



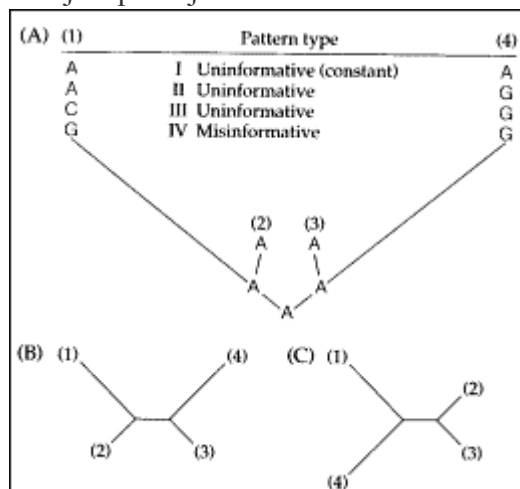
## Metoda największej wiarygodności - ucieczka ze strefy Felsensteina

Krzysztof Spalik  
1.II.2004

Tekst na podstawie kursu przeprowadzonego w Instytucie Zoologicznym Uniwersytetu Wrocławskiego we Wrocławiu w dniu 11 maja 2002.

### Strefa Felsensteina, czyli kiedy parsymonia może zawodzić

Kiedy triumf metod kladystycznych w taksonomii i filogenetyce wydawał się niewątpliwy, pojawił się niemiły zgrzyt. Jego autorem był Joe Felsenstein (1978), który zadał fundamentalne pytanie o spójność metody opartej na parsymonii. Metoda filogenetyczna jest uważana za spójną wtedy, kiedy wraz z napływem nowych danych otrzymujemy drzewa coraz bliższe prawdziwej filogenezie. Felsenstein wykazał, że w pewnych warunkach metody kladystyczne są niespójne. Co gorsza, są pozytywnie mylące: im więcej danych zbieramy, tym silniej wsparte jest określone drzewo filogenetyczne, ale jest to drzewo błędne!



**Ryc. 1.** Przykładowa filogeneza czterech gatunków (A; przedstawione są różne możliwości zmian w liniach 1 i 4) i jej dwie rekonstrukcje (prawdziwa B i fałszywa C).

Rekonstrukcje są drzewami niezakorzenionymi (Swofford i in. 1996).

Rozważmy przypadek drzewa czterech taksonów (ryc. 1 A).

Na drzewie tym zaznaczono długości gałęzi, które odzwierciedlają liczbę

podstawień w sekwencji, a także - przykładowo - zmiany jednego nukleotydu. U wspólnego przodka była to adenina. Po rozdzieleniu jego linii na dwie potomne bardzo szybko nastąpiły kolejne dywergencje.

Na pierwszych gałęziach nie zaszły żadne zmiany i w rozważanej pozycji u obu potomków pozostała adenina. U dalszych potomków wystąpiło silne zróżnicowanie tempa podstawień nukleotydów. Dwie gałęzie są bardzo krótkie - nie zaszły tam prawie żadne zmiany - natomiast dwie są długie, tzn. sekwencje DNA tych taksonów są silnie zmienione. Rozpatrzmy, jakie będą możliwe stany analizowanej cechy (czyli pojedynczej pozycji w sekwencji) u tych taksonów i jakie będą tego konsekwencje dla oszacowania filogenezy metodą parsymonii.

1) Nukleotyd w danej pozycji nie zmieni się - pozostanie adenina. Pozycja ta jest zatem stała, czyli nieinformacyjna filogenetycznie.

2) Jeśli nukleotyd ulegnie podstawieniu na jednej gałęzi, to taka pozycja będzie również nieinformacyjna filogenetycznie. Według zasady parsymonii ważna jest bowiem tylko wspólnota posiadania tej samej cechy zaawansowanej ewolucyjnie.

3) Jeśli na obu gałęziach adenina będzie podstawiona przez różne nukleotydy, np. przez cytozynę i guaninę, to także ta pozycja będzie nieinformacyjna.

4) Z punktu widzenia parsymonii informacyjny filogenetycznie będzie jedynie przypadek, w którym adenina zmieni się na guaninę (albo inną zasadę) równocześnie na obu gałęziach. Ale właśnie ten przypadek jest mylący. Wskazuje bowiem na bliskie pokrewieństwo taksonów (1) i (4), czyli na drzewo C.

