Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
Instytut Informatyki

Autoreferat rozprawy doktorskiej

Algorytmy modelowania ewolucji stochastycznych systemów genetycznych o dużej złożoności

mgr inż. Tomasz Wojdyła

Promotor: prof. dr hab. inż. Marek Kimmel

Gliwice, październik 2011

1 Wprowadzenie

Pytanie o pochodzenie człowieka jest jednym z najważniejszych pytań nurtujących ludzi od dawna. Pierwsze naukowe podstawy w próbie odpowiedzi na to pytanie zostały sformułowane dopiero w połowie XIX wieku przez takich uczonych jak Charles Darwin [12], Alfred Wallace [13] oraz Gregor Mendel [35]. Od tego czasu naukowcy próbują zidentyfikować i poznać wszystkie aspekty mechanizmów stojących za ewolucją organizmów. Szczególnie intensywny wzrost zainteresowania genetyką pojawił się w latach 80-tch XX wieku wraz z wprowadzeniem powszechnego dostępu do technik informatycznych oraz rozwojem nowoczesnej biologii (np. metod sekwencjonowania DNA). Obecny rozwój genetyki stymulowany jest przez coraz to większe nakłady finansowe. Naukowcy mają do dyspozycji dane genetyczne pochodzące od tysięcy osobników (nie tylko ludzkich). Dane te są rezultatem wielu projektów prowadzonych na całym świecie (np. The Human Genome Project [26], The International HapMap Project [9] lub The 1000 Genome Project [10]). Pomimo tego, iż nasza wiedza o podstawowych siłach genetycznych i interakcjach występujących między nimi znacznie wzrosła od czasów Darwina, olbrzymia złożoność tychże procesów wciąż stanowi przeszkodę w poznaniu odpowiedzi na bardziej szczegółowe pytania. Wykorzystując wyniki obserwacji rzeczywistych osobników możemy jednak konstruować modele genetyczne, które, w ramach przyjętych założeń i ograniczeń, pozwolą nam wyjaśnić zjawiska występujące w procesie ewolucji.

Możemy wyróżnić dwa główne kryteria podziału stochastycznych modeli ewolucyjnych. Ze względu na perspektywę czasu modele dzielą się na retrospektywne (ang. backward-time) oraz prospektywne (ang. forward-time). W koncepcyjnie prostszych modelach prospektywnych zwykle symulowana jest cała populacja od wybranego momentu w przeszłości do czasu aktualnego. Wyniki uzyskiwane są zwykle na podstawie próbki (ang. sample) wylosowanej z ostatniego pokolenia. Modele retrospektywne bazują na teorii koalescencji (ang. coalescent theory) [25, 31]. W odróżnieniu do metod prospektywnych, w modelach retrospektywnych uwaga skupiona jest jedynie na wybranej grupie osobników (próbce). W modelach tych modelowanie przebiega w dwóch etapach. Najpierw tworzone jest drzewo genealogiczne wybranej próbki z korzeniem będącym ostatnim wspólnym

przodkiem (ang. Most Recent Common Ancestor, MRCA) wszystkich osobników wchodzących w skład próbki. Następnie, bazując na otrzymanym drzewie, do modelu dodawana jest informacja genetyczna (np. na gałęziach drzewa rozmieszczane są mutacje, zwykle z prawdopodobieństwem proporcjonalnym do długości gałęzi). Modele te wymagają często dodatkowych założeń i są dostosowane jedynie dla wąskiego przedziału wartości niektórych parametrów (np. niski współczynnik rekombinacji). Wspomniane ograniczenia oraz ciągły wzrost mocy obliczeniowej komputerów powodują, iż metody prospektywne zyskują na znaczeniu (simuPop [42], EASYPOP [3], TreesimJ [40]).

Drugim kryterium podziału modeli stochastycznych jest sposób uzyskiwania wyników. Ze względu na złożoność rozwiązywanych problemów, bardzo duża część aktualnie wykorzystywanych modeli oparta jest na metodach symulacyjnych. Podstawowym mechanizmem używanym w tego rodzaju modelach jest metoda Monte Carlo [36, 37]. Dokładne wyliczenia, zwykle niemożliwe do realizacji w sensownym czasie, zastępowane są przez uśrednione wyniki z wielokrotnie powtórzonych realizacji uproszczonego modelu. Oczywistą wadą metod symulacyjnych jest niedokładność otrzymanych rezultatów wynikająca z konieczności dokonania szeregu przybliżeń i wykorzystania heurystyk. Co więcej, poprawność skomplikowanych metod symulacyjnych powinna być zweryfikowana za pomocą metod analitycznych przynajmniej dla najprostszych przypadków. Z tego też powodu metody analityczne stanowią często platformę testową dla metod opartych o symulacje komputerowe.

2 Zakres i cel pracy

W niniejszej rozprawie skupiamy uwagę na teoretycznych (nie opartych o metody symulacyjne) systemach genetycznych modelujących złożone mechanizmy genetyczne. Systemy tego typu wymagają często, w celu uzyskania jakichkolwiek przydatnych i interesujących wyników, wykonania złożonych obliczeń niemożliwych do realizacji za pomocą klasycznych metod analitycznych.

Celem pracy jest pokazanie, iż systemy oparte na metodach analitycznych i realizowane za pomocą wyrafinowanych, specjalnie dedykowanych programów komputerowych mogą być wykorzystane do

rozwiązania pewnych złożonych problemów genetycznych.

Co więcej, działanie tak skonstruowanych systemów może, dla pewnych klas problemów, przewyższać sposób działania metod symulacyjnych. To znaczy, umożliwić uzyskanie dokładniejszych wyników w porównywalnym lub krótszym czasie działania. W rozprawie (rozdziały 4-6) przedstawione i szczegółowo omówione zostały trzy złożone systemy genetyczne.

