wnioskowania filogenetycznego, szczególnie na niskim poziomie zmienności sekwencji, głównie ze względu na niewielkie wymagania obliczeniowe oraz dość dobrze zbadane właściwości, w przeciwieństwie to tak modnej obecnie analizy bayesowskiej (patrz niżej).

MODELE SUBSTYTUCJI NUKLEOTYDÓW

Sposobem na uniknięcie efektu przyciągania się długich gałęzi jest uwzględnienie w szacowaniu filogenezy całkowitej liczby zmian, które na danej gałęzi zaszły, uwzględniając podstawienia wielokrotne i rewersje. Wymaga to zastosowania określonego modelu ewolucji DNA, czyli modelu substytucji nukleotydów. Z modeli takich korzystają metody odległościowe, największej wiarygodności oraz bayesowska.

Ewolucję sekwencji nukleotydowych można przedstawić w postaci modeli matematycznych, które mają uzasadnienie biologiczne oraz są możliwe do implementacji algorytmicznej. Od czasu publikacji pierwszego modelu JUKESA i CANTORA (1969), zakładającego jednakowe prawdopodobieństwo substytucji między wszystkimi czterema nukleotydami, opisano wiele modeli, które odchodzą od tych mało realistycznych założeń. Doprowadziło to w konsekwencji do stworzenia kilkudziesięciu modeli ewolucji DNA. Najbardziej złożony model – GTR+I+Γ [ang. General Time Reversible + Invariant (positions) + Gamma (distribution)] posiada 12 wolnych parametrów. Dziesięć z nich pozwala na przyporządkowanie różnego prawdopodobieństwa podstawienia jednego nukleotydu drugim (przy czym prawdopodobieństwa substytucji np. A $\rightarrow$ T i T $\rightarrow$ A są identyczne, a więc macierz podstawień nukleotydów jest symetryczna) oraz określenie frekwencji poszczególnych nukleotydów. Pozostałe dwa parametry pozwalają na wprowadzenie do modelu procentu miejsc niezmiennych (I) oraz zróżnicowanego tempa substytucji w różnych częściach danej sekwencji, opisanego za pomocą rozkładu *gamma* (Γ). Wiele modeli można wyprowadzić z GTR poprzez uproszczenie jego założeń. Duża liczba modeli o różnej liczbie parametrów umożliwia matematyczny opis sekwencji pełniących róż-

norodne role w genomie. Warto wspomnieć, że istnieją także inne modele ewolucji, które wykorzystywane są do rekonstrukcji filogenezy na podstawie sekwencji specyficznych cząsteczek, takich jak RNA czy białka.

W celu zobiektywizowania procesu wyboru odpowiedniego modelu substytucji, wykorzystuje się kilka metod: LRT (ang. Likelihood-Ratio Test), AIC (ang. Akaike Information Criterion), BIC (ang. Bayesian Information Criterion). Wszystkie one pozwalają na wybranie najprostszego modelu dobrze opisującego analizowane dane. Procedura ta jest standardowo wykonywana przed użyciem metody filogenetycznej, która wymaga modelu ewolucji.

METODY ODLEGŁOŚCIOWE

Szacowanie filogenezy metodami odległościowymi wymaga dwóch kroków: obliczenia odległości genetycznej pomiędzy parami sekwencji, a następnie rekonstrukcji drzewa na podstawie macierzy odległości za pomocą określonego algorytmu. Najczęściej stosowaną metodą odległościową jest metoda łączenia sąsiadów (ang. Neighbour-Joining, NJ). Jedną z podstawowych zalet tej techniki jest jej szybkość obliczeniowa, nawet dla setek przyrównanych sekwencji. Uzyskujemy jednak tylko jedno drzewo, podczas gdy może istnieć wiele innych, równie dobrych drzew (o równie prawdopodobnej topologii). Dlatego też wykorzystanie tej metody jest ograniczone głównie do szybkiego oszacowania suboptymalnej zazwyczaj filogenezy. Służy ona do zgrubnej analizy danych, znajduje też zastosowanie do obliczenia wartości funkcji wiarygodności w procedurze wyboru modelu substytucji (np. w programie ModelTest) albo dostarcza drzewa stanowiącego punkt startowy do dalszych przeszukiwań (np. w metodzie maksymalnej wiarygodności).

