Streszczenie rozprawy doktorskiej pod tytułem

“Wykrywanie horyzontalnego transferu genów”

Paweł Górecki

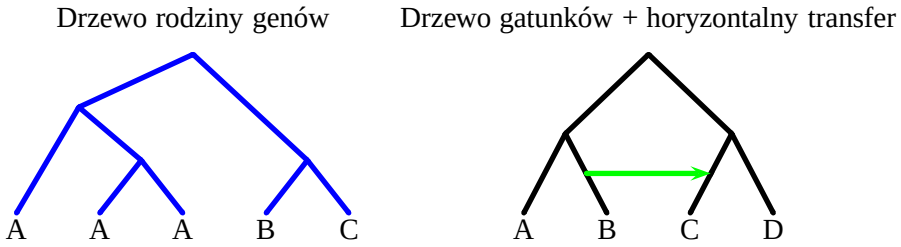
1 Wprowadzenie

Tematem niniejszej rozprawy są zagadnienia z pogranicza biologii molekularnej, matematyki i informatyki dotyczące własności modeli ewolucyjnych uwzględniających horyzontalny transfer genów (w skrócie HGT) i praktycznych metod jego wykrywania.

Współczesnie metody rekonstrukcji drzew ewolucyjnych gatunków są następujące: korzystając ze znanych sekewncji DNA lub sekwencji aminokwasów oblicz drzewa ewolucyjne dla rodzin genów (nazywane *drzewami genów*), a następnie znajdź drzewo ewolucyjne gatunków (nazywane *drzewem gatunków*) “najbardziej podobne w pewnych sensie” do tego zbioru drzew rodzin genów. Trudność tego zadania polega na możliwych różnicach pomiędzy drzewami genów, które wynikają z prostego faktu, że ewolucja genów zwykle przebiega inaczej niż ewolucja gatunków. Stąd mamy dwa ważne problemy:

- rekonstrukcja drzewa gatunków ze zbioru drzew genów,
- uzgodnienie danego drzewa gatunków z danym drzewem genów.

W tej rozprawie zajmujemy się głównie tym drugim problemem, który można nieformalnie nazywać *wbudowywaniem drzewa (genów) w drzewo (gatunków)*. To pozornie dziwaczne sformułowanie ma swoje uzasadnienie biologiczne: gatunki można w pewnym sensie traktować jak “pojemniki” dla genów, co przy uwzględnieniu zależności ewolucyjnych (wyrażanych w postaci drzew) daje wyżej wspomniane wbudowywanie drzew ewolucyjnych.

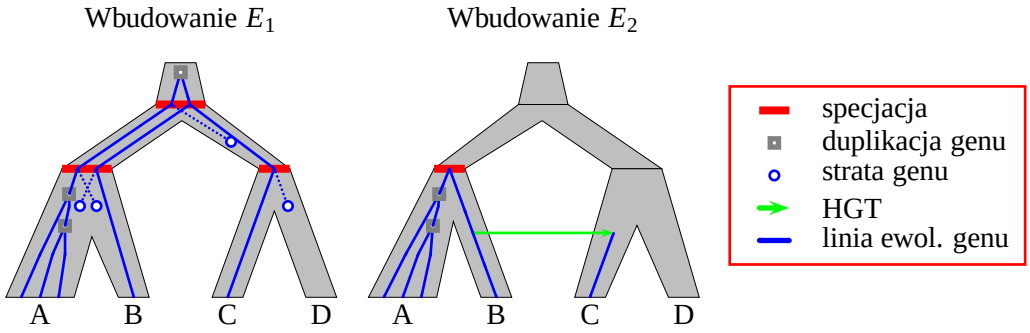


Rysunek 1: Jak uzgodnić te drzewa?

Rozważmy przykład problemu uzgadniania przedstawiony na Rysunku 1. Liście grafów z tego rysunku są etykietowane nazwami gatunków *A*, *B*, *C* i *D*. Prawe drzewo posiadające jeden horyzontalny transfer, jest nazywane w tej pracy *grafem gatunków*. Ponadto mamy drzewo ewolucyjne rodziny 5 genów (lewe drzewo). Zauważmy, że etykiety liści drzewa genów są nazwami gatunków. Oznacza to, że

odpowiednia sekwencja genu pochodzi od gatunku o danej nazwie. Zatem, mamy tu 3 geny pochodzące od gatunku A, po jednym od B i C, natomiast gatunek D nie ma genów reprezentowanych w tej rodzinie (być może ten gatunek posiada geny w tej rodzinie ale są jeszcze nieznane). Zauważmy, że HGT na drzewie gatunków oznacza pewną hipotezę, która może być wykorzystana do odszukania pewnego scenariusza ewolucyjnego uzgadniającego te dwa drzewa.