Pierwszy model jest modelem Morana [38, 51] z mutacjami, dryfem genetycznym oraz rekombinacją miedzy wieloma loci. Wprowadzenie rekombinacji do modelu genetycznego powoduje ogromny wzrost złożoności modelu, uniemożliwiając często jego analizę. Z tego też powodu badania nad wzajemnymi relacjami między mutacjami, dryfem oraz rekombinacjami zaczęły być realizowane dość późno (lata osiemdziesiąte poprzedniego wieku). Dodatkowo, znaczna część prowadzonych badań dotyczy efektu rekombinacji jedynie w kontekście retrospektywnej teorii koalescencji [1, 14, 20, 25, 51]. Podejście retrospektywne, mimo iż znacznie szybsze od podejścia prospektywnego, nie pozwala na dokładne modelowanie różnych aspektów związanych z rekombinacją, szczególnie gdy rozważany jest system z wielokrotnie zależnymi loci (ang. multilinked loci). Wraz ze wzrostem mocy obliczeniowej, coraz częściej do modelowania rekombinacji jest wykorzystywane podejście prospektywne [3, 21, 23, 24, 40, 42] lub mieszane [41]. Mimo znacznego postępu w wyjaśnieniu zagadnień związanych z rekombinacją, wiele pytań wciąż pozostaje bez odpowiedzi. Nas szczególnie interesuje asymptotyczne zachowanie modelu rekombinacji typu crossover. Koncentrujemy się na problemie rozróżnialności (ang. identifiability). Dokładniej, chcemy znaleźć odpowiedź na pytanie, czy populacja może osiągnąć stan, w którym jest nieodróżnialna od populacji, która wyewoluowała jedynie pod wpływem mutacji oraz dryfu. Podobne analizy dynamiki rekombinacji w modelu Morana, aczkolwiek koncentrujące się na nieco innych aspektach, są przedstawione w [1, 2].

Znajomość rozkładu czasu do MRCA w danej populacji dostarcza informacji o historii tej populacji. Informacja o czasie do MRCA służy również do estymacji innych istotnych parametrów danej populacji (na przykład czas do MRCA jest ściśle związany ze stopniem pokrewieństwa osobników wybranych z danej populacji). Znaleziony czas do MRCA pewnej części populacji może być wykorzystany w anali-

zie całej populacji. Dlatego też metody wyznaczania czasu do wspólnego przodka są od dawna obiektem wzmożonych badań [19, 31, 33] i wiele aspektów wyznaczania tego czasu dla prostych modeli jest obecnie bardzo dobrze znanych [51]. W prostych modelach, opartych przede wszystkim na modelu Wrighta-Fishera, przyjmujemy zwykle stały lub inny dobrze znany (głównie eksponencjalny) scenariusz zmiany rozmiaru populacji. W ostatnich latach czas do MRCA jest wyznaczany dla coraz bardziej złożonych modeli. Jako przykład, możemy wspomnieć o podejściach dyfuzyjnych zastosowanych do modelu Wrighta-Fishera [43, 45] lub modelu gałązkowego [15]. Interesującym zagadnieniem jest jednak badanie czasu do MRCA w populacjach, w których nie zakładamy żadnych modeli zmiany ich rozmiaru. W rozdziale piątym dysertacji wyznaczamy czas do MRCA dużej próbki osobników wylosowanych z populacji, która wyewoluowała zgodnie z dowolnym (ale znanym) scenariuszem wzrostu.

Ślady zależności między różnymi gatunkami lub populacjami (również tymi wymarłymi) są widoczne w genomach ich osobników (lub przodków tych osobników). Rezultaty badań tych zależności pozwalają na znalezienie odpowiedzi na wiele pytań, od tych wynikających z prozaicznej ciekawości historią danych populacji, aż do znacznie ważniejszych dotyczących przykładowo szczegółow związanych z genealogią mutacji, co może być wykorzystane w metodach budowy genetycznej mapy (ang. gene mapping methods) mutacji odpowiedzialnych za występowanie rzadkich chorób genetycznych [54]. Niestety, wyjaśnienie wszystkich szczegółów wspomnianych zależności bazując na pewnej (zwykle nielicznej) próbce osobników nie jest łatwe. Podstawowym podejściem [22] wykorzystywanym w tym celu jest estymacja wybranych parametrów opisujących te zależności poprzez symulowanie próbek pasujących do danych rzeczywistych. Naukowcy zakładają zwykle kilka możliwych scenariuszy demograficznych i porównują wyniki uzyskane dla każdego z nich z rzeczywistymi danymi. W rozdziale szóstym rozprawy prezentujemy efektywną, ogólną metodę modelowania złożonych sieci demograficznych realizowaną metodami niesymulacyjnymi. Poza standardowymi zdarzeniami w sieci demograficznej (takimi jak podział pojedynczej populacji na dwie, połączenie się dwóch populacji lub migracje między populacjami), podstawowa wersja naszego modelu uwzględnia dryf wewnątrz populacji, zmianę rozmiaru populacji w czasie oraz dowolny markowski

model mutacji. Jako wynik otrzymujemy rozkład łączny (ang. joint distribution) pary osobników wylosowanych z dwóch, niekoniecznie różnych, populacji. Bazując na znajomości dokładnych wartości rozkładu łącznego, możliwe jest wyznaczenie innych parametrów opisujących zależności między modelowanymi populacjami.