METODA NAJWIĘKSZEJ WIARYGODNOŚCI

Stosowana powszechnie w statystyce metoda największej wiarygodności pomaga oszacować prawdopodobieństwo obserwowanych danych (w naszym przypadku przyrównanych sekwencji), kiedy parametry modelu są znane. Zmieniając wartości parametrów możemy znaleźć taki ich zbiór, który daje nam najwyższą wiarygodność opisu naszych danych — innymi słowy, poszukujemy parametrów, dla których funkcja wiarygodności osiąga maksimum. W przypadku rekonstrukcji drzew filogenetycznych poszukiwanymi wartościami są topologia drzewa filogenetycznego oraz parametry wybranego modelu ewolucji DNA, niezbędne dla oszacowania długości gałęzi. Drzewo o najwyższej wartości funkcji wiarygodności uważane jest za najlepsze. Jednym z podstawowych argumentów za użyciem tej metody jest możliwość elastycznego wprowadzania różnych założeń w postaci parametrów oraz znane własności statystyczne. Problemem jest jednak obliczeniowa czasochłonność. Spowodowane jest to dużą liczbą parametrów do optymalizacji oraz ogromną liczbą możliwych drzew do sprawdzenia.

METODA BAYESOWSKA

Metoda bayesowska stała się obecnie najczęściej stosowaną techniką rekonstrukcji drzew filogenetycznych. Aby zrozumieć zasady leżące u podstaw tej metody, należy poznać dwa wzory z rachunku prawdopodobieństwa: wzór na prawdopodobieństwo całkowite i wzór Bayesa. Warto tutaj posłużyć się przykładem niezwiązanym z filogenetyką. Wyobraźmy sobie dwie urny, jedna zawiera 4 białe kule i jedną czarną, zaś druga 2 białe i 3 czarne. Wiemy także, że szansa wylosowania pierwszej urny równa się 2/3, zaś urny drugiej 1/3. Jakie jest prawdopodobieństwo wylosowania kuli białej? Jak widać, mamy tutaj dwie tury losowań, pierwsza dotyczy wylosowania urny, a druga losowania kuli. Oznaczmy zdarzenie wylosowania kuli białej literą A, natomiast wybór urny – literą H. Zdarzenie H jest rozbite na dwa wykluczające się zdarzenia – wybór urny pierwszej (H_1) lub wybór urny drugiej (H_2). Na wartość prawdopodobieństwa wyboru kuli białej składać się

będzie prawdopodobieństwo wylosowania kuli białej z pierwszej urny $P(A|H_1)$ ważone przez prawdopodobieństwo wylosowania tej urny $P(H_1)$ oraz prawdopodobieństwo wylosowania kuli z drugiej urny $P(A|H_2)$ ważone przez prawdopodobieństwo wyboru tej urny $P(H_2)$. Uogólniając na dowolną liczbę wykluczających się zdarzeń H_i, uzyskujemy wzór na prawdopodobieństwo całkowite:

$$P(A) = \sum P(A|H_i)P(H_i).$$

Prawdopodobieństwo całkowite obliczamy wtedy, kiedy znamy procedurę doświadczenia i pytamy o jego najbardziej prawdopodobny wynik. Możemy jednak problem odwrócić — znamy wynik doświadczenia, a chcemy zapytać o jego przebieg. Przykładowo, wiemy, że została wylosowana kula biała. Jakie jest prawdopodobieństwo, że wylosowano ją z pierwszej urny, czyli jakie jest prawdopodobieństwo zdarzenia H_1, jeśli wiemy że zaszło A? Prawdopodobieństwo $P(H_1|A)$ jest iloczynem prawdopodobieństwa wyboru pierwszej urny $P(H_1)$ i wylosowania kuli białej z tej urny $P(A|H_1)$, podzielonym przez prawdopodobieństwo całkowite wylosowania kuli białej. Uogólniając dla dowolnej liczby zdarzeń, prawdopodobieństwo to można zapisać jako