Na Rysunku 2 przedstawiamy dwa wbudowania drzewa genów w drzewo gatunków z transferem genów. Można pokazać, że E_1 jest minimalnym wbudowaniem o koszcie[1] 7 (4 straty + 3 duplikacje genów) w tzw. modelu duplikacji i strat, w którym nie uwzględnia się HGT. Analogicznie, można pokazać, że E_2 jest minimalnym wbudowaniem o koszcie 3 (1 HGT + 2 duplikacje) dla modelu z transferem genów. Te dwa powyższe modele są obiektem badań tej rozprawy.



Rysunek 2: Przykłady wbudowań drzewa genów w drzewo gatunków (z Rys. 1)

W pierwszej części tej pracy przedstawiony jest nowy model drzew ewolucyjnych, nazywanych DLS-drzewami, w którym uwzględniane są duplikacje i straty genów (ewolucja genów) oraz specjacje (ewolucja gatunków). DLS-drzewa reprezentują ewolucję genów w kontekście ewolucji gatunków. Z DLS-drzewa można odtworzyć wbudowania (Rysunek 2).

Jest to model bazowy, który w następnej części będzie uzupełniony o horyzontalny transfer genów. W tej części pokazujemy wiele ciekawych własności tego modelu. W szczególności pokazujemy związki DLS-drzew z drzewami uzgadniającymi (Page, 1994). Dzięki tym związkom rozwiązujemy niektóre otwarte problemy postawione w (Page and Charleston, 1997b) dla drzew uzgadniających.

W drugiej części tej pracy przedstawiony jest nowy model drzew ewolucyjnych, nazywanych H-drzewami, w którym uwzględniane są duplikacje, straty i horyzontalny transfer genów (ewolucja genów) oraz specjacje (ewolucja gatunków). H-drzewa są rozszerzeniem DLS-drzew. Jednakże ze względu na transfer, który umożliwia dodatkowe warianty uzgadniania, ich struktura jest znacznie bardziej skomplikowana i ciekawsza. Analogicznie do DLS-drzew, H-drzewa można przekształcać w kontekście grafu gatunków do wbudowań (tak jak na Rysunku 2). W tej części pokazujemy analogiczne własności tego modelu. W szczególności pokazujemy związki H-drzew z drzewami uzgadniającymi z transferem (Górecki, 2004).

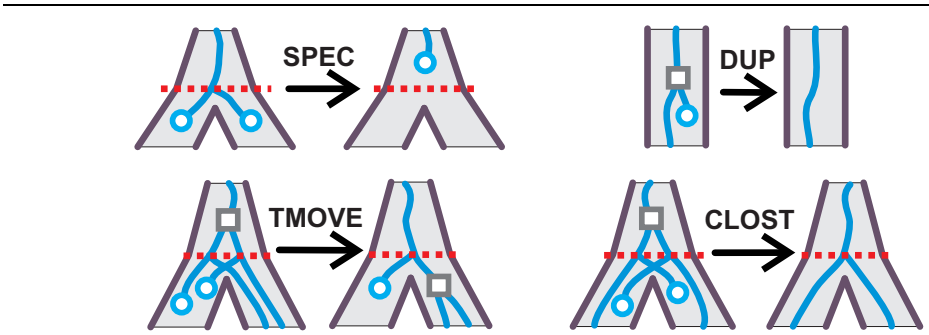
[1]Koszt będący sumą duplikacji i strat jest nazywany kosztem mutacyjnym.

W rozdziale 4 przedstawiony jest algorytm o wielomianowej złożoności czasowej i pamięciowej, obliczający dla danego drzewa genów i grafu gatunków tzw., *minimalny ważony koszt* H-drzewa. W szczególności jest to algorytm pozwalający na weryfikowanie hipotez horyzontalnego transferu genów.

W następnej części streszczenia przedstawiamy bardziej szczegółowo wyniki z poszczególnych rozdziałów.