W gałęziach ewolucyjnych 1 i 4 zmian było wiele (są "długie"), takich przypadków przypadkowej zbieżności jest więc w nich z pewnością więcej, niż w "krótkich" gałęziach 2 i 3. Im więcej danych zbierzemy, tym bardziej możemy być utwierdzeni, że to właśnie drzewo C jest prawdziwe, podczas gdy będzie to artefakt - jesteśmy po prostu w tzw. strefie Felsensteina. Tą nazwą określa się zbiór topologii rzeczywistego drzewa filogenetycznego (strefę w przestrzeni możliwych topologii), w którym jego odtworzenie jest bardzo trudne.

Można próbować podważyć te rozważania, wskazując, że założenie o tak dużych różnicach w tempie ewolucji poszczególnych gałęzi jest nierealistyczne. Niestety, Hendy i Penny (1989) wykazali, że nawet w wypadku zegara molekularnego, czyli stosunkowo wyrównanego tempa ewolucji na poziomie molekularnym, taki efekt występuje. Nazwali go long branch attraction, czyli przyciąganiem się długich gałęzi. Efekt ten spotykamy, kiedy w próbie taksonów pewne gałęzie ewolucyjne są lepiej reprezentowane niż inne (próba jest niereprezentatywna).

Odpowiedni wybór taksonów nie zawsze jednak zależy od badacza. Na przykład nie uzyskamy bardziej reprezentatywnej próby miłorzębowych, ponieważ miłorzęb dwukłapowy jest jedynym współcześnie żyjącym przedstawicielem tej licznej kiedyś grupy.

Na ile powszechnie spotyka się efekt przyciągania się długich gałęzi? Prawdopodobnie dość często. Istnieją uzasadnione obawy, że taką sytuację mamy w wypadku roślin lądowych - do tej pory nie wiemy na przykład, która grupa mszaków jest grupą siostrzaną roślin naczyniowych, różne sekwencje DNA dają bowiem odmienne oszacowania.

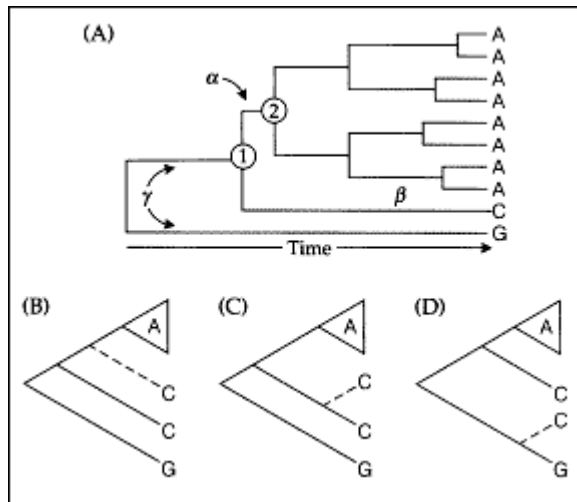
Istnieją metody ucieczki ze strefy Felsensteina. Przyjmują one pewne założenia co do ewolucji na poziomie molekularnym i biorą pod uwagę nie tylko obserwowane zmiany nukleotydów, ale także szacują liczbę wielokrotnych podstawień, które są jedną z przyczyn szumu filogenetycznego. Taką metodą jest m.in. maximum likelihood, czyli metoda największej wiarygodności.

### **Parsymonia nie uwzględnia tempa ewolucji**

Zacznijmy od pokazania, jaka jest różnica między parsymonią a metodą największej wiarygodności (ryc. 2). Drzewo A pokazuje filogenezę sekwencji DNA kilku gatunków. Jak w poprzednio analizowanym przykładzie, zaznaczono zmiany nukleotydów w jednym miejscu sekwencji. Aczkolwiek to drzewo jest narysowane ze strzałką czasu, z formalnego punktu widzenia nie jest zakorzenione. Innymi słowy, nie jest rozłamana gałąź zaznaczonej literą ?. Dla tych rozważań nie jest ważne, w której części tej gałęzi - dolnej czy górnej - zaszły zmiany.