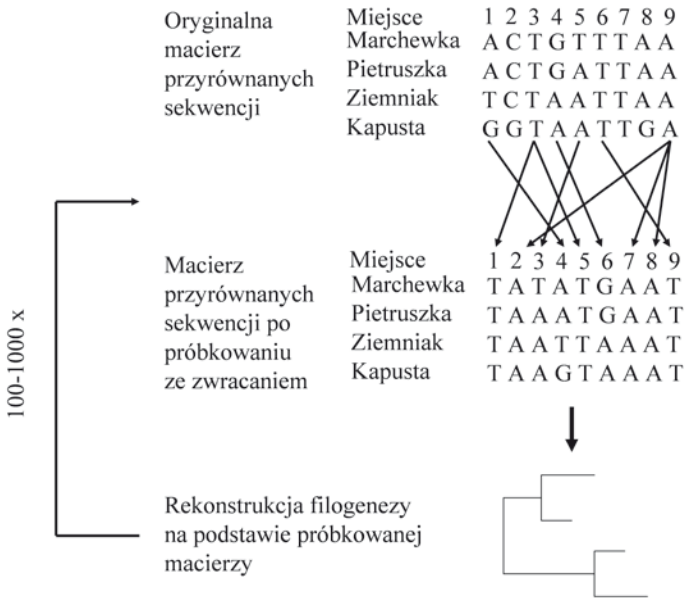
$$P(H_j|A) = P(A|H_j)P(H_j) / P(A).$$

Jest to właśnie wzór Bayesa. Jeśli zdarzenie H jest naszą hipotezą badawczą, to wzór Bayesa pozwala nam obliczyć jej prawdopodobieństwo *a posteriori*, czyli po zajściu zdarzenia A, pod warunkiem że znamy $P(H_i)$, czyli prawdopodobieństwo tej hipotezy *a priori* (przed doświadczeniem — w naszym przypadku jest to wiedza o prawdopodobieństwie wylosowania poszczególnych urn).

Aby przełożyć ten przykład na język filogenetyki, wystarczy za zdarzenie A podstawić nasze dane wyjściowe, czyli przyrównane sekwencje, zaś za hipotezę H — drzewo filogenetyczne wraz z długościami gałęzi. Wtedy można zadać pytanie: jakie jest prawdopodobieństwo poszczególnych drzew filogenetycznych przy danym zestawie przyrównanych sekwencji. Mimo prostoty wzoru Bayesa, jego zastosowanie w filogenetyce napotyka na poważne problemy, a mianowicie na kwestię wyboru wartości prawdopodobieństwa *a priori* dla stawianej hipotezy, czyli drzew filogenetycznych, oraz na pytanie, jak sprawdzić wszystkie możliwe drzewa. W drzewie filogenetycznym można wyróżnić: topologię (kolejność rozgałęzień) oraz długości gałęzi, które określone są przez parametry modelu substytucji nukleotydów. Musimy więc nadać prawdopodobieństwo *a priori* wszystkim składnikom budującym filogenezę. Ponieważ zazwyczaj nie mamy żadnej wiedzy na ten temat, przyjmujemy często tzw. wartości nieinformacyjne *a priori*, które nie wpływają na prawdopodobieństwo *a posteriori* — a przynajmniej nie powinny wpływać, co niestety nie jest do końca prawdą. Oprócz wybrania odpowiedniego rozkładu *a priori*, pojawia się także problem przeszukiwania kombinacji wszystkich parametrów. Przy bardziej skomplikowanych modelach, do których należy rekonstrukcja filogenezy, statystyka bayesowska posiłkuje się algorytmem Monte Carlo z wykorzystaniem łańcuchów Markowa (ang. Markov Chain Monte Carlo, MCMC). Algorytm ten działa w ten sposób, że przeszukuje przestrzeń wszystkich możliwych filogenez, pobierając z niej próby. Zatrzymuje się jednak najdłużej w tym miejscu przestrzeni, w którym drzewa filogenetyczne mają najwyższe prawdopodobieństwo *a posteriori*. Drzewa o najwyższym prawdopodobieństwie zostaną próbkowane wielokrotnie — i właśnie stosunek liczby próbkowań, w których uzyskano dane drzewo, do ich ogólnej liczby, to właśnie prawdopodobieństwo *a posteriori* danego drzewa. Jeśli nasze dane niosą ze sobą dużo informacji, w wyniku działania algorytmu otrzymamy niewielką liczbę drzew o wysokim prawdopodobieństwie i niewiele różniących się od siebie.