Podsumowując, przedstawione w niniejszej rozprawie modele służą udowodnieniu słuszności nastepujących tez:

- **Możliwe jest, stosując analityczny system genetyczny oparty na matematycznym modelu Morana, rozstrzygnięcie kwestii rozróżnialności rekombinacji, przynajmniej w znaczeniu relacji ograniczonych do zbioru rozkładów łącznie charakteryzujących stany alleliczne w dowolnej liczbie różnych loci.**

- **Możliwe jest, wykorzystując analityczny rekurencyjny system genetyczny, wyznaczenie czasu do MRCA próbki o znacznym rozmiarze wylosowanej z dużej populacji ludzkiej, która wyewoluowała zgodnie z dowolnym, ale danym, scenariuszem zmiany rozmiaru populacji.**

- **Możliwe jest zbudowanie systemu genetycznego nie opartego na metodach symulacyjnych, który modeluje zależności między populacjami lub gatunkami w złożonej sieci demograficznej i zastępuje, przynajmniej w pewnych zastosowaniach, symulacyjne modele sieci demograficznej.**

3 Wyniki

Model Morana

Nasz model jest uogólnieniem na s loci modelu z dwoma loci przedstawionego w [5, 30]. Zakładamy, iż każdy z $2N$ osobników z populacji opisany jest przez s zmiennych losowych X_{ab} reprezentujących pojedynczy locus, gdzie $1 \leq a \leq 2N$ jest numerem osobnika oraz $1 \leq b \leq s$ jest numerem locusa. Mutacje modelujemy jako niezależne zmiany wartości zmiennych X. Każdy z osobników ma swój czas życia określony rozkładem wykładniczym z parametrem $\frac{2}{\lambda}$. W chwili śmierci osobnik jest

zastępowany nowym osobnikiem. Z prawdopodobieństwem $1 - r$ nie dochodzi do rekombinacji i nowy osobnik jest jednym z osobników z populacji (każdy z nich może być wybrany z prawdopodobieństwem $\frac{1}{2N}$). W przypadku rekombinacji, wybieramy miejsce wystąpienia rekombinacji (prawdopodobieństwo wystąpienia rekombinacji po locusie i jest równe r_i) oraz losujemy dwa osobniki j i k. Nowy osobnik powstaje poprzez połączenie loci znajdujących się na lewo od miejsca rekombinacji z osobnika j z pozostałymi loci wybranymi z osobnika k. Jak widać, ewoluujące osobniki mogą zawierać loci pochodzące od różnych osobników z pierwszego pokolenia. Wprowadzamy system rozkładów $\{D_{a_1 \ldots a_s}\}$ opisujących tak ewoluujące grupy osobników. W danym rozkładzie $a_i = a_j$ oznacza, iż loci na pozycjach i oraz j pochodzą od tego samego osobnika. Wprowadzamy regularny indeks rozkładów spełniający następujące własności:

1. a_1 is 1,
2. $a_\alpha \leq \max(a_1, \ldots, a_{\alpha-1}) + 1$, $\alpha \geq 2$;

Rozkłady sortujemy leksykograficznie od $D_{11\ldots1}$ do $D_{12\ldots s}$. Liczba wszystkich rozkładów w zależności od liczby loci jest liczbą Bella ϖ_s [18]. Tworzymy dyskretny łańcuch Markova z rozkładami jako stanami, przejściami między stanami w chwili śmierci osobnika oraz macierzą przejścia $\Theta = (1 - r)\Theta_0 + \sum_{i=1}^{s-1} r_i \Theta_i$. Otrzymujemy następujące wyrażenie na ewolucję systemu:

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = \mathcal{G}D(t) + \lambda N \Theta D(t) - \lambda N D(t), t \geq 0, \tag{1}$$

gdzie D jest kolumnowym wektorem rozkładów, a $\mathcal{G}$ jest generatorem operacji mutacji. Korzystając z faktu, iż łańcuch Markova określony za pomocą macierzy Θ jest ergodyczny (dowód w dysertacji), równanie (1) pozwala nam otrzymać następującą zależność dla dużych wartości t:

$$D_{1\ldots1}(t) \sim S(t) \sum_{\iota=1\ldots1}^{12\ldots s} \pi_\iota D_\iota(0), \tag{2}$$

gdzie $S(t)$ oznacza półgrupę mutacyjną, a π jest rozkładem stacjonarnym Θ. Wynika stąd, że asymptotycznie efekt rekombinacji jest nierozróżnialny od efektu

mutacji i dryfu. Jest to bardzo ciekawy, chociaż nieco paradoksalny, rezultat. Musimy jednak być świadomi, że rezultat ten dotyczy jedynie asymptotycznego zachowania modelu. Dodatkowo, system rozkładów D użyty w modelu nie jest kompletny. Zbiór rozkładów opisujących zależności w modelu charakteryzuje łączne stany allelyczne na wielu loci na jednym lub wielu chromosomach, ale nie określa łącznych stanów allelycznych na pojedynczym locusie na jednym lub więcej chromosomach. Oznacza to, że prawdopodobieństwa takie jak $P[X_{11} = x_{11}; X_{12} = x_{12}; X_{23} = x_{23}]$ są określone w naszym systemie, ale prawdopodobieństwa takie jak $P[X_{11} = x_{11}; X_{21} = x_{21}; X_{23} = x_{23}]$ nie. Mimo to, system jest wystarczająco bogaty aby wyznaczyć zarówno wszystkie możliwe wielopunktowe nierównowagi sprzężeń (ang. multipoint linkage disequilibrium), jak i ich wariancje i kowariancje [53].

W rozprawie badamy algorytmy pozwalające na zbudowanie macierzy Θ. Efektywne zarządzanie rozkładami D zapewnia specjalna funkcja mieszająca (ang. hashing function) oparta na programowaniu dynamicznym. Nasza metoda wyznacza wartość pojedynczej macierzy Θ_i w czasie $O(s^4\varpi_s + \varpi_s^2)$ wykorzystując $20\varpi_s^2$[B] pamięci, co jest wystarczające do otrzymania wyników dla $s \leq 9$. Wykorzystanie odpowiedniej implementacji macierzy rzadkich skutkowałoby niewielkim zwiększeniem tego limitu (kosztem pogorszenia złożoności czasowej).

