Biologia molekularna *in silico*.

Jerzy TYSZKIEWICZ

O porównywaniu sekwencji biologicznych

Biologia molekularna jest wielkim obszarem dla zupełnie nowych zastosowań matematyki i informatyki. Badania biologiczne nie dzielą się już, jak to było dawniej, na *in vivo* i *in vitro*, ale doszła trzecia możliwość: *in silico*. Z wielkiego bogactwa metod stosowanych w biologii obliczeniowej zdecydowałem się opisać Czytelnikom *Delty* kilka zagadnień dotyczących badania podobieństwa sekwencji biologicznych, w tym jeden ciekawy algorytm. Wiedza o podobieństwie takich sekwencji (białek, DNA czy genomów) jest pomocna w odtwarzaniu dziejów ewolucji.

Aby uświadomić sobie, jak jest to trudne, wyobraźmy sobie historyków, którzy by chcieli opisać dzieje Europy wyłącznie na podstawie wiedzy o współczesnych granicach między państwami. Takie jest właśnie nasze położenie np. przy badaniu ewolucji bakterii, które właściwie nie pozostawiają skamieniałości.

Podobieństwo białek

Patrząc na białko oczyma biologa molekularnego, widzimy przede wszystkim ciąg złożony z symboli; te symbole to 20 aminokwasów, które stanowią materiał budulcowy wszystkich białek we wszystkich znanych nam organizmach.

Białka zmieniają się w toku ewolucji – są różne mechanizmy, które powodują mutacje w sekwencjach DNA, które je kodują. W wyniku mutacji w białku wymianie mogą podlegać pojedyncze aminokwasy, jak też mogą ginąć lub pojawiać się całe fragmenty, które zostają wstawione/usunięte spomiędzy już istniejących fragmentów sekwencji.

W biologii obliczeniowej rozważa się podobieństwo dwóch sekwencji biologicznych, w tym przypadku – białek. Jest to liczba, która jest tym większa, im bliższe sobie są dwie sekwencje. Jeśli pewna grupa białek pochodzących od różnych organizmów jest wzajemnie do siebie podobna, to dokładne wartości podobieństw pozwalają odtworzyć w przybliżeniu drzewo ewolucyjne tych organizmów.

Miarą podobieństwa dwóch białek jest ważona liczba wymienionych wyżej operacji: wymian, usunięć i wstawień, niezbędnych do zamiany jednego białka w drugie. Zakładamy, że szansa zajścia każdej ze zmian jest niska, więc im więcej ich potrzeba, tym mniejsza szansa, że oba białka pochodzą od wspólnego przodka. *Ważona* oznacza, że nie wszystkie operacje „kosztują” tyle samo.

Bardziej formalnie, w celu porównania dwie sekwencje mogą być ze sobą *uliniowione*, co polega na wstawieniu do nich spacji (czyli wolnych miejsc, oznaczanych poniżej kreskami), by osiągnęły identyczną długość, napisaniu ich pod sobą i policzeniu kar i nagród. Najlepiej rozważyć to na przykładzie. W uliniowieniu $\begin{matrix}\mathrm{C}&\mathrm{N}&\mathrm{G}&-&\mathrm{T}\\-&\mathrm{N}&\mathrm{G}&\mathrm{V}&\mathrm{T}\end{matrix}$ mamy trzy nagrody za zgodność liter i dwie kary za wstawienie spacji. Uliniowienie to reprezentuje hipotezę, że w drodze od wspólnego przodka do obu aktualnych białek były dwa wydarzenia utraty/wstawienia aminokwasów. W uliniowieniu $\begin{matrix}\mathrm{A}&\mathrm{N}&\mathrm{G}&\mathrm{V}&\mathrm{V}\\-&\mathrm{N}&\mathrm{G}&\mathrm{V}&\mathrm{T}\end{matrix}$ mamy trzy nagrody za zgodność liter, jedną karę za niezgodność liter i jedną karę za wstawienie spacji; hipoteza zakłada trzy mutacje: dwie utraty/wstawienia oraz jedną wymianę aminokwasu.

Nagrody to punkty dodatnie, kary to punkty ujemne, a *wartość* uliniowienia to suma kar i nagród dla wszystkich miejsc. Liczba ta mierzy łatwość zajścia mutacji odpowiadających uliniowieniu. Dane nam są tylko sekwencje białek, a poszukujemy ich uliniowień o maksymalnych wartościach, czyli tych, które najłatwiej mogły zajść. Właśnie maksymalna wartość uliniowienia to miara podobieństwa tych dwóch białek.