OSZACOWANIE WEWNĘTRZNEGO WSPARCIA WĘZŁÓW

Metody rekonstrukcji drzew filogenetycznych, takie jak metoda największej parsymonii, największej wiarygodności oraz odległościowe traktowane są jako tzw. oszacowania punktowe. Oznacza to, że przy odpowiednio dużej liczbie danych (i silnym sygnale filogenetycznym) otrzymujemy tylko jedno drzewo, które jest najlepsze przy danym kryterium rekonstrukcji. Nasuwa się zatem pytanie, jak ocenić niepewność w oszacowaniu poszczególnych kladów na tym drzewie. Do tego celu najczęściej wykorzystuje się metodę *bootstrap*. Metoda ta polega na losowaniu ze zwracaniem poszczególnych miejsc w ma-



Rycina 3. Konstrukcja próbki *bootstrap* polegająca na losowaniu ze zwracaniem z oryginalnej macierzy przyrównanych sekwencji. Powstałą macierz wykorzystuje się do rekonstrukcji filogenezy. Całą procedurę powtarza się setki lub tysiące razy.

cierzy przyrównanych sekwencji do momentu utworzenia nowej macierzy o tej samej liczbie miejsc (kolumn w macierzy), jak w oryginalnej. Na podstawie tej nowej macierzy rekonstruowana jest filogeneza według takiego samego kryterium, jak w wypadku danych oryginalnych. Cały ten cykl próbkowania powtarza się setki lub tysiące razy, a następnie dla każdego kladu występującego w drzewie pierwotnym zlicza się procent drzew, w których dany klad wystąpił – jest to właśnie wartość wsparcia *bootstrap* dla danego węzła (Ryc. 3).

Warto zauważyć, że jednym z założeń tej metody jest to, że miejsca w przyrównanych sekwencjach są próbkami niezależnymi. Jednak bardzo często poszczególne miejsca są ze sobą skorelowane. Przykładowo, w sekwencjach kodujących skorelowane są miejsca należące do tego samego kodonu, natomiast w sekwencjach, które nie kodują białka, ale po transkrypcji przybierają określoną i funkcjonalnie ważną strukturę przestrzenną (rRNA, introny, transkrybowane przestrzenie międzygenowe itd.), skorelowane są fragmenty tworzące struktury dwuniciowe (np. w tzw. „spinkach do włosów"). W takim wypadku metoda *bootstrap* może prowadzić do błędnego oszacowania wsparcia węzłów.

FILOGENEZA I CZAS EWOLUCYJNY

Zaproponowana przez ZUCKERKANDLA i PAULINGA (1965) hipoteza zegara molekularnego zakłada, że tempo ewolucji jest stałe w czasie oraz pomiędzy gałęziami drzewa filogenetycznego. Do takich założeń doprowadziły autorów wcześniejsze obserwacje dotyczące badań nad cytochromem *c* (MARGOLIASH 1963) oraz fibrynopeptydami (DOOLITTLE i BLOMBACK 1964), które sugerowały, że różnice między peptydami są mniej więcej proporcjonalne do czasu dywergencji między gatunkami. Hipoteza zegara molekularnego otrzymała także wsparcie w postaci neutralnej teorii ewolucji molekularnej KIMURY (1983). Od początku jednak zdawano sobie sprawę, że każdy taki „zegar" odmierza czas w różnym tempie w różnych liniach filogenetycznych, a także może przyspieszać lub zwalniać. Założenie ścisłego zegara molekularnego jest w rzeczywistości wyjątkowo rzadko spełnione, zazwyczaj tylko dla niewielkich grup blisko spokrewnionych gatunków. Badania nad tempem ewolucji molekularnej pokazały, że jest ono skorelowane z czasem generacji — im krótszy czas generacji, tym szybsze tempo substytucji. U roślin czas generacji związany jest z formą życiową (drzewa i krzewy żyją dłużej niż rośliny zielne), co przekłada się na związek między formą życiową a tempem ewolucji molekularnej (SMITH i DONOGHUE 2008). Aby uwzględnić te zjawiska przy szacowaniu czasu rozejścia się organizmów, osłabiono założenia zegara, tworząc grupę metod określanych wspólną nazwą „rozluźnionego zegara molekularnego" (ang. relaxed molecular clock). Opracowano

różne podejścia do tego zagadnienia, np. zakładając autokorelację tempa substytucji w liniach filogenetycznych (co ma uzasadnienie, jeśli tempo substytucji jest skorelowane z czasem generacji) albo przyjmując, że tempo to jest niezależne i próbkowane z rozkładu log-normalnego. Wszystkie te metody pozwalają na uzyskanie chronogramu, a zatem drzewa, w którym długości gałęzi są proporcjonalne do czasu.