Analiza wartości współczynnika Dobrušina dla przypadku $s = 3$ oraz wartości odstępu spektralnego (ang. spectral gap) w przypadku ogólnym sugerują eksponencjalną szybkość zbieżności macierzy Θ. Przeprowadzone porównanie naszego modelu z modelem rekombinacji Hudsona [25] bazującym na modelu Wrighta-Fishera pokazuje, że system oparty na modelu Morana wykazuje większą korelację czasu do MRCA na dwóch loci. Jest to bardzo ważny, chociaż dość intuicyjny, fakt (wynika on z dodatkowej zależności między osobnikami związanej z wprowadzeniem do modelu Morana czasu życia osobników). Wykorzystując interpolację średniokwadratową otrzymujemy następującą wartość tej korelacji:

$$Cor(t_1, t_2) = \frac{R + 32}{R^2 + 10R + 32}, \tag{3}$$

gdzie t_1, t_2 są czasami do MRCA na obu loci, a $R = 4Nr$.

Wyznaczanie czasu do MRCA

Załóżmy, iż populacja ewoluuje od pokolenia $t = 1$ do aktualnego pokolenia T. Rozmiar populacji w pokoleniu t jest dany jako N_t. W aktualnym pokoleniu losujemy n osobników z populacji. Interesować nas będzie dokładny rozkład czasu do MRCA wylosowanej próbki.

Niech $\alpha_{t,k}$ jest prawdopodobieństwem, że osobniki z próbki mają dokładnie k przodków w czasie t. Oczywiście, $\alpha_{T,n} = 1$ oraz $\alpha_{T,i} = 0$ dla $i \neq n$. Wartości α spełniają następującą zależność rekurencyjną:

$$\alpha_{t,k} = \sum_{i=k}^{n} \alpha_{t+1,i} q_{i,k,t}, \tag{4}$$

gdzie $q_{m,k,t}$ jest prawdopodobieństwem, iż m osobników wylosowanych z populacji w pokoleniu $t+1$ ma dokładnie k przodków w pokoleniu t [4] i wynosi:

$$q_{m,k,t} = \frac{S_{m,k}\binom{N_t}{k}k!}{N_t^m}, \tag{5}$$

gdzie $S_{m,k}$ jest liczbą Stirlinga drugiego rodzaju [18].

Korzystając z wartości α, szukany rozkład czasu do MRCA jest dany jako:

$$P(\tau_{n,T} = t) = (\alpha_{T-t,1} - \alpha_{T-t+1,1}) \tag{6}$$

W celu wyznaczenia wartości q korzystamy z następujących zależności rekurencyjnych (dowód poprawności w rozprawie):

$$q_{1,1,t} = 1, \qquad 1 \leq t \leq T \tag{7}$$

$$q_{i+1,i+1,t} = q_{i,i,t}\frac{N_t - i}{N_t}, \qquad 1 \leq t \leq T, 1 \leq i < n \tag{8}$$

$$q_{i+1,k,t} = \frac{W_{i,k}}{N_t} q_{i,k,t}, \qquad 1 \leq t \leq T, 1 \leq k \leq i < n \tag{9}$$

gdzie $W_{i,k} = \frac{S_{i+1,k}}{S_{i,k}}$.

$$W_{i,1} = 1, \qquad 1 \leq i \leq n \tag{10}$$

$$W_{i,i} = W_{i-1,i-1} + i, \qquad 2 \leq i \leq n \tag{11}$$

$$W_{i,k} = k + W_{k-1,k-1} \prod_{j=k}^{i-1} \frac{W_{j,k-1}}{W_{j,k}}, \qquad 2 \leq k < i \leq n. \tag{12}$$

Zależności (4) oraz (7)-(12) pozwalają, korzystając z programowania dynamicznego, na wyznaczenie wartości α w czasie $\mathrm{O}(n^3 + n^2 T)$. Użycie liczb $W_{i,k}$ pozwala uniknąć konieczności operowania dużymi liczbami Stirlinga; korzystając z faktu, iż $\frac{S_{n,k-1}}{S_{n,k}}$ jest ściśle malejący z $n \to \infty$ [7] pokazujemy, iż $W_{n,k} < n^2$. Wynika stąd, iż nasza metoda pozwala na bardzo szybkie wyznaczenie rozkładu czasu do MRCA nawet dla próbki $n \approx 10^3$ oraz okresie czasu porównywalnym do czasu życia gatunku ludzkiego.

Nasze podejście różni się znacząco od zwykle używanych w przypadku dużej próbki lub długiego odcinka czasu metod opartych na aproksymacji dyfuzyjnej ciągłego procesu koalescencji. Przykładowo, w Polański i inni [44] model oparty na aproksymacji dyfuzyjnej został użyty w celu estymacji historii populacji na podstawie różnicy osobników w próbce. Inny ciekawy model zaproponował Takahata [48, 49] w celu estymacji czasu do MRCA próbki wylosowanej z populacji o stałym rozmiarze, ale ewoluującej pod wpływem silnej selekcji. Takahata analizując przeżywalność starych linii genealogicznych estymuje wartości ściśle związane z wartościami α dostępnymi bezpośrednio w naszym modelu. Nasz model jest prosty i wystarczająco szybki aby z powodzeniem zastępować podejście dyfuzyjne w zakresie n nie przekraczającym 10^4. Dodatkowo, w przeciwieństwie do metod opartych na aproksymacji dyfuzyjnej, nasz model działa dobrze dla małych populacji. Znajomość dokładnych wartości α może być również pomocna w metodach estymacji historii całej populacji na podstawie historii próbki [34] lub też w analizie dynamiki zmiany MRCA w czasie [43].

W dysertacji użyliśmy naszej metody do wyznaczenia czasu do MRCA dla populacji świata i Polski otrzymując dość paradoksalny rezultat stwierdzający, iż wspólny przodek obu tych populacji pojawił się przed pojawieniem się naszego gatunku. Musimy jednakże pamiętać, iż otrzymana wartość dotyczy fragmentu genomu, który nie podlegał ani rekombinacji, ani znaczącej selekcji.