Dla uproszczenia rozważań założymy, że wszystkie zgodne litery są nagradzane tak samo, wszystkie zamiany są karane tak samo, a wstawianie spacji jest tańsze od zamiany, ale już dwie spacje są droższe niż zamiana. Spacji naprzeciw spacji w ogóle nie wolno ustawiać.

Załóżmy, przykładowo, 1 za zgodność, -3 za niezgodność, -2 za spację. Wówczas pierwsze uliniowienie ma wartość -1, a drugie -2. (O tym, skąd się biorą wartości kar i nagród, powiemy poźniej.) Skoro miarą podobieństwa dwóch białek jest najlepszy wynik możliwy do osiągnięcia dla uliniowienia tych sekwencji, należy więc rozpatrzyć wszystkie możliwe uliniowienia, polegające na różnym sposobie wstawienia spacji, i wybrać to o najlepszym wyniku.

Fundamentalne znaczenie ma więc szukanie optymalnych uliniowień, a jest to trudne, bo jest ich, w przypadku dwóch sekwencji długości n każda, wykładniczo wiele i nawet już dla niewielkich wartości n szukamy małej igły w gigantycznym stogu siana. Na szczęście istnieje sprytny algorytm pozwalający ją szybko odnaleźć. Otóż uliniowienia można przedstawiać w postaci diagramów.

Intuicyjnie, każda kolumna i każdy wiersz w tabeli odpowiada stanowi po wypisaniu w sekwencji odpowiedniej litery opisującej wiersz/kolumnę. Nieopisany wiersz i kolumna odpowiadają stanowi przed wypisaniem pierwszego znaku. Przejścia do sąsiednich pól odpowiadają dopisywaniu znaków do sekwencji: ruch w prawo to dopisanie symbolu kolumny do pionowej sekwencji i „nic” (czyli spacji) do drugiej; w dół to, dualnie, dopisanie symbolu

wiersza do poziomej sekwencji i „nic” do pionowej; w końcu ukośnie w prawo w dół dopisuje po jednej literze do obu sekwencji. Widać, że drogi w tabelce odpowiadają dokładnie możliwym uliniowieniom.

Kluczem do algorytmu jest pomysł, że w pola tabeli wpisujemy dwie informacje: wartość optymalnego uliniowienia początkowych fragmentów sekwencji, które kończą się na literach opisujących kolumnę i wiersz pola, oraz informację, skąd do tego pola przyszliśmy, aby ten optymalny koszt uzyskać.

Wartość uliniowienia całych sekwencji znajdziemy w prawym dolnym narożniku tabeli.

Prześledźmy to na przykładzie. Najpierw wpisujemy do pierwszego wiersza i pierwszej kolumny wartości optymalnych uliniowień, które kończą się w danym polu tabeli. Te wartości łatwo obliczyć, bo to koszty wstawiania coraz dłuższych ciągów samych spacji.

		C	N	G	T
	0	$-2_{\leftarrow}$	$-4_{\leftarrow}$	$-6_{\leftarrow}$	$-8_{\leftarrow}$
N	$-2_{\uparrow}$				
G	$-4_{\uparrow}$				
V	$-6_{\uparrow}$				
T	$-8_{\uparrow}$				

		C	N	G	T
	0	$-2_{\leftarrow}$	$-4_{\leftarrow}$	$-6_{\leftarrow}$	$-8_{\leftarrow}$
N	$-2_{\uparrow}$	$-3_{\nwarrow}$	$-1_{\nwarrow}$	$-3_{\leftarrow}$	$-5_{\leftarrow}$
G	$-4_{\uparrow}$	$-5_{\nwarrow,\uparrow}$			
V	$-5_{\uparrow}$				
T	$-6_{\uparrow}$				

		C	N	G	T
	0	$-2_{\leftarrow}$	$-4_{\leftarrow}$	$-6_{\leftarrow}$	$-8_{\leftarrow}$
N	$-2_{\uparrow}$	$-3_{\nwarrow}$	$-1_{\nwarrow}$	$-3_{\leftarrow}$	$-5_{\leftarrow}$
G	$-4_{\uparrow}$	$-5_{\nwarrow,\uparrow}$	$-3_{\uparrow}$	$0_{\nwarrow}$	$-2_{\leftarrow}$
V	$-6_{\uparrow}$	$-7_{\nwarrow,\uparrow}$	$-5_{\uparrow}$	$-2_{\uparrow}$	$-3_{\nwarrow}$
T	$-8_{\uparrow}$	$-9_{\nwarrow,\uparrow}$	$-7_{\uparrow}$	$-4_{\uparrow}$	$-1_{\nwarrow}$

W narożnym polu wpisujemy 0, bo odpowiada ono uliniowieniu pustego ciągu z pustym ciągiem. Teraz wpisujemy wartości do drugiego wiersza. W polu (N,C) wpisujemy -3 za przyjście z pola narożnego, bo przyjście z lewej i z góry daje wartość -4, czyli gorszą, i odnotowujemy, skąd przyszlśmy. W (N,N) wpisujemy -1 za przyjście z (N,C), bo odbieramy nagrodę 1 za identyczność liter, w (N,G) dostajemy -3 za przyjście z (N,N), a w (N,T) mamy -5 przychodząc z (N,G).