**Ryc. 2.** Przykład ilustrujący różnice w podejściu do rekonstrukcji filogenezy między metodami opartymi na parsymonii a największej wiarygodności (Swofford i in. 1996).

Jeśli do zbioru taksonów z ryciny (A) dołączymy obiekt z cytozyną w analizowanej pozycji sekwencji, to według kryterium parsymonii wszystkie drzewa (B)-(D) są równocenne, natomiast funkcja wiarygodności wskazuje na drzewo (C).



Cyframi 1 i 2 zaznaczono węzły wspólnych przodków dwu grup gatunków. Spróbujmy dociec, jaki nukleotyd występował w rozważanej pozycji u tych przodków. W wypadku przodka 2 sprawa jest oczywista - zgodnie z zasadą parsymonii powinna tam być adenina.

Ale co z przodkiem 1? Ponieważ na każdej z trzech gałęzi wychodzących z tego węzła znajduje się inny stan (A, C lub G), wszystkie trzy rekonstrukcje są możliwe. Każda z nich wymaga tylko dwóch podstawień.

Zastanówmy się, gdzie dodana byłaby sekwencja, która ma cytozynę w analizowanej pozycji (ryc. 1, drzewa B, C i D). Pod względem parsymonii równie dobre jest dołączenie tej sekwencji do gałęzi  $\alpha$ ,  $\beta$ , lub  $\gamma$ . Nie wymaga to żadnej dodatkowej zmiany. Podobnie byłoby, gdyby ta sekwencja miała guaninę. Gdzie natomiast najlepiej byłoby dołączyć sekwencję, która miałaby tyminę w tej pozycji? Ją również można dodać do każdej gałęzi tego drzewa, w każdym przypadku bowiem wymaga to jednej dodatkowej zmiany. Podobnie jest z sekwencją, która ma w tej pozycji adeninę - ona także może być dołączona wszędzie.

Wpłynie to jednak na rekonstrukcję wspólnego przodka. Jeśli sekwencję z adeniną dołączymy do gałęzi  $\beta$  lub  $\gamma$ , to musimy założyć, że przodek nr 1 miał także adeninę w tej pozycji. A zatem analizowane miejsce jest informacyjne filogenetycznie jedynie w stosunku do sekwencji z cytozyną i guaniną, natomiast nic nie mówi o przypuszczalnym położeniu sekwencji z tyminą i adeniną.

Zauważmy, że w naszych rozważaniach dotyczących parsymonii ani razu nie pojawił się czas. Metoda parsymonii go zupełnie nie uwzględnia. Ważna jest tylko liczba zmian na całym drzewie.

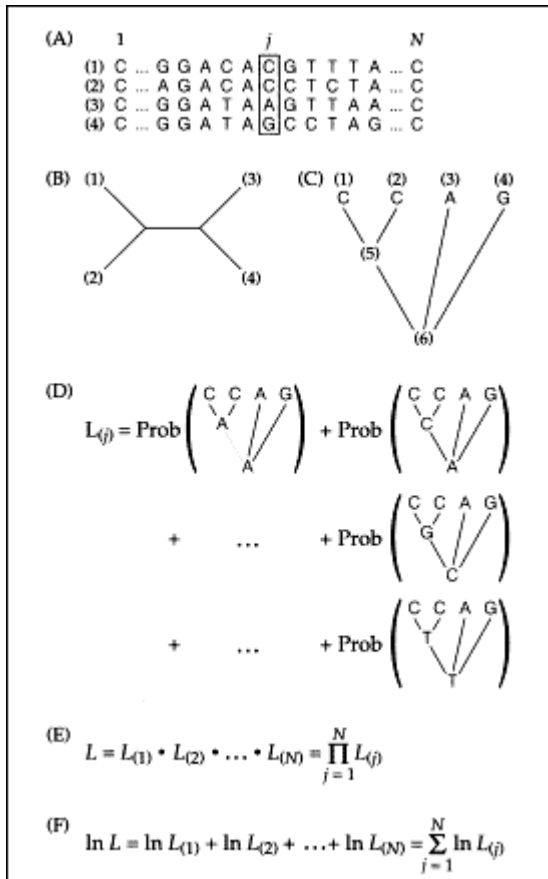
### Metoda największej wiarygodności przyjmuje model podstawiania nukleotydów

Inną perspektywę przyjmuje metoda największej wiarygodności. Stosuje ona odmienne kryterium oceny drzewa. Wybieramy takie drzewo, dla którego prawdopodobieństwo osiągnięcia obserwowanego rozkładu wartości cech na wierzchołkach gałęzi jest najwyższe. Aby oszacować to prawdopodobieństwo, musimy dokonać założeń o przebiegu zmian ewolucyjnych - modelu substytucji nukleotydów w sekwencji.

