

Záró jelentés

Sztochasztikus modellek a bioinformatikában és többváltozós ökológiai adatfeltárásban

A pályázat során végzett kutatásainkat négy csoportba oszthatjuk:

1. Új algoritmusok kidolgozása
2. Új Markov lánc Monte Carlo metódusok kidolgozása
3. A metódusok elméleti jóságának az elemzése
4. A kidolgozott metódusok alkalmazásai

A pályázati időszakban 3 új algoritmust dolgoztunk ki.

a) Genomátrendezések vizsgálatára használt Markov lánc Monte Carlo tranzíciós kernelére dolgoztunk ki egy új algoritmust. Az új algoritmus segítségével $O(n)$ időben tudunk random transzpozíciókat és invertált transzpozíciókat mintavételezni, amelyek a valóság és kívánalom gráfban a körök számát növelik. A naív algoritmus $O(n^3)$ időben fut, így jelentős gyorsítást sikerült elérni.

b) Megadtunk egy algoritmust, amely génduplikáció-géntörlés-horizontális géntranszfert modellező időfolytonos Markov modellben számolja ki egy evolúciós fa likelihoodját. Az algoritmus futási ideje $O(NM^2)$, ahol N a fajok száma, M pedig az össz-kópiaszám a leveleken összegezve. Az algoritmus Java programozási nyelven lett implementálva, és akadémiai célú kutatásokra szabadon elérhető.

c) Megadtunk egy olyan algoritmust, amely egy fokszámsorozathoz tartozó összes gráfrealizációt előállítja. Bár ezen gráfrealizációk száma exponenciálisan nőhet a fokszámsorozat hosszával, az algoritmusunk két előállítás között csak polinom időben fut.

5 új Markov lánc Monte Carlo metódust dolgoztunk ki.

a) A SimulFold egy olyan MCMC módszer, amely szimultán RNS-ek álcsumókat megengedő térszerkezeteinek, többszörös szekvenciaillesztéseinek és evolúciós kapcsolatainak együttes Bayesi eloszlásából mintavételez.

b) A StatAlign programcsomagunk evolúciós fák és többszörös szekvenciaillesztések együttes Bayesi eloszlásából mintavételez. A beszúrás-törlés modell megenged hosszú beszúrások és törlések egyetlen lépésben való bekövetkezését is, ezáltal az affín résbüntetési optimalizálási alapú szekvenciaillesztési módszer sztochasztikus változatának fogható fel.

c) Kidolgoztunk egy olyan MCMC módszert, amely vegetációs térképek dinamikáját interpolálja és extrapolálja. Az eljárást több hazai és nemzetközi konferencián bemutattuk, a módszert bemutató kézirat a közeljövőben kerül benyújtásra.

d) Kidolgoztunk egy párhuzamos MCMC módszert két unikromoszómális genom legrövidebb reverziós útjainak a mintavételezésére.

e) A BigFoot programcsomag transzkripció faktor kötőhelyeket prediktál összehasonító genomikai alapon, úgyszintén MCMC metódust használva.

1 cikkünk került elfogadásra, amelyben megmutatjuk, hogy egy régebben általunk is és más kutatócsoportok által is alkalmazott MCMC módszer lassan keveredik unikromoszómális genomok legrövidebb reverziós útjainak a mintavételezésére. Valójában ez az eredmény sarkalt minket arra, hogy az előző rész d) pontjában ismertetett eljárást kidolgozzuk.

3 olyan munkánk került publikálásra, amelyben a kidolgozott metódusokat használjuk.

a) Megmutattuk, hogy a StatAlign programcsomag segítségével lehet fehérjék másodlagos térszerkezetét prediktálni. Megmutattuk, hogy a predikció pontossága korrelál az adott

szekvenciaillesztési oszlop poszterior valószínűségével. Mivel ez a poszterior valószínűség algoritmikusan könnyen számolható, meg tudjuk határozni azokat a pozíciókat, ahol a predikciónk nagy valószínűséggel pontos, és azokat a pozíciókat, ahol a predikció bizonytalan.

b) Genomátrendező MCMC módszerrel baktérium populációk genomátrendeződési dinamikáját vizsgáltuk. Megmutattuk, hogy az inter-replicore inverziók gyakoribbak az intra-replicore inverzióknál.

c) Vizsgáltuk baktériumok és ősbaktériumok genomjainak a géntartalomváltozását.