W przypadku braku kompletnej informacji na temat demografii populacji w całym badanym okresie, nasza metoda może wciąż być użyta pod warunkiem, iż w modelu

uwzględnione zostaną wszystkie ważne wydarzenia demograficzne. W tym przypadku, brakujące dane (rozmiary populacji) pomiędzy dwoma kolejnymi wydarzeniami mogą być interpolowane, na przykład za pomocą eksponenty, i nie powinno to mieć dużego wpływu na otrzymane wyniki. Model może być również użyty jako platforma testowa służąca do weryfikacji nieznanych scenariuszy demograficznych na podstawie danych genetycznych.

Nasz model zastosowaliśmy również do badania czasu do MRCA w populacji powstałej zgodnie z procesem Galtona-Watsona [28, 52]. Genealogie testowe stworzyliśmy wykorzystując specjalnie na ten cel przygotowaną platformę zawierającą, m.in., nowy algorytm, który pozwala na wydajne czasowo i pamięciowo zarządzanie niewymarłymi liniami genealogicznymi powstałymi w procesie Galtona-Watsona. Porównanie naszej metody z metodami symulacyjnymi wykazuje zalety zaproponowanego podejścia (szybkość działania oraz zmniejszenie wariancji rozkładu) kosztem uwzględnienia niepełnej informacji o danej genealogii (wykorzystujemy jedynie informację o rozmiarze populacji nie biorąc pod uwagę dokładnej postaci drzewa genealogicznego).

Model sieci demograficznej

Przez sieć demograficzną rozumiemy zbiór populacji, które ewoluują od czasu $t_0 = 0$ z pojedynczej populacji. Populacje w sieci charakteryzowane są za pomocą rozkładów łącznych cech (alleli) populacji z sieci. Precyzyjniej, dla każdej pary populacji (x, y) wyznaczamy w czasie t rozkład $R_{xy}(t) = \{r_{xy}[a, b](t)\}$ określający prawdopodobieństwo wystąpienia cechy typu b w losowo wybranym osobniku z populacji y przy założeniu, że losowo wybrany osobnik z populacji x ma cechę typu a. Na podstawie znajomości tak określonych rozkładów łącznych jesteśmy w stanie obliczyć większość powszechnie używanych parametrów opisujących relacje między populacjami.

Zakładamy, iż w sieci mogą wystapić trzy rodzaje dyskretnych zdarzeń: (i) połączenie (ang. merge) dwóch populacji w jedną, (ii) wyodrębnienie się nowej populacji poprzez podział (ang. split) populacji na dwie oraz (iii) migracje między populacjami w sieci. Zdarzenia te występują w czasach t_i, $1 \leq i \leq I$, gdzie t_I jest

aktualnym czasem. W czasie między wydarzeniami dyskretnymi $[t_i, t_{i+1})$ sieć ewoluuje pod wpływem dryfu genetycznego i mutacji zgodnie z modelem przedstawionym w [6]. Ewolucja ta jest opisana równaniem Lyapunova [16]:

$$\frac{\mathrm{d}R_{ab}(t)}{\mathrm{d}t} = Q_a^T R_{ab}(t) + R_{ab}(t)Q_b + \frac{\delta_{ab}}{N_a(t)}(\Pi(t) - R_{ab}(t)), \tag{13}$$

gdzie $t \in [t_i, t_{i+1})$, Q_a jest macierzą intensywności mutacji w populacji a w danym przedziale czasowym, $N_a(t)$ jest rozmiarem populacji a, $\Pi(t)$ jest macierzą diagonalną z wartościami na przekątnej $\pi_{jj}(t)$ będącymi prawdopodobieństwami wystąpienia cechy j w populacji a w czasie t, δ jest deltą Kroneckera oraz Q^T oznacza transpozycję macierzy Q.

Przyjmujemy, iż przestrzeń stanów allelycznych, opisany za pomocą wybranego modelu mutacji, nie zmienia się w sieci. Intensywności mutacji (wartości w macierzy Q) mogą się różnić między populacjami lub w różnych przedziałach czasowych. Model dopuszcza dowolny scenariusz zmiany rozmiaru populacji.

Operacje podziału i złączenia zmieniają liczbę populacji w sieci. Oznaczmy liczbę populacji w sieci między zdarzeniami i oraz $i+1$ za pomocą κ_i. Jeśli pewna populacja w czasie po zdarzeniu t_i ma indeks k, to indeks tej populacji w czasie przed zdarzeniem $t_i - 0$ oznaczamy jako k'. Podobnie, macierze $R_{ab}(t_i)$ oraz $R_{a'b'}(t_i - 0)$ oznaczają rozkłady łączne między tymi samymi populacjami natychmiast po i natychmiast przed wydarzeniem i. Jeśli wydarzeniem tym jest podział, wtedy:

$$R_{ab}(t_i) = R_{a'b'}(t_i - 0). \tag{14}$$

Jeśli wydarzeniem i jest złączenie dwóch populacji, allele na chromosomie mogą być wybrane z obu łączących się populacji x i y z prawdopodobieństwami równymi odpowiednio p i $q = 1 - p$, gdzie $p = \frac{N_x(t_i-0)}{N_x(t_i-0)+N_y(t_i-0)}$. Skutkuje to następującą zmiano wartości rozkładów łącznych:

$$R_{ab}(t_i) = \begin{cases} R_{a'b'}(t_i - 0) & x \neq a',\ x \neq b' \\ pR_{a'b'}(t_i - 0) + qR_{yb'}(t_i - 0) & a' = x,\ b' \neq y \\ pR_{a'b'}(t_i - 0) + qR_{a'y}(t_i - 0) & b' = x,\ a' \neq y \\ p^2 R_{xx}(t_i - 0) + 2pqR^+_{xy}(t_i - 0) + q^2 R_{yy}(t_i - 0) & a' = x,\ b' = y \end{cases} \tag{15}$$

gdzie $2R^{+}_{ab}(t) = R_{ab}(t) + R_{ba}(t)$.