Przyjmijmy model najprostszy:

- (1) tempo podstawiania (substytucji) jest jednakowe dla wszystkich par nukleotydów,
- (2) wszystkie typy podstawień są jednakowo prawdopodobne,
- (3) spodziewana liczba podstawień na dowolnej gałęzi jest funkcją tempa substytucji i długości tej gałęzi (czasu od rozejścia się linii ewolucyjnych).

Te założenia to w zasadzie uproszczony model Jukes-Cantora. Na razie zakładamy także, iż tempo substytucji jest jednakowe w całym drzewie.



Ryc. 3. Obliczanie funkcji wiarygodności (Swofford i in. 1996).

(A) - macierz przyrównanych sekwencji

(B) - jedno z możliwych niezakorzenionych drzew

(C) - drzewo zakorzenione

(D) - suma prawdopodobieństw dla danej pozycji sekwencji

(E) - wartość funkcji wiarygodności to iloraz wartości z wszystkich pozycji sekwencji

(F) - ponieważ jest to bardzo mały ułamek, przedstawia się go w postaci logarytmu naturalnego.

Miarą jakości drzewa jest suma prawdopodobieństw zliczona po wszystkich cechach (ryc. 3).

Na przykład liczymy prawdopodobieństwo osiągnięcia obserwowanego rozkładu nukleotydów na gałęziach drzewa przy założeniu, że przodek nr 2 miał adeninę, a następnie, że miał tyminę, cytozynę lub guaninę. Sumujemy te prawdopodobieństwa - to jest właśnie funkcja wiarygodności. Należy pamiętać,

że funkcja ta nie określa prawdopodobieństwa prawdziwości drzewa, ale jest miarą prawdopodobieństwa osiągnięcia obserwowanego rozkładu stanów cech przy założeniu, że dane drzewo jest prawdziwe.

Zmiany ewolucyjne są rzadkie, a zatem historie ewolucyjne, w których zaszło mniej zmian są bardziej prawdopodobne niż takie, w których tych zmian zaszło więcej. U przodka nr 2 z ryc. 2, bardziej prawdopodobne jest występowanie adeniny niż innego nukleotydu i ono najwięcej wnosi do funkcji wiarygodności. W tym wypadku metoda największej wiarygodności nie odbiega od parsymonii.

Rozważmy teraz węzeł nr 1 (ryc. 2). Wychodzą z niego trzy gałęzie, na każdej z nich występuje inny nukleotyd: A, C i G. Zastanówmy się najpierw, który stan - A czy C - jest bardziej prawdopodobny. Jeśli przodek nr 1 miał adeninę, jak przypuszczamy, to pomiędzy tym węzłem a wierzchołkiem drzewa z cytozyną musiała zajść jakaś zmiana. Ta zmiana mogła zajść albo na gałęzi  $\alpha$ , albo na gałęzi  $\beta$ . Na której z nich jest bardziej prawdopodobna? Gałąź  $\alpha$  jest stosunkowo krótka, natomiast gałąź  $\beta$  jest znacznie dłuższa. Długość gałęzi odzwierciedla liczbę podstawień, a zatem jest bardziej prawdopodobne, że rzeczona zmiana zaszła na gałęzi  $\beta$  niż na gałęzi  $\alpha$ . Pamiętajmy jednak, że wciąż mówimy o prawdopodobieństwie. Metoda największej wiarygodności nie określa, że ta zmiana zaszła na gałęzi  $\beta$ , ponieważ jest to bardziej prawdopodobne. Jedynie przy liczeniu funkcji wiarygodności dla danego drzewa, taka hipoteza wniesie najwięcej do obliczanej sumy prawdopodobieństw.