2 DLS-drzewa

W rozdziale 2 pracy przedstawiamy nowy model drzew ewolucyjnych oparty na koncepcjach modelu duplikacji i strat (bez transferu genów). Przedstawiamy definicję DLS-drzewa, które jest modelem ewolucji drzewa genów w kontekście drzewa gatunków przy założeniu, że dopuszczalne są tylko duplikacje i straty genów (ewolucja genów) oraz specjacje (ewolucja gatunków). DLS-drzewa pozwalają na modelowanie dowolnych scenariuszy ewolucyjnych, nie tylko tych minimalnych (por. drzewa uzgadniające (Page, 1994)). Pokazujemy jak z DLS-drzewa można otrzymać drzewo genów i drzewo gatunków. Definiujemy system przepisywania DLS-drzew zawierający 4 reguły: SPEC, DUP (typu I, usuwające "proste" wystąpienia strat genów) oraz CLOST, TMOVE (typu II). Ich interpretacje biologiczne przedstawione są na Rysunku 3 (kwadrat oznaczają duplikację, kółko stratę, a przerywana linia specjację). Zauważmy, że redukcja zmniejsza koszt mutacyjny i rozmiar drzewa.



Rysunek 3: Interpretacje biologiczne reguł dla DLS-drzew

Będziemy mówić, że dwa DLS-drzewa są równoważne jeśli jedno może być przekształcone w drugie przez zastosowanie zero lub większej liczby redukcji w dowolnym kierunku. Ten system posiada znaczące własności biologiczne i matematyczne. W szczególności przekształcanie zachowuje drzewa genów i gatunków. System posiada własność konfluencji i silnej normalizacji, w szczególności w każdej klasie równoważnych DLS-drzew istnieje jednoznaczna postać normalna (tzn. nieredukowalne DLS-drzewo), łatwo osiągalna przez redukcje w naszym systemie. Łatwo stąd wynika, że koszt mutacyjny i rozmiar drzewa w postaci normalnej jest minimalny w klasie równoważnych DLS-drzew (patrz także Twierdzenie 1).

Oprócz drzew w postaci normalnej, analizujemy także ogólne własności DLS-drzew. W klasie równoważnych drzew wyróżniamy ważny podzbiór drzew semi-normalnych, tzn. niezawierających redeksów reguł typu I. Pokazujemy, że równoważne drzewa semi-normalne można przedstawić w skończonego postaci diagramu (DAG-u), gdzie krawędziami są redukcje typu II. Pokazujemy, że system z odwróconymi regułami typu II jest silnie normalizowanly, a jednoznaczną postacią normalną w takim systemie reguł jest *drzewo tłuste*, które wszystkie duplikacje ma w "czubku". Zatem, każdy taki diagram drzew semi-normalnych posiada jeden korzeń (drzewo tłuste) oraz jeden liść (postać normalną). W szczególności, z naszej analizy wynika, że minimalny koszt duplikacyjny (czyli liczba duplikacji) w klasie drzew równoważnych może posiadać także drzewo nie będące w postaci normalnej.

Pokazujemy jak z danego drzewa genów i danego drzewa gatunków rekonstruować drzewa w postaci normalnej.