Aby przełożyć długości gałęzi drzewa filogenetycznego na czas absolutny potrzebujemy tzw. punktów kalibracyjnych. Musimy bowiem pamiętać, że na długość gałęzi wpływają dwa czynniki — tempo substytucji nukleotydów oraz czas. Załóżmy na przykład, że dwie sekwencje DNA różnią się między sobą podstawieniami w 10% miejsc. Jeśli tempo substytucji wynosiło 1% miejsc (pozycji w sekwencji) na milion lat, to ich wspólny przodek żył pięć milionów lat temu, ale równie dobrze obie sekwencje mogły ewoluować pięć razy szybciej przez milion lat. Sytuację tę można porównać do próby oszacowania czasu jazdy, bazując tylko i wyłącznie na wskazaniu licznika przejechanych kilometrów. Aby wykalibrować zegar molekularny, potrzebujemy datowania jakiegoś zdarzenia w przeszłości. Najlepiej, jeśli jest to skamieniałość, którą można przypisać konkretnej gałęzi wewnętrznej na drzewie filogenetycznym. Umiejscawiamy ją w węźle, z którego dana gałąź wychodzi albo do którego wchodzi (to temat do osobnej dyskusji), dzięki czemu możemy datować pozostałe węzły. W ostatnich latach nastąpił duży postęp w rozwoju metod szacowania czasów dywergencji, w tym bazujących na wnioskowaniu bayesowskim. Umożliwiają one wprowadzenie niepewności datowania punktów kalibracyjnych w postaci odpowiedniego rozkładu prawdopodobieństwa *a priori*, a w wyniku uzyskuje się nie tylko punktowe oszacowanie wieku poszczególnych węzłów, ale i rozkład gęstości prawdopodobieństwa tego oszacowania.

Różnice między datowaniem za pomocą ścisłego i rozluźnionego zegara molekularnego dobrze ilustruje przykład roślin kwiatowych. Wykorzystując różne sekwencje i ścisły zegar molekularny oszacowano ich wiek na 420–350 mln lat, 354–300 lub 200 mln lat, a zatem te datowania były nie tylko niezgodne ze sobą, ale i z danymi kopalnymi, albowiem sugerowały, że rośliny okrytozalążkowe powstały nie tylko znacznie wcześniej niż na to wskazują ich najstarsze skamieniałości, ale nawet wcześniej niż dotychczasowe oszacowania wieku roślin nasiennych, wynoszące około 390–350 mln lat. Większość datowań korzystających z rozluźnionego zegara molekularnego waha się natomiast w granicach 180–140 mln lat. Na podstawie danych kopalnych powstanie roślin kwiatowych szacowano na około 131–125 mln lat temu, kiedy to pojawiają się charakterystyczny dla nich pyłek oraz *Archaefructus* — najstarsze pozostałości rośliny zielnej.

Trzeba jednak nadmienić, że szacowanie czasu dywergencji za pomocą zegara molekularnego ma także swoich zdecydowanych przeciwników. Wskazują oni na arbitralność wielu decyzji, które trzeba podjąć przy takim wnioskowaniu, jak np. przypisanie skamieniałości do określonego węzła oraz wybór rozkładu *a priori* w analizie bayesowskiej, które znacząco wpływają na końcowy wynik. Przykładowo, w naszych badaniach nad roślinami z plemienia Oenantheae z rodziny baldaszkowatych, zmieniając przypisany punktom kalibracyjnym typ rozkładu prawdopodobieństwa *a priori* z równomiernego na log-normalny uzyskaliśmy dramatycznie różne oszacowania — 21 lub 45 mln lat — dla tego samego zbioru danych. Pokazuje to, że do wyników szacowania bezwzględnego czasu ewolucyjnego należy podchodzić z dużą ostrożnością, zwłaszcza jeśli służą one dalszemu wnioskowaniu, np. biogeograficznemu.