W podobny sposób rozważamy, na której gałęzi zaszła zmiana prowadząca do guaniny - na gałęzi  $\beta$  czy  $\alpha$ . Jest bardziej prawdopodobne, że na gałęzi  $\alpha$ , ponieważ jest ona dłuższa od gałęzi  $\beta$ . A zatem wartość funkcji daje nam porządek najbardziej prawdopodobnych rozwiązań. Najbardziej prawdopodobny stan u przodka 1 to adenina, potem cytozyna, a na końcu guanina. Jakie to ma znaczenie? Wróćmy do naszego oryginalnego problemu - w

którym miejscu będzie dołączona sekwencja z cytozyną w określonej pozycji sekwencji? Przypominam, że w świetle parsymonii drzewa A, B i C są równoprawne. Natomiast w świetle metody największej wiarygodności nie są. Drzewo B wymagałoby rozcięcia gałęzi ? i umiejscowienia zmiany na tej krótkiej gałęzi, co jest mało prawdopodobne. Drzewo D nie wymagałoby rozcięcia tej gałęzi, ale wymagałoby zmiany na niej. Oczywiście możnaby uniknąć podstawienia na gałęzi ?, jeśli założymy dodatkową, niezależną zmianę na tej dołączonej gałęzi. Jest to jednak jeszcze mniej prawdopodobne. A zatem, najbardziej prawdopodobne będzie drzewo C. Zauważmy, że przewaga drzewa C znika, jeśli wydłużamy gałąź ?, np. jeśli przyjmiemy, że tempo ewolucji na tym odcinku jest szybsze.

Podsumowując, metoda największej wiarygodności, w odróżnieniu od metody opartej na parsymonii, uwzględnia długość gałęzi drzewa filogenetycznego. Dlatego też jest spójna. Taki układ bowiem, jaki pojawia się na drzewie w wypadku strefy Felsensteina, jest bardzo prawdopodobny, a zatem będzie wychwycony. Swofford i in. (1996) uważają, że metoda największej wiarygodności jest najlepszą metodą szacowania filogenezy nie tylko z uwagi na spójność. Ma niższą wariancję od innych metod, co znaczy, że jest najmniej wrażliwa na błąd pobierania próby, a także jest bardziej odporna na odstępstwa od założeń o modelu ewolucji.

### Wybór właściwego modelu, czyli diabeł tkwi w szczegółach

Pominęliśmy jak dotąd bardzo ważny etap - w jaki sposób oblicza się częst-kowe prawdopodobieństwa składające się na wartość funkcji wiarygodności dla określonego drzewa. Aby to obliczyć, musimy przyjąć pewne założenia o substytucji nukleotydów, czyli model ewolucji (neutralnej) na poziomie molekularnym. Szczegółowa prezentacja poszczególnych modeli wykracza poza ramy tego omówienia. Warto jednak wiedzieć, że podstawą każdego modelu jest macierz tempa substytucji.

