Filogenetyka molekularna

Paweł Bujnowski, Tomasz Mularczyk
Listopad 2004

Filogenetyka - dyscyplina zajmująca się odtwarzaniem dróg rozwoju rodowego poszczególnych grup organizmów

Filogeneza – rodowy rozwój, historia rodowa. Filogeneza wynika ze zmian w przebiegu ontogenez kolejnych pokoleń organizmów.

Ontogeneza – rozwój osobniczy [liczony od powstania jego zalążka aż do naturalnej śmierci]

Filogenetyka molekularna – badanie podobieństw i zależności między organizmami lub pomiędzy genami i ich produktami – białkami.

Drzewo filogenetyczne – próba graficznej reprezentacji przebiegu historii ewolucji. Może dotyczyć pojedynczego organizmu bądź całej ich grupy [jedna Afryka vs. wiele regionów].

Podobieństwo genów ze względu na pochodzenie:

Ortologi – geny różnych organizmów, które pełnią te same funkcje [np. hemoglobina człowieka i konia], ale rozwinęły się niezależnie.

Paralogi – geny w jednym organizmie będące skutkiem duplikacji u jakiegoś przodka.

Ksenologi – geny jednego organizmu sprawiające wrażenie podobieństwa tego organizmu do innego, posiadającego podobne geny, aczkolwiek nie należące do oryginalnego organizmu.

Więcej o drzewach filogenetycznych:

- buduje się je w oparciu o ortologi (ew. paralogi)
- węzeł drzewa to gatunek albo gen
- wierzchołki wewnętrzne to hipotetyczni „przodkowie”
- są ukorzenione (skierowane) bądź nie – wtedy interesuje nas tylko, jak bardzo gatunki lub geny są spokrewnione, a nie jak przebiegała ich hipotetyczna ewolucja.
- większość metod pozwala stworzyć drzewo nieukorzenione, a korzeń uzyskać poprzez wprowadzenie niespokrewnionego indywiduum.
- drzewa te zwykle są binarne
- wagi na krawędziach oznaczają odległość między gatunkami bądź genami

Metody rekonstrukcji drzew:

- **Oparte na cechach –** oddzielne badanie każdej cechy (w przypadku DNA oznacza to każdy nukleotyd w sekwencji)
- **Oparte na odległościach –** wejściem dla algorytmu jest macierz odległości między gatunkami / sekwencjami / genami. Wynik powstaje poprzez uliniowienia sekwencji bądź oszacowanie za pomocą matematycznych modeli.

Dlaczego wyznaczanie odległości genetycznej między organizmami nie jest trywialne ?

Rodzaje zmian mutacyjnych w sekwencjach DNA:

- **substytucja –** zamiana jednego nukleotydu na inny
- **delecja –** usunięcie spójnego ciągu nukleotydów
- **insercja –** wstawienie fragmentu nici
- **inwersja –** odwrócenie kierunku pewnego fragmentu nici

Dodatkowa klasyfikacja mutacji:

- **tranzycja –** zamiana puryny (adenina lub guanina) na inną purynę lub zamiana pomiędzy pirymidami (tymina i cytozyna)
- **transwersja –** zamiana pomiędzy tymi grupami

Więcej o mutacjach:

- tranzycje zachodzą częściej niż transwersje
- jeśli nukleotyd zostanie zastąpiony w ten sposób, że kodon ciągle będzie reprezentował ten sam aminokwas, to mutacja jest **synonimiczna**
- mutacja powodująca przekształcenie aminokwasu w kodon stopu jest **nonsensowna**
- insercje i delecje dotyczą zwykle spójnych fragmentów DNA (zwykle niekodujących)
- insercje i delecje mogą powodować ogromne zmiany poprzez przesunięcie ramki odczytu

Zjawisko duplikacji genów:

- zduplikowany gen występuje w organizmie w dwóch lub więcej kopiach
- kopie ewoluują niezależnie od siebie
- prowadzi to do powstania spokrewnionych genów
- zjawisko to powoduje spore problemy w tworzeniu drzew filogenetycznych

Metody rekonstrukcji drzew z macierzy odległości

Paweł Bujnowski, Tomasz Mularczyk 2004

Na metodę rekonstrukcji drzewa składają się:

- **kryterium optymalności** – mierzy dopasowanie posiadanych przez nas danych (zwykle odległości ewolucyjnych do danego drzewa)
- **strategia poszukiwań** – optymalnego drzewa w przestrzeni wszystkich drzew filogenetycznych
- jakie **założenia mechanizmu** ewolucji przyjmujemy

Trudno porównywać ze sobą metody, które różnią się w tych trzech składowych.