Zdarzenie migracji w czasie t_i jest opisane macierzą $M(t_i) = \{m_{xy}(t_i)\}$, $0 \leq x, y < \kappa_i$. Każda z wartości m_{xy}, $0 \leq m_{xy} \leq 1$, $m_{xx} = 0$ równa jest współczynnikowi migracji z populacji x do y. Rozmiar populacji migrującej w czasie t_i z populacji x do y jest równy $m_{xy}(t_i)N_x(t_i - 0)$. Zdarzenie migracji realizowane jest w dwóch krokach:

- Z każdej populacji wyodrębniamy za pomocą operacji podziału część, która migruje. Rozmiar wyodrębnionej podpopulacji z populacji x wynosi $N_x(t_i - 0)\sum_{k=0}^{\kappa_i - 1} m_{xk}(t_i)$. Zakładamy, iż podpopulacja wyodrębniona z populacji x ma indeks x''.

- Każdą z κ_i populacji x'' otrzymanych w poprzednim kroku dzielimy $\kappa_i - 1$ razy wyodrębniając pojedyncze migracje z jednej populacji do drugiej. Po każdym takim podziale wykonujemy operację złączenia z docelową populacją.

Powyższy schemat wymaga wykonania κ_i^2 podziałów i $\kappa_i(\kappa_i - 1)$ złączeń i przechowywania maksymalnie $2\kappa_i + 1$ populacji w tym samym czasie.

W ramach dysertacji stworzony został program realizujący opisywany model. Parametrem wejściowym programu jest skrypt opisujący sieć demograficzną. Do numerycznego rozwiązania ewolucji populacji, danej równaniem (13), wykorzystujemy algorytm Rungego-Kutty czwartego rzędu (RK4) [17] z adaptacyjnym doborem kroku metodą Casha-Karpa [8]. Wszystkie operacje na macierzach rzadkich (w tym operacja mnożenia macierzy) realizowane są w czasie nie gorszym niż kwadratowy. Szacujemy, iż złożoność czasowa modelu wynosi $\mathrm{O}(\kappa^2 kr(60 + 8c)N_{\mathbb{A}}^2)$, gdzie k jest liczbą podstawowych zdarzeń dyskretnych (podziałów i złączeń), a r jest średnią liczbą kroków algorytmu RK4 dla pojedynczego przedziału czasowego (zwykle $r \ll 100$, szczególnie przy wykorzystaniu algorytmu adaptacyjnego doboru kroku). Model może być stosowany nawet dla $N_{\mathbb{A}} \approx 1000$ oraz $\kappa > 10$.

W rozprawie prezentujemy kilka rozszerzeń modelu:

- wprowadzamy nowy model ewolucji populacji dla mikrosatelit, zastępując rozkład łączny liczby powtórzeń tandemowych (13) ich różnicą [29]

- wyznaczamy rekurencyjną zależność na rozkład łączny w próbce większej niż dwa

- wprowadzamy algorytm kompresji wykorzystywany przy modelowaniu długich sekwencji haplotypowych ograniczający rozmiar przestrzeni stanów z 2^s do około $s^2/4$, gdzie s jest liczbą nukleotydów w sekwencji

Inne możliwe rozszerzenia modelu mogą uwzględniać, na przykład, dodanie nowych mechanizmów genetycznych. Wprawdzie zastosowanie koalescencyjnego schematu rekombinacji [25] nie jest możliwe (gdyż wprowadza zależność rozkładu łącznego dla próbki o rozmiarze n od rozkładu dla próbki o rozmiarze $n+1$), ale możliwe jest dodanie modelu rekombinacji z naszego modelu Morana. Selekcja może być uwzględniona w modelu na dwa sposoby [39] używając albo selekcyjnego grafu przodków (ang. ancestral selection graph) [32], albo tzw. koalescencji strukturalnej (ang. structured coalescent) [27].

Nasza metoda różni się od zwykle używanego podejścia symulacyjnego. Oczywistą przewagą naszej metody jest to, iż otrzymane wyniki są dokładne. Na podstawie otrzymanych wartości rozkładów łącznych alleli jesteśmy w stanie wyznaczyć praktycznie wszystkie najważniejsze parametry opisujące populacje i ich relacje. Pomimo pewnych ograniczeń wydajnościowych, nasza metoda może być z powodzeniem wykorzystywana do analizy rzeczywistych danych genetycznych. Szczególnie dotyczy to modelu mikrosatelit (liczba powtórzeń tandemowych rzadko osiąga wartości większe niż 100).

W rozprawie prezentujemy kilka możliwych zastosowań naszej metody. Podstawowy obszar wykorzystania modelu leży w szacowaniu wartości różnych parametrów w wybranych scenariuszach demograficznych. W dysertacji szacujemy: parametry w równowadze mutacji i dryfu dla prostego modelu SNP, w tym nierównowagę sprzężeń (ang. linkage disequilibrium), różnicę w parach osobników (ang. pairwise difference) w długich sekwencjach haplotypowych, odległość Slatkina R_{ST} [46] między dwoma populacjami oraz obciążenie próbkowania (ang. ascertainment bias) B [11] dla modelu mikrosatelit. Eksperymenty te pozwoliły nam uzyskać kilka interesujących rezultatów. Wyniki uzyskane z modelu obciążenia próbkowania sugerują, iż współczynnik mutacji w mikrosatelitach u człowieka jest wyższy niż u szym-

pansa. Dodatkowo, badamy wpływ górnej granicy liczby powtórzeń tandemowych występujących na locusie mikrosatelitarnym na wartość B. Istnienie takich granic, związane z różną dynamiką procesów zachodzących na locusie u różnych gatunków, jest głównym powodem występowania zjawiska ociążenia próbkowania [50]. Jako wynik otrzymujemy, iż ustawienie tej granicy na wartość powyżej 30 nie wpływa na B.