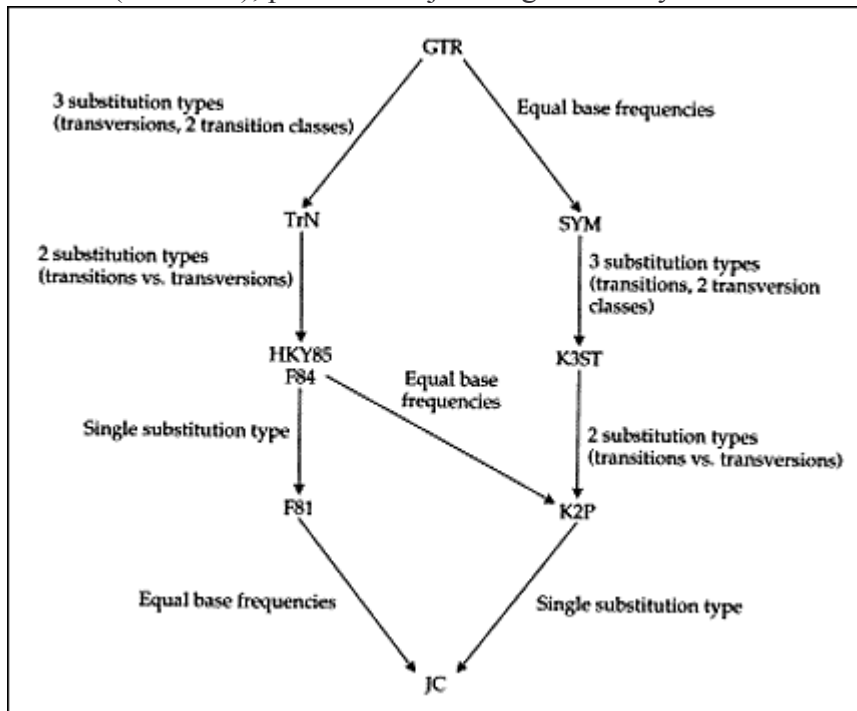
Najbardziej ogólną postać tej macierzy przedstawia ryc. 4. Tempo substytucji jednego nukleotydu przez drugi zależy od średniego tempa substytucji  $\Gamma$ , stałej dla każdego typu podstawienia oraz częstości podstawianego nukleotydu. Ponieważ nukleotydów jest cztery, w najbardziej ogólnym modelu substytucji mamy 12 typów podstawień (nie licząc podstawień synonimicznych). Zwykle jednak zakłada się pełną odwracalność ewolucji (model GTR - general time reversible). Macierz jest zatem symetryczna wzdłuż przekątnej, a tym samym otrzymujemy sześć typów substytucji. Niemalże wszystkie modele substytucji DNA są specjalnymi przypadkami modelu GTR (ryc. 5).

Na przykład jeśli w modelu GTR ograniczymy liczbę typów przekształceń z sześciu do trzech: transwersji (podstawienia puryny przez pirymidynę i odwrotnie) dwóch typów tranzycji (podstawienia jednej puryny przez drugą i jednej pirymidyny przez drugą), to otrzymamy model Tamury i Nei. Jeśli będziemy rozpatrywać tylko tranzycje i transwersje, to otrzymamy model Hasegawa-Kishino-Yano z 1985 r. albo model Felsensteina z 1984 r. Jeśli te modele uprościmy, zakładając tylko jeden rodzaj podstawienia, to otrzymamy model Felsensteina z 1981 r. A jeśli w tym modelu założymy dodatkowo, że częstości występowania wszystkich nukleotydów są równe, to dojdziemy do najprostszego modelu Jukes-Cantora. Modele substytucji używane są nie tylko w metodzie największej wiarygodności, ale i w metodach odległościowych.

$$Q = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{pmatrix}$$

Ryc. 4. Macierz tempa substytucji nukleotydów (Swofford i in. 1996). Rzędy i kolumny od lewego górnego rogu dotyczą kolejno adeniny, cytozyny, guaniny i

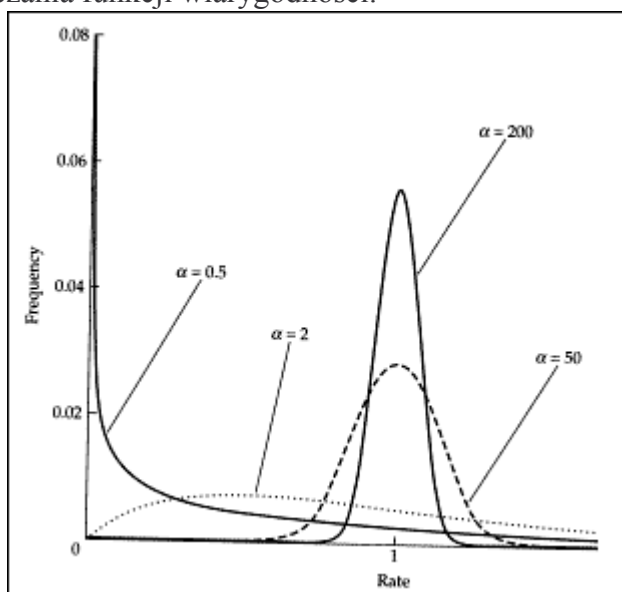
tyminy. Parametry:  $m$  - bezpośrednie tempo podstawień;  $a-l$  - stałe dla każdego typu podstawienia (razem 12);  $p$  - frekwencja danego nukleotydu.



Ryc. 5. Zależności między najpowszechniej używanymi modelami substytucji nukleotydów (Swofford i in. 1996).