FILOGENEZA JAKO PODSTAWA BIOLOGII PORÓWNAWCZEJ I EWOLUCYJNEJ

Drzewa filogenetyczne są wykorzystywane nie tylko do weryfikacji systemu klasyfikacji organizmów, ale także do rekonstrukcji ich ewolucji — i właśnie takie zastosowanie jest najbardziej ekscytujące. Ze względu na niekompletność zapisu kopalnego, zwłaszcza w wypadku organizmów lądowych, często jedynym sposobem wnioskowania o historii ewolucyjnej organizmów jest właśnie drzewo filogenetyczne i współczesna różnorodność organizmów, czyli dane neontologiczne, nazywane tak dla odróżnienia od danych paleontologicznych. Analizując rozkład cech na liściach drzewa, możemy zrekonstruować sta-

ny tych cech w jego wewnętrznych węzłach. Podobnie jak w wypadku nukleotydów, rekonstrukcji tych można dokonać za pomocą różnych metod, w tym największej parsymonii, największej wiarygodności lub analizy bayesowskiej.

Do czego może się przydać taka analiza? Czasem chcemy po prostu dobrze wyjaśnić ewolucję danej grupy organizmów, pokazać kolejne etapy jej różnicowania się lub uzyskiwania określonych adaptacji. Czasem interesuje nas koewolucja określonych cech — chcielibyśmy się na przykład dowiedzieć, czy istnieją pewne syndromy adaptacyjne do określonych warunków, czyli grupy współewoluujących cech. Innym razem chcemy sprawdzić, czy uzyskanie określonej nowości ewolucyjnej zbiega się na drzewie filogenetycznym z radiacją danej grupy organizmów. Możliwości wykorzystania wiedzy o ewolucji cech jest wiele.

Badania porównawcze prowadzono już od dawna, ale przed rozwojem filogenetyki molekularnej miały one wątpliwą wartość. Biologia porównawcza kręciła się w błędnym kole, albowiem dysponując jedynie danymi fenotypowymi wykorzystywała je zarówno do szacowania filogenezy, jak i rekonstrukcji ewolucji cech. Takie podejście obarczone jest poważnym błędem. Jeśli bowiem podobieństwo fenetyczne jest wynikiem ewolucji zbieżnej, to uzyskamy błędną filogenezę i zjawiska konwergencji nie wyłapiemy. Jeśli badamy korelację ewolucyjną cech, to nie możemy jej badać na filogenezie uzyskanej z tych cech (albowiem metody filogenetyczne zakładają brak tej korelacji). Dopiero filogenetyka molekularna dostarczyła silnie wspartych drzew uzyskanych na podstawie niezależnych danych i w mniejszym stopniu podatnych na konwegencję.

Przez wiele lat jedyną metodą wykorzystywaną do rekonstrukcji ewolucji cech była metoda największej parsymonii. Jest to stosunkowo prosta i dobra metoda, ale podobnie jak w wypadku rekonstrukcji stanów cech nukleotydów (patrz powyżej) czasem zawodzi, zwłaszcza w dużej skali czasowej. Dlatego też coraz częściej wykorzystywane są inne metody, np. maksymalnej wiarygodności lub bayesowskie. Podobnie jak w wypadku analiz sekwencji, metody te wymagają założenia określonego modelu. Jednym z podstawowych jest model bazujący na ruchach Browna (proces Wienera). Zakłada on, że cecha ewoluuje pod wpływem dryfu genetycznego lub pod wpływem doboru naturalnego, którego kierunek zmienia się w sposób nieprzewidywalny (nie ma doboru kierunkowego). Ponieważ procesy ewolucyjne nie są czysto losowe, poszukiwano także metod, które pozwoliłyby na modelowanie siły doboru i rozluźnienie założenia o czystej losowości. Taki jest np. model bazujący na procesie stochastycznym nazwanym od dwóch holenderskich fizyków procesem Ornsteina-Uhlenbecka. Model ten jest bardziej realistyczny od modelu ruchów Browna, ponieważ ma parametr pozwalający na ograniczenia w zmianach cechy, co pozwala symulować ewolucję pod wpływem doboru naturalnego. Bardzo ciekawy empiryczny test metod rekonstrukcji cech przodków przeprowadzili WEBSTER i PURVIS (2002). Ze względu na bardzo obszerny, niemalże kompletny zapis kopalny ewolucji otwornic (Foraminifera), znali oni wartości cech przodków dla węzłów zrekonstruowanego drzewa filogenetycznego współcześnie żyjących gatunków. Mogli więc porównać oszacowania tych węzłów za pomocą różnych metod ze stanem faktycznym. Okazało się, że najlepiej sprawdziła się metoda bazująca na modelu Ornsteina-Uhlenbecka.