Wartości parametrów estymowane dla założonego scenariusza demograficznego mogą być porównywane z wartościami uzyskanymi z danych genetycznych. W ten sposób parametry te służą nam jako miary do testowania przeszłych nieznanych scenariuszy demograficznych [47]. W dysertacji wykorzystujemy to podejście to zbadania wspólnej historii Słowian i Bałtów na podstawie analizy danych z chromosomu Y. Jako miary używamy odległości R_{ST}. Jako wynik uzyskujemy oszacowanie zależności między współczynnikiem migracji między przodkami współczesnej Polski oraz Bałtami, a usytuowaniem w czasie kilku wydarzeń demograficznych (takich jak, na przykład, wyodrębnienie się Słowian i Bałtów z grupy narodów Indoeuropejskich).

4 Podsumowanie

Przedmiotem rozprawy są złożone systemy genetyczne. Opisane w pracy wyniki eksperymentów analitycznych i numerycznych świadczą o tym, iż metody modelowania takich systemów nie oparte na podejściu symulacyjnym mogą z powodzeniem zastępować metody symulacyjne. Oczywiście, zakres wykorzystania metod analitycznych jest mniejszy. Metody te nie wymagają jednak tak starannej weryfikacji, jak metody symulacyjne. Co więcej, same mogą stanowić platformę testową dla prostszych scenariuszy symulacyjnych.

W dysertacji zaprezentowane zostały trzy złożone systemy stochastyczne rozwiązujące problemy rozróżnialności rekombinacji, wyznaczenia czasu do wspólnego przodka dużej próbki z populacji o dowolnym scenariuszu wzrostu oraz modelowania złożonej sieci demograficznej. Wyniki uzyskane z wykorzystaniem tych modeli, naszym zdaniem, udowadniają prawdziwość postawionych tez. Rezultaty te nie mogłyby jednak być uzyskane bez wykorzystania dedykowanych algorytmów

komputerowych, np. algorytmów zarządzania rozkładami i podziału przestrzeni stanów w modelu Morana, rekurencji opartej na programowaniu dynamicznym w modelu wyznaczającym rozkład czasu do MRCA, czy też algorytmów rozwiązywania ODE, operowania na macierzach rzadkich i efektywnego zarządzania populacjami w modelu sieci demograficznej.

Wyniki uzyskane w rozpawie mogą być punktem wyjścia dla przyszłych badań. W szczególności dotyczy to:

- analizy wybranych modeli mutacji w modelu Morana z rekombinacjami
- uwzględnienia opracowanego modelu rekombinacji w innych modelach (np. w modelu sieci demograficznej)
- wykorzystania algorytmu wyznaczającego rozkład czasu do MRCA do testowania lub estymowania scenariuszy demograficznych
- poszerzenia modelu sieci demograficznej o nowe mechanizmy ewolucyjne (np. selekcję, rekombinację lub draft genetyczny)
- rozwinięcia badań nad optymalizacją modelu sieci demograficznej (np. badanie efektywnego wyznaczania rozkładów łącznych alleli dla większej próbki lub bardziej złożonych modeli mutacji)
- wykorzystania modelu sieci demograficznych w analizie różnych scenariuszy demograficznych

Literatura

[1] Ellen Baake and Inke Herms. **Single-crossover dynamics: finite versus infinite populations**. *Bulletin of Mathematical Biology*, **70(2)**:603–624, 2008.

[2] Ellen Baake and Thiemo Hustedt. **Moment closure in a Moran model with recombination**. *arXiv:1105.0793v1 [math.PR]*, 2011.

[3] Francois Balloux. **EASYPOP (Version 1.7): A computer program for population genetics simulations**. *Journal of Heredity*, **92(3)**:301–302, 2001.

[4] Adam Bobrowski. *Functional analysis for probability and stochastic processes*. Cambridge University Press, 2005.

[5] Adam Bobrowski and Marek Kimmel. **A random evolution related to a Fisher-Wright-Moran model with mutation, recombination and drift**. *Mathematical Methods in the Applied Sciences*, **26**:1587–1599, 2003.

[6] Adam Bobrowski, Marek Kimmel, Ovide Arino, and Ranjit Chakraborty. **A semigroup representation and asymetric behavior of certain statistics of the Fisher-Wright-Moran coalescent**. *Handbook of Statistics*, **19**:215–242, 2001.

[7] Rodney E. Canfield and Carl Pomerance. **On the problem of uniqueness for the maximum Stirling number(s) of the second kind**. *Electronic Journal of Combinatorial Number Theory*, **2(2002), Paper A01 electronic only**:13–13, 2002.

[8] Jeff R. Cash and Alan H. Karp. **A variable order Runge-Kutta method for initial value problems with rapidly varying right-hand sides**. *ACM Transactions on Mathematical Software*, **16(3)**:201–222, 1990.

[9] International HapMap Consortium. **The International HapMap Project**. *Nature*, **426(6968)**:789–796, 2003.

[10] The 1000 Genomes Project Consortium. **A map of human genome variation from population-scale sequencing**. *Nature*, **467(7319)**:1061–1073, 2010.

[11] Allan M. Crawford et al. **Microsatellite evolution: testing the ascertainment bias hypothesis**. *Journal of Molecular Evolution*, **46**:256–260, 1998.

[12] Charles Darwin. *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life.* John Murray, 1859.

[13] Charles Darwin and Alfred R. Wallace. **On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection**. *Linnean Society of London*, **Zoology 3**:46–50, 1858.

[14] Richard Durret. *Probability Models for DNA Sequence Evolution.* Springer, New York, 2002.

[15] Steven N. Evans and Peter L. Ralph. **Dynamics of the time to the most recent common ancestor in a large branching population**. *Annals of Applied Probability*, **20(1)**:1–25, 2010.

[16] Zoran Gajic, Muhammed Tahir, and Javed Qureshi. *Lyapunov matrix equation in system stability and control.* Academic Press, San Diego, 1995.

[17] William C. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations.* Prentice-Hall, 1971.

[18] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics : a foundation for computer science.* Addison-Wesley, 1994.

[19] Robert C. Griffiths. **Lines of descent in the diffusion approximation of neutral Fisher-Wright models**. *Theoretical Population Biology*, **17**:37–50, 1980.