Modele: **GTR** - general time reversible; **TrN** - Tamura i Nei; **HKY85** - Hasegawa-Kishino-Yano 1985; **F84** - Felsenstein 1984; **SYM** - model Zharkikha 1994; **K3ST** - trójparametryczny model Kimury; **K2P** - dwuparametryczny model Kimury; **JC** - Jukes-Cantor.

Macierz substytucji, którą przedstawiliśmy, służy do obliczenia macierzy prawdopodobieństwa zmian jednego nukleotydu w drugi. Ta właśnie macierz jest podstawą wyliczania funkcji wiarygodności.



Ryc. 6. Zmiany kształtu rozkładu  $\alpha$  w zależności od wartości parametru  $\alpha$  (Swofford i in. 1996).

W rozważanym dotychczas modelu założyliśmy, że ewoluują wszystkie miejsca w sekwencji oraz że zmieniają się w takim samym tempie. Założenie takie jest oczywiście błędne. Jeśli to założenie nie jest spełnione, metoda największej wiarygodności jest niespójna, czyli ma taką samą wadę, jak metoda parsymonii. Kiedy bowiem część miejsc pozostaje niezmiennych, to funkcja wiarygodności niedoszacowuje liczbę wielokrotnych podstawień. Tym samym źle wyliczona jest długość gałęzi i błędna filogeneza zostaje oceniona najwyżej. Aby temu zapobiec, można oszacować, jaka część miejsc w sekwencji jest silnie konserwowana i nie przyjmuje żadnych zmian. Można także określić rozkład tempa ewolucji w sekwencji.

Zwykle przyjmuje się, że rozkład ten przybiera postać tzw. rozkładu  $\gamma$ . Rozkład  $\gamma$  jest charakteryzowany przez współczynnik kształtu, określane zwykle literą  $\alpha$ . Na ryc. 6 przedstawiono rozkład częstości tempa podstawiania dla różnych wartości parametru  $\alpha$ . Kiedy  $\alpha$  jest niskie, np. 0,5, zauważymy, że najwięcej jest miejsc, które ewoluują wolno, tzn. tempo ich podstawień (na osi x) jest bliskie zeru. Są jednak także nieliczne miejsca, które ewoluują szybko. Im wyższa wartość  $\alpha$ , tym bardziej ujednotolica się tempo ewolucji. Na przykład dla wartości tego współczynnika równej 200, wszystkie pozycje są podstawiane w tempie zbliżonym do 1.

Wybór możliwości jest zatem bardzo duży - kilka podstawowych typów modeli, każdy z możliwością oszacowania (lub nie) miejsc niezmiennych oraz zróżnicowaniem rozkładu tempa podstawień (z różnym parametrem kształtu rozkładu). Na szacowanie filogenezy wpływają również częstości nukleotydów i tempa poszczególnych typów podstawień. Który model wybrać? Odpowiedź nie jest prosta.

Nie należy się kierować samą wartością funkcji wiarygodności. Drzewa uzyskane za pomocą bardziej złożonych modeli, czyli z większą liczbą parametrów, zawsze mają wyższą wartość tej funkcji niż drzewa bazujące na prostszych modelach. Bardziej złożone modele (z większą liczbą stopni swobody) są jednak wrażliwsze na błąd próby. Innym ograniczeniem jest czas obliczeń. Metoda największej wiarygodności jest najbardziej złożoną obliczeniowo metodą szacowania filogenezy. A im więcej parametrów, tym więcej obliczeń do wykonania.

Podobnie jak w wypadku innych modeli, można stosować do ich porównania test wiarygodności albo np. kryterium informacyjne Akaike'go. Trzeba zaznaczyć, że są to środki pomocnicze - badanie ewolucji nie poddaje się w pełni statystyce, dotyczy bowiem odtwarzania przeszłości.

### Literatura

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology* 27, 401-410.

Hendy, M.D. & Penny, D. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* 38, 297-309.

Swofford, D.L., Olsen, G.J., Waddell, P.J., & Hillis, D.M. 1996. Phylogenetic inference. In D. M. Hillis, C. Moritz, B.K. Mable (ed.), *Molecular systematics*. 2nd ed. 407-514. Sinauer Associates, Sunderland.