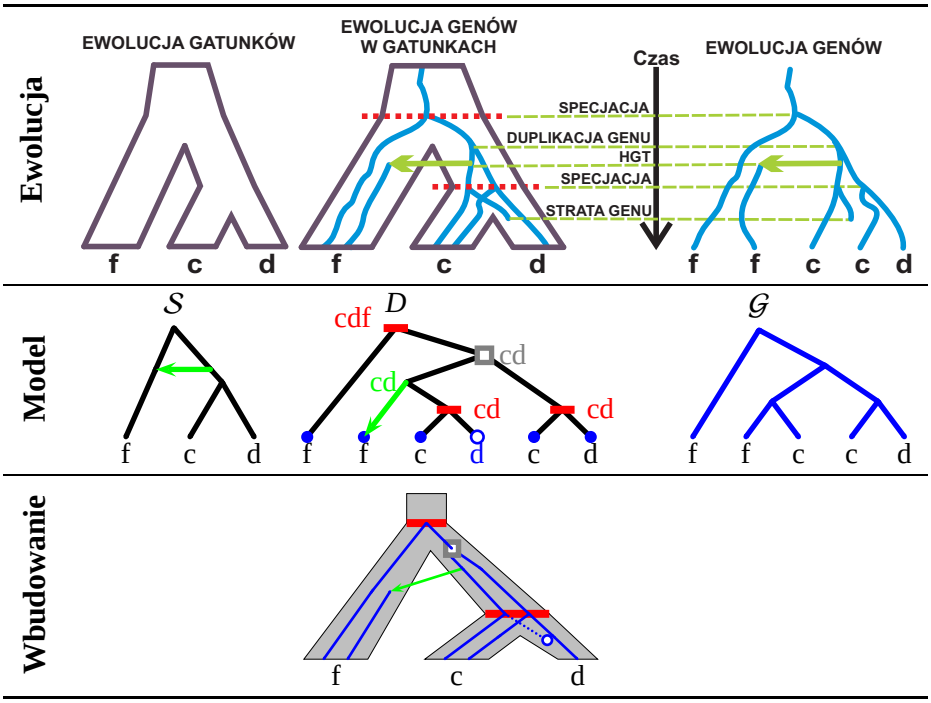
W ostatniej części tego rozdziału pokazujemy związki DLS-drzew z modelem duplikacji i strat. Ten model, zwany też modelem drzew uzgadniających, jest stosunkowo dobrze poznanym (Page, 1994; Mirkin *et al.*, 1995; Eulenstein and Vingron, 1998; Zhang, 1997; Page and Charleston, 1997a; Eulenstein *et al.*, 1998; Page and Charleston, 1997b; Ma *et al.*, 1998; Bonizzoni *et al.*, 2003; Górecki and Tiuryn, 2005). Drzewo uzgadniające, będące kluczowym pojęciem tego modelu, zdefiniowane przez Page'a (1994) było traktowane jako minimalne w sensie rozmiaru, kosztu etc. Jednakże przez wiele lat nie było dowodu tych własności, które jako otwarte problemy były postawione w pracy (Page and Charleston, 1997b). Dopiero w pracy (Bonizzoni *et al.*, 2003) autorzy rozwiązują problem dotyczący minimalizacji drzewa uzgadniającego w sensie rozmiaru. W ostatniej części rozdziału o DLS-drzewach pokazujemy, że drzewo uzgadniające można łatwo przekształcić do DLS-drzewa w postaci normalnej. W szczególności, z wyników przedstawionych w tym rozdziale rozwiązujemy problemy przedstawione w (Page and Charleston, 1997b) dla drzew uzgadniających.

3 H-drzewa

W rozdziale 3 pracy przedstawiamy nowy model drzew ewolucyjncych oparty na koncepcjach modelu duplikacji, strat i horyzontalnego transferu genów (Charleston, 1998; Hallett and Lagergren, 2001; Addario-Berry *et al.*, 2003; Górecki, 2003, 2004; Hallett *et al.*, 2004). W pierwszej części definiujemy pojęcie grafu gatunków, które jest modelem ewolucji gatunków z horyzontalnymi transferami genów. Nieformalnie możemy napisać: *graf gatunków = drzewo gatunków + horyzontalne transfery genów* (patrz Rysunek 1). Jednak nie wszystkie transfery są poprawne - nie mogą one naruszać czasowych zależności, np. niedozwolone jest krzyżowanie w czasie. Analiza tych własności wykazuje, że dla każdego grafu gatunków istnieje relacja częściowego porządku na transferach nazywana *relacją zależności*. Analizujemy też sytuację odwrotną, tzn. gdy zbiór transferów jest określony i odpowiadamy na pytanie "*Kiedy możemy dodać go do drzewa gatunków tak by powstał graf gatunków?*".

Przedstawiamy definicję H-drzewa, które jest modelem ewolucji drzewa genów w kontekście grafu gatunków przy założeniu, że dopuszczalne są tylko duplikacje, straty i horyzontalny transfer genów (ewolucja genów) oraz specjacje (ewolucja gatunków). H-drzewo jest rozszerzeniem DLS-drzewa. Ale oprócz transferów, H-drzewo posiada dodatkowo relację częściowego porządku na transferach nazywaną *horyzontalną zależnością*.

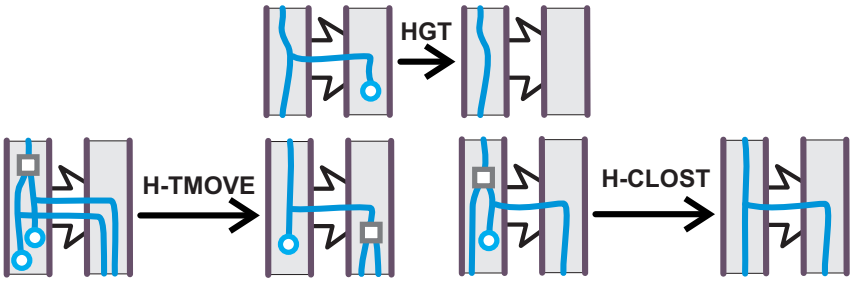
Na Rysunku 4 pokazujemy jak przykładowa ewolucja trzech gatunków i pięciu genów pochodzących od tych gatunków jest reprezentowana w modelu H-drzew (w modelu DLS-drzew sytuacja jest analogiczna ale bez HGT).