Warto wspomnieć, że rekonstrukcja cech przodków nie musi się ograniczać tylko do cech fenotypowych organizmu, ale może dotyczyć jego środowiska życia albo zasięgu geograficznego. Takie pytania rodzą się w badaniach biogeografii historycznej, paleoekologii lub uwarunkowań kladogenezy. Badając np. zmiany tempa dywersyfikacji — czyli wypadkowej specjacji i wymierania — pytamy, który z czynników odpowiada za to zjawisko. Najczęściej wymienia się dwa typy uwarunkowań, które mogą mieć wpływ na zmiany tempa dywersyfikacji:

a) uwarunkowania wewnętrzne, jakimi są inherentne właściwości organizmów sprzyjające ewolucyjnemu różnicowaniu się; zwraca się szczególną uwagę na kluczowe innowacje adaptacyjne — u roślin są to cechy związane z morfologią kwiatów, formą życiową oraz typem owocu i związanym z nim mechanizmem rozsiewania się;

b) uwarunkowania zewnętrzne, jakimi są np. czynniki geograficzne i klimatyczne; powstawanie barier sprzyja specjacji, natomiast zanikanie barier ułatwia migracje; takie bariery mogą powstawać wskutek zjawisk geologicznych (wędrówki kontynentów, zanikanie i pojawianie się pomostów lądowych, zmiany poziomu mórz, orogeneza itd.) albo klimatycznych (bariery termiczne, zlodowacenia i ustępowanie gatunków do ostoi itd.); zmiany

klimatyczne powodują wymieranie starych gatunków, a także powstawanie nowych.

W obydwu przypadkach często nie mamy wiedzy paleontologicznej na temat warunków, w jakich występował, lub cech, jakie posiadał przodek badanych gatunków. Jeśli umiemy odpowiednio zakodować cechy, w tym ekologiczne, oraz wybrać odpowiedni model zmian wzdłuż gałęzi drzewa filogenetycznego, to można taką rekonstrukcję przeprowadzić. Pozwoli ona na ustalenie, ile razy i w którym momencie nastąpiło przejście do innych warunków ekologicznych. Na przykład, HARDY i LINDER (2005) wykorzystując kilka metod, zrekonstruowali najbardziej prawdopodobne warunki ekologiczne, w jakich żył przodek rodzaju *Thamnochortus* z Afryki Południowej. Okazało się, że żył on w typie siedliska, jakie występuje dzisiaj w południowo-zachodniej, górzystej części florystycznego regionu przylądkowego w Afryce Południowej, a jego potomkowie skolonizowali siedliska o niższej amplitudzie opadów atmosferycznych i położone niżej, przystosowali się także do większego spektrum warunków glebowych.

W POSZUKIWANIU DRZEWA ŻYCIA

Rozwój metod molekularnych, w tym wysoko wydajnego sekwencjonowania, stwarzają filogenetyce molekularnej nowe, niezwykłe możliwości. Narodziła się filogenomika – analizująca nie poszczególne sekwencje, ale całe genomy, np. mitochondrialne albo chloroplastowe. Dużym osiągnięciem było zsekwencjonowanie kompletnego genomu mitochondrialnego neandertalczyka oraz porównanie go z genomami współczesnych ludzi (GREEN i współaut. 2008). Pozwoliło to na oszacowanie czasu rozejścia się *Homo sapiens* i *Homo neanderthalensis* na 660 ± 140 tys. lat temu — znacznie dokładniejsze i z mniejszym błędem niż poprzednie oszacowania, bazujące na pojedynczych sekwencjach. Warto zauważyć, że sygnał filogenetyczny zawarty w genomach to nie tylko sekwencje poszczególnych odcinków, ale także informacja o zmianach strukturalnych — o duplikacjach i utracie genów, zmianach ich położenia, fuzjach, transferze poziomym itd.

Niekwestionowane sukcesy filogenetyki molekularnej skłaniają do zadania pytania, czy poznamy kiedyś kompletne drzewo życia. Pomijając fakt, że nie znamy jeszcze wszystkich gatunków żyjących na Ziemi, a wiele z nich wyginie, zanim je opiszemy, to jest to przedsięwzięcie możliwe do wykonania. Pamiętajmy jednak, że będzie to drzewo przybliżone, albowiem — jak to już zaznaczyliśmy — nie zawsze w materiale genetycznym organizmów zachował się czytelny sygnał filogenetyczny, a metody rekonstrukcji filogenezy niekiedy zawodzą. Tym niemniej, warto próbować.