Zabieramy się za kolejny wiersz: w (G,C) mamy -5, ale tam można przyjść na dwa sposoby, uzyskując tyle właśnie punktów. Potem uzupełniamy analogicznie kolejno następne wiersze.

Poczynając teraz od pola (T,T) i idąc pod prąd strzałek, odczytujemy wspak jedyną najlepszą drogę dojścia: (N,C),(N,N),(G,G),(V,G), (T,T) (jedyną w tym przypadku; w ogólności optymalnych dróg może być więcej). Zgadza się ona z uliniowieniem podanym w przykładzie.

Widać natychmiast, że szukanie optymalnych uliniowień tą metodą wspaniale nadaje się do powierzenia komputerowi, który staje się w badaniach historii życia równie ważny jak pipeta i szalka Petriego.

Ewolucja i tablice substytucyjne

Gen, czyli sekwencja DNA, która koduje białko, ma specjalną strukturę: składa się z trójek symboli spośród A, C, G i T (oznaczają one adeninę, cytozynę, guaninę i tyminę); taka trójka to kodon. Każdy kodon wyznacza jeden z 20 aminokwasów, z których utworzone zostanie potem kodowane przez gen białko. Zmiany w DNA zachodzą z grubsza losowo. Jednak zmiany w białkach są w znacznym stopniu hamowane przez mechanizmy ewolucji. Jeśli w genie zajdzie jakaś mutacja powodująca poważną zmianę kodowanego białka, to może ono utracić swoją biologiczną rolę, wystawiając swego nosiciela na podwyższone ryzyko śmierci, utrudniając tym samym przekazanie mutacji na potomków. Ten wpływ ewolucji powoduje zróżnicowanie szybkości mutacji różnych aminokwasów: zamiana X na Y może być łatwo utrwalana, gdy X i Y są podobne, a bardzo trudno, gdy się znacznie różnią. Zatem szansa na zamianę X na Y zależy w istocie i od X, i od Y.

Do badania podobieństwa białek stosuje się więc w rzeczywistości uliniowienie, w którym nagrody za zgodność i kary za niezgodność aminokwasów zależą od tego, które konkretnie są to aminokwasy. Wartości te bierze się ze specjalnych *tablic substytucyjnych*, których opracowanie jest osobnym, poważnym zadaniem badawczym. Istnieje wiele tablic o ugruntowanym znaczeniu, jak PAM i BLOSUM, ale ciągle opracowuje się nowe, np. dostosowane do porównywania specyficznych grup białek. Również kara za użycie spacji jest zwykle inna niż w naszym uproszczonym modelu – zależy od ilości spacji wstawianych obok siebie. Natomiast sam algorytm nie różni się co do swojej głównej idei od tego, który dla uproszczonego przypadku opisałem powyżej. Szczegóły są, oczywiście, trochę inne.

Podobieństwo białek to narzędzie do odtwarzania ich pochodzenia ewolucyjnego, ale ma też inne zastosowania: jednym z najważniejszych jest przewidywanie struktury przestrzennej białek. Otóż liniowa cząsteczka białka, zaraz po jej syntezie, zaczyna się spontanicznie zwijać do wysoce

skomplikowanej struktury przestrzennej. Dopiero po jej osiągnięciu białko może pełnić swoją biologiczną funkcję. Przewidywanie tej struktury na podstawie znajomości tylko sekwencji to wielkie wyzwanie. Gdybyśmy potrafili to robić, to, oprócz wielu innych rzeczy, moglibyśmy też konstruować nowe lekarstwa w komputerze, bez trudnych i czasochłonnych badań „w probówce". Podobieństwo pomaga tu trochę: białka o podobnych sekwencjach często zwijają się do podobnych struktur przestrzennych, a nawet orientacyjna sugestia co do kształtu cząsteczki jest ogromnie ważna przy próbach jej precyzyjnej rekonstrukcji, bo chroni nas przed szukaniem „na manowcach".