Rysunek 4: Przykładowa ewolucja genów i gatunków, interpretacja w modelu (graf gatunków $\mathcal{S}$, H-drzewo D, drzewo genów $\mathcal{G}$) oraz wbudowanie

H-drzewa, podobnie jak DLS-drzewa, pozwalają na modelowanie dowolnych scenariuszy ewolucyjnych, nie tylko minimalnych. Pokazujemy jak z H-drzewa otrzymać drzewo genów (takie drzewo genów będzie nazywane *zgodnym* z danym H-drzewem). Ale okazuje się, że H-drzewo może reprezentować ewolucję genów w kontekście wielu grafów gatunków (co jest zgodne z interpretacją biologiczną H-drzewa). Określamy pojęcie *zgodnego* grafu gatunków wykorzystując m.in. horyzontalną zależność i relację zależności. Nieformalnie można napisać, że H-drzewo jest zgodne z danym grafem gatunków jeśli istnieje interpretacja dla tego H-drzewa w postaci wbudowania w ten graf gatunków (por. Rysunki 2 i 4).

Analogiczne do reguł dla DLS-drzew, definiujemy system przepisywania H-drzew zawierający 7 reguł. Oprócz reguł z Rysunku 3 dodajemy jedną regułę typu I: HGT oraz dwie reguły typu II: H-CLOST i H-TMOVE. Ich interpretacje biologiczne przedstawione są na Rysunku 5.



Rysunek 5: Interpretacje biologiczne dodatkowych reguł dla H-drzew

Dowodzimy analogicznych własności dla H-drzew: zachowywanie przez redukcję drzewa genów oraz zbioru zgodnych grafów gatunków, konfluencja, silna normalizacja. Definiujemy drzewa tłuste, semi-normalne i analizujemy ich diagramy. Otrzymujemy analogiczne wyniki dotyczące minimalizacji kosztów w klasie równoważnych H-drzew (Wniosek 42 z Twierdzenia 41):

Twierdzenie 1 *Dla H-drzewa T, nieredukowalne H-drzewo otrzymane z T przez redukcje systemu jest jednoznacznym H-drzewem o najmniejszym koszcie (liczonym jako suma horyzontalnych transferów, duplikacji i strat genów) w klasie wszystkich H-drzew równoważnych T.*

Pokazujemy jak z danego drzewa genów $\mathcal{G}$ i danego grafu gatunków $\mathcal{S}$ rekonstruować drzewa w postaci normalnej. W tym celu wprowadzamy pojęcie scenariusza (Górecki, 2004), który odpowiada wyborowi pewnego schematu użycia transferów z grafu gatunków. Okazuje się, że dla każdego scenariusza ξ istnieje H-drzewo Ψ_ξ (efektywnie konstruowalne) w postaci normalnej zgodne z $\mathcal{G}$ oraz $\mathcal{S}$. Ponadto, pokazujemy także (patrz Twierdzenie 81):

Twierdzenie 2 *Niech $\mathcal{G}$ będzie drzewem genów, a $\mathcal{S}$ grafem gatunków, takim że wszystkie gatunki z $\mathcal{G}$ występują w $\mathcal{S}$. Wtedy, dla każdego drzewa T zgodnego z $\mathcal{G}$ i $\mathcal{S}$, istnieje scenariusz ξ, taki że albo $\Psi_\xi = T$ albo T jest poddrzewem Ψ_ξ.*

W ten sposób otrzymujemy pełną klasyfikację, a także sposób na otrzymywanie H-drzew w postaci normalnej zgodnych z danym drzewem genów i danym grafem gatunków. Z Twierdzenia 1 wynika, że poszukiwanie H-drzew minimalizujących koszt powinno być ograniczone do drzew w postaci normalnej. Ponadto, z Twierdzenia 2 otrzymujemy, że powinniśmy rozważać tylko te H-drzewa w postaci normalnej, które nie zawierają poddrzew w postaci normalnej zgodnych z $\mathcal{G}$ i $\mathcal{S}$.

W ostatniej części tego rozdziału pokazujemy związki z drzewami uzgadniającymi (Górecki, 2004).