Charakterystyka dobrych metod rekonstrukcji:

- **efektywność** – złożoność obliczeniowa algorytmów używanych w metodzie. Sama ewaluacja kryterium optymalności nie stanowi problemu, gorzej z przestrzenią rozwiązań [rozwiązuje się stosując heurystyki lub próbkowanie przestrzeni]
- **zgodność** – wraz ze wzrostem danych powinniśmy zbliżać się do drzewa optymalnego
- **zbieżność –** jak szybko zbiega do optymalnego drzewa [jak długich sekwencji musimy użyć]
- **stabilność** – wynik metody powinnien zostać nieznacznie zmieniony przy niedużym zaburzeniu danych wejściowych

Metoda najmniejszych kwadratów:

- znamy odległości między każdymi parami gatunków / sekwencji [zgromadzone w macierzy odległości D]
- konstruujemy drzewo ważone, które zawiera w liściach rozważane sekwencje i jest zgodne z macierzą D [odległości jako sumy wag po ścieżce łączącej liście]

Wady tej metody:

- zwykle nie istnieje zgodne z macierzą odległości drzewo [szuka się drzewa, które indukuje odległości najbliższe tym dostarczonym w macierzy D]
- problem znalezienia takiego ważonego drzewa jest NP trudny

Będziemy zatem poszukiwać heurystyki, które działają w czasie wielomianowym

- Odległości ultrametryczne
- Odległości addytywne

Metoda odległości ultrametrycznych:

- bazuje na hipotezie zegara molekularnego [częstotliwość zmian sekwencji ma być niezmienna w czasie]
- liczba mutacji jest wprost proporcjonalna do długości przedziału czasowego
- niestety, nie jest to prawdą [częstotliwości mutacji są różne dla różnych białek oraz dla różnych części tego samego białka]
- jeśli znamy liczbę mutacji od chwili rozejścia się dwóch sekwencji, to zakładamy, że w obu z nich zaszła taka sama ich ilość i była ona proporcjonalna do czasu, który upłynął od momentu specjacji
- jeśli trzymamy się tego założenia, to drzewa od korzenia do każdego z liści są równe

Algorytm UPGMA:

- Unweighted Pair Group Method using Arimethic averages.
- łączymy wspólnym przodkiem najbliższe sobie sekwencje
- z połączenia powstaje klaster, który uczestniczy w dalszym łączeniu
- klastry łączymy ze sobą aż powstanie z nich tylko jeden, wynikowy
- Odległość między klastrami wyliczamy z wzoru:

$$Dij = 1 / (|Ci| * |Cj|) * \text{suma } pCi \text{ suma } qCj \; Dpq$$

Kroki algorytmu UPGMA:

1.w wynikowym drzewie przypisz liściom gatunki
repeat
2.znajdź Ci oraz Cj o najmniejszej odległości Dij
3.połącz je w jeden klaster Ck
4.dodaj w drzewie wierzchołek odpowiadający nowemu klastrowi i nadaj wagi jego wagę do synów jako Dij/2
5.policz odległośći między nowym klastrem a pozostałymi jako:
Dkl = (Ci / (Ci + Cj)) Dil + (Cj / (Ci + Cj)) Djl
6.usuń z macierzy D kolumny i wiersze odpowiadające klastrom Ci oraz Cj oraz dodaj kolumnę i wiersz dla nowego klastra Ck
until pozostanie jeden klaster

Odległości addytywne:

- dane ultrametryczne to przypadek idealny – bardzo rzadko spotykany w praktyce
- **addytywność** to słabsze założenie – wynik algorytmu nie będzie zwykle drzewem binarnym [ale może]
- nie oczekujemy już, że zmiany zachodzą w tych samych odstępach czasu
- w drzewie występują wierzchołki nieetykietowane [nie jest zwarte]
- na wejściu macierz odległości między elementami [o diagonalnych elementach równych 0]

Kroki algorytmu Neighbor-Joining:

1.w wynikowym drzewie przypisz liściom gatunki
repeat
2.znajdź i oraz j o najmniejszej wartości
Dij – Ui – Uj
3.połącz je w jeden klaster Ck
4.dodaj w drzewie wierzchołek odpowiadający nowemu klastrowi. Nowe wagi:
dik = 1/2 Dij + 1/2 (Ui – Uj)
djk = 1/2 Dij + 1/2 (Uj – Ui)
5.policz odległości między nowym klastrem Ck, a pozostałymi:
Dkl = 1/2 (Dil + Djl – Dij)
6. Usuń z macierzy D kolumny i wiersze odpowiadające klastrom Ci oraz Cj oraz dodaj kolumnę i wiersz dla nowego klastra Ck
until pozostanie tylko jeden klaster

Właściwości drzewa addytywnego:

- Dla dowolnych trzech liści i, j, l mamy:

$$Dil = Dik + Dkl, \quad Djl = Djk + Dkl, \quad Dij = Dik + Dkj$$

$$Dkl = 1/2\ (Dil + Djl - Dij)$$

- Dla dowolnych czterech liści i, j, k, l dwie spośród poniższych odległości muszą być równe i większe od trzeciej:

$$Dij + Dkl, \quad Dik + Djl, \quad Dil + Djk$$