PHYLOGENY ESTIMATION AND PHYLOGENETIC INFERENCE IN EVOLUTIONARY STUDIES

Summary

Modern phylogenetics, although rooted in Darwin's and Haeckel's ideas on evolutionary relationships among organisms, dates back to the second half of the 20th century and the advance of numerical methods in taxonomy. Its beginnings were marked by a fierce debate between phenetics and cladistics but at present it incorporates a diverse array of methods including those based on distance and clustering algorithms, parsimony, maximum likelihood and Bayesian statistics. The phylogeny of extant organisms is usually inferred using molecular markers, because they are genetic, less arbitrary (do not require arbitrary coding), more additive, less prone to convergence and more universal than traditional morphological markers. Phylogenies inferred using molecular data are usually more stable and have better internal support than those obtained from morphology. However, the informed user of phylogenetics methods must be aware of their assumptions and caveats. The chosen sequences must be orthologous (resulting from a speciation event), as opposed to paralogous (resulting from a duplication event); choosing orthologous sequences does not guarantee that the phylogenetic signal is undisturbed. Reversals, multiple hits and parallel substitutions may result in a higher similarity of sequences than expected from their evolutionary history and therefore affect the phylogenetic reconstructions. Moreover, trees inferred from molecular data are usually gene trees rather than species trees. There are several processes that may result in discordance between a gene tree and an organism tree including interspecific hybridisation, horizontal gene transfer, incomplete lineage sorting and selection for allele polymorphism. The most commonly used phylogenetic methods include those based on parsimony, distance and clustering, maximum likelihood and Bayesian statistics. The last three employ nucleotide

substitution models. Each method is based on certain evolutionary assumptions that may not necessarily apply to a given data set. Noteworthy are recent advances in methods of inferring divergence times using relaxed molecular clock. In evolutionary biology, molecular phylogenies are widely used in comparative studies, historical biogeography and for analysing character state evolution.

LITERATURA

CAMIN J. H., SOKAL R. R., 1965. *A method for deducing branching sequences in phylogeny.* Evolution 19, 311–326.

DOOLITTE R. F., BLOMBACK B., 1964. *Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications.* Nature 202, 147–152.

GREEN R. E., MALASPINAS A.-S., KRAUSE J., BRIGGS A. W., JOHNSON P. L., UHLER C., MEYER M., GOOD J. M., MARICIC T., STENZEL U., PRÜFER K., SIEBAUER M., BURBANO H. A., RONAN M., ROTHBERG J. M., EGHOLM M., RUDAN P., BRAJKOVIĆ D., KUĆAN Z., GUSIĆ I., WIKSTRÖM M., LAAKKONEN L., KELSO J., SLATKIN M., PÄÄBO S., 2008. *A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing.* Cell 134, 416–26.

GREHAN J. R., SCHWARTZ J. H., 2009. *Evolution of the second orangutan: phylogeny and biogeography of hominid origins.* J. Biogeograph. doi:10.1111/j.1365–2699.2009.02141.x.

HARDY C. R., LINDER H. P., 2005. *Intraspecific variability and timing in ancestral ecology reconstruction: A test case from the Cape Flora.* Systematic Biol. 54, 299–316.

JUKES T. H., CANTOR C. R., 1969. *Evolution of protein molecules.* [W:] *Mammalian protein metabolism.* MUNRO H. N. (red.). Academic Press, New York, 21–123.

KIMURA M., 1983. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

MARGOLIASH E., 1963. *Primary structure and evolution of cytochrome C.* Proc. Natl. Acad. Sci. USA 50, 672–679.

SMITH S. A., DONOGHUE M. J, 2008. *Rates of molecular evolution are linked to life history in flowering plants.* Science 322, 86–89.

WEBSTER A. J., PURVIS A., 2002. *Testing the accuracy of methods for reconstructing ancestral states of continuous characters.* Proc. R. Soc. Lond. Series B 269, 143–149.

ZUCKERKANDL E., PAULING L., 1965. *Evolutionary divergence and convergence in proteins.* [W:] *Evolving genes and proteins.* BRYSON V., VOGEL H. J. (red.). Academic Press, New York, 97–166.