4 Algorytm i przykłady

W rozdziale 4 przedstawiony jest algorytm o wielomianowej złożoności czasowej i pamięciowej, obliczający dla danego drzewa genów $\mathcal{G}$ i grafu gatunków $\mathcal{S}$ mini-

malny *ważony koszt* H-drzewa w zbiorze H-drzew zgodnych z $\mathcal{G}$ i $\mathcal{S}$. Ten algorytm może być łatwo rozszerzony do wersji generującej optymalne (tzn., posiadające ten minimalny ważony koszt) H-drzewo. W szczególności jest to algorytm pozwalający na weryfikowanie hipotez horyzontalnego transferu genów. Oprócz algorytmu przedstawiamy prosty biologiczny przykład i kilka sztucznych przykładów (w dodatku).

5 Podsumowanie

Wyniki zawarte w tej pracy mogą być stosowane nie tylko do weryfikowania lub wykrywania hipotez HGT ale także w innych systemach posiadających cechy "drzew w drzewach" (w naszym przypadku mamy system typu *gen-gatunek*). Np. w biogeografii (system typu *organizm-obszar*), gdzie transfer odpowiada przemieszczeniu się organizmów na nowy obszar (Nelson and Platnick, 1981; Swenson *et al.*, 2001; Arvestad *et al.*, 2004), w parazytologii (system typu *gospodarz-pasożyt*), gdzie transfer odpowiada przemieszczeniu się pasożyta na nowy gatunek (Page, 1993, 1994).

W ramach prac nad rozprawą powstał system komputerowy do obliczania i wizualizacji wyników, w którym zaimplementowano wszystkie algorytmy przedstawione w pracy. Wkrótce będzie dostępny w sieci.

Literatura

Addario-Berry, L., Hallett, M., and Lagergren, J., 2003. Towards identifying lateral gene transfer events. In *Pacific Symposium on Biocomputing*, 279–290.

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B., 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *RECOMB 2004*.

Bonizzoni, P., Vedova, G., and Dondi, R., 2003. Reconciling gene trees to a species tree. *Algorithms and Complexity, Proceedings of the 5th Italian Conference (CIAC 2003)* 2653, 120–131.

Charleston, M. A., 1998. Jungles: A new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences* 149, 191–223.

Eulenstein, O., Mirkin, B., and Vingron, M., 1998. Duplication-based measures of difference between gene and species trees. *Journal of Computational Biology* 5, 135–148.

Eulenstein, O. and Vingron, M., 1998. On the equivalence of two tree mapping measures. *DAMATH: Discrete Applied Mathematics and Combinatorial Operations Research and Computer Science* 88.

Górecki, P., 2003. Single step reconciliation algorithm for duplication, loss and horizontal gene transfer model. In *Proceedings of ECCB 2003*. Paris.

Górecki, P., 2004. Reconciliation problems for duplication, loss and horizontal gene transfer. In *RECOMB 2004*, 316–325. San Diego.

Górecki, P. and Tiuryn, J., 2005. On the structure of reconciliations. *LNCS* 3388, 42–54.

Hallett, M. and Lagergren, J., 2001. Efficient algorithms for lateral gene transfer problems. In *RECOMB 2001*, 149–156. ACM Press, New York.

Hallett, M., Lagergren, J., and Tofigh, A., 2004. Simultaneous identification of duplications and lateral transfers. In *RECOMB 2004*, 316–325. San Diego.

Ma, B., Li, M., and Zhang, L., 1998. On reconstructing species trees from gene trees in term of duplications and losses. In *RECOMB 1998*, 182–191.

Mirkin, B., Muchnik, I., and Smith, T. F., 1995. A biologically consistent model for comparing molecular phylogenies. *J. of Comput. Biol.* 2, 493–507.

Nelson, G. and Platnick, N. I., 1981. *Systematics and Biogeography: Cladistics and Vicariance*. Columbia University Press, New York.

Page, R., 1993. Parasites, phylogeny and cospeciation. *International Journal of Parasitology* 23, 449–506.

Page, R. D. M., 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology* 43, 58–77.

Page, R. D. M. and Charleston, M. A., 1997a. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* 7, 231–240.

Page, R. D. M. and Charleston, M. A., 1997b. Reconciled trees and incogruent gene and species trees. *Mathematical Hierarchies and Biology, DIMACS Series in Mathematics and Theoretical Computers Science* 37.

Swenson, U., Backlund, A., McLoughlin, S., and Hill, R. S., 2001. *Nothofagus* biogeography revisited with special emphasis on the enigmatic distribution of subgenus *Brassosphora* in New Caledonia. *Cladistics* 17, 28–47.

Zhang, L., 1997. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology* 4, 177–188.