DNA

Mechanizm ewolucyjny hamowania mutacji działa dużo słabiej na poziomie DNA, bo kodonów jest 64, a aminokwasów tylko 20, więc jeden aminokwas jest zwykle kodowany przez kilka różnych kodonów (maksymalnie do 6). Mutacje DNA nie zmieniające kodowanego białka są dla ewolucji oczywiście obojętne, więc nie są też przez nią hamowane. Można uliniawiać białka, można też i sekwencje DNA. Ale, wobec obojętności ewolucji na niektóre zmiany DNA, nie istnieją sensowne tablice substytucyjne dla DNA, a wyniki uliniowienia sekwencji DNA są dużo trudniejsze do zinterpretowania. Dlatego też większość badań nad ewolucją opiera się na uliniowieniu białek. Jednak dla fragmentów DNA, które nie kodują białek, porównywanie DNA to jedyna droga do odtworzenia ich historii.

Genomy

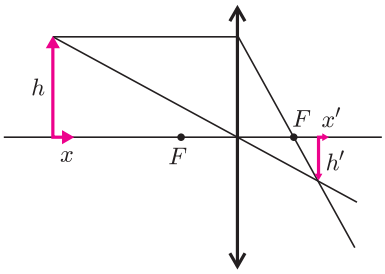
Na całej nici DNA niektóre geny są na tyle podobne, że można je praktycznie uważać za kopie tego samego genu. (Oczywiście podobieństwo rozumiemy jako istnienie uliniowienia o wysokiej wartości dla kodowanych przez nie białek – wiemy już, jak to stwierdzić.) Czyli teraz cała nić prezentuje nam się jako ciąg symboli, którymi są poszczególne geny. To kolejna sekwencja biologiczna, którą można analizować. Cały genom (bo tak będziemy mówić o sekwencji genów) też podlega zmianom. Jednak nie zachodzą w nim lokalne, punktowe mutacje, a raczej fundamentalne przebudowy całej sekwencji. Możliwe jest wycięcie całego fragmentu genomu i wklejenie go z powrotem w to samo miejsce, ale w odwrotnej kolejności. Podobnie, wycięty fragment może być przeniesiony w inne miejsce i włączony w oryginalnej lub odwrotnej kolejności.

Miarą podobieństwa dwóch genomów jest liczba operacji jednego z wybranych typów, niezbędnych do przekształcenia jednego z nich w drugi. Miary podobieństwa tego typu są wykorzystywane do odtwarzania drzew ewolucyjnych organizmów, opartych na całych genomach. Niestety, algorytmy używane do obliczania podobieństwa genomów są dużo bardziej skomplikowane niż dla DNA, a niektóre z problemów obliczeniowych są wręcz NP-trudne (patrz artykuł Damiana Niwińskiego w *Delcie* 7/2000). W końcu może się też pojawić kolejny problem: drzewa ewolucyjne organizmów zrekonstruowane na podstawie białek nie zawsze pokrywają się z opartymi na genomach. Co z tym zrobić, to już raczej temat na kolejny artykuł...

Zadania

Redaguje Ewa CZUCHRY

F 581. Znaleźć i porównać powiększenia podłużne $\alpha = x'/x$ i poprzeczne $\beta = h'/h$ dla cienkiej soczewki.



Rozpatrzyć przypadek przedmiotu o małych wymiarach podłużnych.
Rozwiązanie na str. 13

F 582. Na osi optycznej soczewki, w odległości równej dwóm ogniskowym, umieszczono kulkę. Jaką postać ma obraz tego przedmiotu?
Rozwiązanie na str. 13

Redaguje Mikołaj ROTKIEWICZ

M 1003. W probówce znajduje się 9 bakterii typu A i 11 typu B. W każdej sekundzie losowo wybrana bakteria dzieli się na dwie takie same. Po 60 sekundach losujemy bakterię z probówki. Obliczyć prawdopodobieństwo tego, że będzie to bakteria typu A.
Rozwiązanie na str. 3

M 1004. W probówce znajdują się bakterie typu A i B. W każdej sekundzie dochodzi do podziału losowo wybranej bakterii. Bakterie typu A i B dzielą się odpowiednio na m i n identycznych bakterii, $m > n > 0$. Niech p_k oznacza prawdopodobieństwo tego, że losowo wybrana w k-tej sekundzie bakteria jest typu A. Zbadać monotoniczność ciągu p_k.
Rozwiązanie na str. 8

M 1005. W pewnej rodzinie mąż i żona zawarli następującą umowę:
- żona zmywa naczynia zawsze w dwa kolejne dni, po czym zmywać musi mąż,
- jeśli któregoś dnia naczynia zmywał mąż, to o tym, kto będzie zmywał następnego dnia, decyduje rzut monetą,
- pierwszego dnia zmywanie naczyń rozpoczyna mąż.

Niech p_k będzie prawdopodobieństwem tego, że w k-tym dniu obowiązywania umowy naczynia zmywa mąż.
Oblicz $\lim_{k\to\infty} p_k$.
Rozwiązanie na str. 9