[20] Robert C. Griffiths. **Neutral two-locus multiple allele models with recombination**. *Theoretical Population Biology*, **19(2)**:169–186, 1981.

[21] Frederic Guillaume and Jacques Raugemont. **Nemo: an evolutionary and popuation genetics programming framework**. *Bioinformatics*, **22(20)**:2556–2557, 2006.

[22] Silvia Guimaraes et al. **Genealogical Discontinuities among Etruscan, Medieval, and Contemporary Tuscans**. *Molecular Biology abd Evolution*, **26(9)**:2157–2166, 2009.

[23] Jody Hey. **FPG: a computer program for forward population genetic simulation**. http://lifesci.rutgers.edu/~heylab/HeyLabSoftware.htm.

[24] Clive J. Hoggart et al. **FREGENE: software for simulating large genomic regions**. http://www.ebi.ac.uk/projects/BARGEN/download/FREGEN/.

[25] Richard R. Hudson. **Properties of a neutral allele model with intragenic recombination**. *Theoretical Population Biology*, **23(2)**:183–201, 1983.

[26] Entire issue of Science. **The human genome**. *Science*, **291(5507)**:1145–1434, 2001.

[27] Norman L. Kaplan, Thomas Darden, and Richard R. Hudson. **The coalescent process in models with selection**. *Genetics*, **120**:819–829, 1988.

[28] Marek Kimmel and David E. Axelrod. *Branching Processes in Biology*. Springer Varlag, New York, 2002.

[29] Marek Kimmel et al. **Signatures of population expansion in microsatellite repeat data**. *Genetics*, **148**:1921–1930, 1998.

[30] Marek Kimmel and Joanna Polańska. **A model of dynamics of mutation, genetic drift and recombination in DNA-repeat genetic loci**. *Archives of Control Sciences*, **9(XVL)**:143–157, 1999.

[31] John F. C. Kingman. **The coalescent**. *Stochastic Processes and their Applications*, **13(3)**:235–248, 1982.

[32] Stephen M. Krone and Claudia Neuhauser. **Ancestral processes with selection**. *Theoretical Population Biology*, **51(3)**:210–237, 1997.

[33] Ray A. Littler. **Loss of variability at one locus in a finite population**. *Mathematical Biosciences*, **25**:151–163, 1975.

[34] Yosef E. Maruvka, Nadav M. Shnerb, Yaneer Bar-Yam, and John Wakeley. **Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate**. *Molecular Biology and Evolution, In Press*, **28(5)**:1617–1631, 2011.

[35] Gregor J. Mendel. **Versuche uber Pflanzen-Hybriden**. *Verhandlungen des naturforschenden Vereines in Brűnn*, **IV(1865)**:3–47, 1865.

[36] Nicholas Metropolis. **The beginning of the Monte Carlo method**. *Los Alamos Science*, **15**:125–130, 1987.

[37] Nicholas Metropolis and Stanisław Ulam. **The Monte Carlo method**. *Journal of the American Statistical Association*, **44(247)**:335–341, 1949.

[38] P. A. P. Moran. **A general theory of the distribution of gene frequencies**. *Proceedings of the Royal Society, B Biological Sciences*, **149(934)**:113–116, 1958.

[39] Magnus Nordborg. *Coalescent Theory. In: Handbook of Statistical Genetics, David J. Balding, Martin Bishop, Chris Cannings eds.* John Wiley and Sons, Chichester, 2001.

[40] Brendan O'Fallon. **TreesimJ: a flexible, forward time population genetic simulator**. *Bioinformatics*, **26**:2200–2201, 2010.

[41] Badri Padhukasahasram et al. **Exploring population genetic models with recombination using efficient forward-time simulations**. *Genetics*, **178(4)**:2417–2427, 2008.

[42] Bo Peng and Marek Kimmel. **simuPop: A forward-time population genetics simulation environment**. *Bioinformatics*, **21**:3686–3687, 2005.

[43] Peter Pfaffelhuber and Anton Wakolbinger. **The process of most recent common ancestors in an evolving coalescent**. *Stochastic Processes and their Applications*, **116(12)**:1836–1859, 2006.

[44] Andrzej Polański, Marek Kimmel, and Ranajit Chakraborty. **Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data**. *Proceedings of the National Academy of Science of the United States of America*, **95**:5456–5461, 1998.

[45] Damien Simon and Bernard Derrida. **Evolution of the most recent common ancestor of a population with no selection**. *Journal of Statistical Mechanics*, **(2006) P05002**:10.1088/1742–5468/2006/05/P05002, 2006.

[46] Montgomery Slatkin. **A measure of population subdivision based on microsatellite allele frequencies**. *Genetics*, **139**:457–462, 1995.

[47] Mark Stoneking and Johannes Krause. **Learning about human population history from ancient and modern genomes**. *Nature Reviews Genetics*, **12**:603–614, 2011.

[48] Naoyuki Takahata. **A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism**. *Proceedings of the National Academy of Science of the United States of America*, **87**:2419–2423, 1990.

[49] Naoyuki Takahata. *Evolutionary Genetics of Human Paleo-Populations. In: Mechanisms of Molecular Evolution, Naoyuki Takahata and Andrew G. Clark eds.* Japan Scietific Societies Press, Tokio, 1993.

[50] Edward J. Vowels and William Amos. **Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences**. *Molocular Biology Evolution*, **23(3)**:598–607, 2006.

[51] John Wakeley. *Coalescent Theory: An Introduction*. Ben Roberts Rublishing, 2008.

[52] Henry W. Watson and Francis Galton. **On the probability of the extinction of families**. *Journal of the Anthropological Institute of Great Britain*, **4**:138–144, 1874.

[53] Bruce S. Weir. *Genetic data analysis II: methods for discrete population genetic data*. Sinauer Associates Inc, 1996.

[54] Carsten Wiuf. *Highly Structured Stochastic Systems, chapter 14*. Oxford University Press, 2003.