Atasal Dizi Tahmini - Bash Temelli Kisa Uygulama

Evrimsel Genombilim Kis Okulu, Hacettepe Biyoloji Bolumu, 2018.

Sorular ve komutlarda bulduğunuz yanlislar icin: ozan.bozdag@gmail.com

A) Giris

Atasal dizi tahmini temelde 3 adim iceriyor: 1) sequence alignment 2) phylogenetic tree yapimi (bizim uygulamamizda 'maximum likelihood' yontemi ile) 3) ilk iki veriyi (alignment+filogeni) bir araya getirip atasal dizilerin tahmin (!) edilmesi.

1 no'lu adimi t-coffee kullanarak yapiyoruz.

2 no'lu adimi 'maximum likelihood' ile agac cizimi / hesabi yapan bir program kullanabilirsiniz (orn. PhyML). Fakat biz adim #2 ve #3'u tek bir seferde yapabildigi icin iqtree'yi kullaniyoruz/kullandik. Boylece #2 & #3 ayni komut ile gerceklestirilecek.

Not: iqtree'nin bir avantaji sekans verisine uygun olabilecek nukleotid substitution modelleri arasindan secim yapabilmesi. Beyaz tahta uzerinde substitution modellerinin neden evrimsel agac ciziminde ve dal uzunluklarinin kalibre edilmesinde onemli oldugunu konusmustuk! Eger farkli programlar ile subsitution model secimi yapmak istiyorsaniz buna bakabilirsiniz (alignment yaptiktan sonra cikan output'u model testi yapan programa input olarak veriyorsunuz, ve size gereken modeli oneriyor, ilgili modeli sonraki adimda kullanacaginiz filogeni programina bildirmeniz gerekiyor – her program size sunulan model secenegini icermeyebilir!): https://academic.oup.com/mbe/article/25/7/1253/1045159

Cok onemli uyari: Kendi calistiginiz sekans verileri uzerinde alignment, filogeni, ve atasal dizi tahmini yapmaya baslamadan once kullanacaginiz programlarin 'el kitaplarini' (manual) okumaniz bilimsel etik ve tutarlilik nedeniyle gerekli.

https://media.readthedocs.org/pdf/tcoffee/latest/tcoffee.pdf
http://www.iqtree.org/doc/iqtree-doc.pdf

B. Teori

Asagida cok kis bir iki komut uygulayarak modern primatlarin ADH4 amino asid dizilerini kullanarak atasal ADH4 enzimi amino asit dizilerini tahmin edicegiz. ADH4 enzim aktivitesinin evrimsel olarak nasil degistigini arastiran cok ilginc bir calismayi temel aldik: http://www.pnas.org/content/pnas/112/2/458.full.pdf

Carrigan et al 2015 "Hominids adapted to metabolize ethanol long before human-directed fermentation" www.pnas.org/cgi/doi/10.1073/pnas.1404167111

Hedef: Calismada onemli oldugu bulunan A294V mutasyonunu atasal dizilerde bulmayi hedefliyoruz (makalede Figur 1). Not: kis okulu sirasinda atasal dizi tahminini grup calismasi olarak secen arkadaslariniz bu onemli mutasyonu dogru evrimsel yerlesimi ile buldular!

C. Uygulama - BASH komut dizini uzerinde uygulanacak -

BASH komutlari '$' isareti ile basliyor.

1. ADIM: T COFFEE ile alignment

http://www.tcoffee.org/Projects/tcoffee/papers/tcoffee.pdf

T-coffee linux son surumunu indirin:

http://www.tcoffee.org/Projects/tcoffee/#DOWNLOAD

Dosya uzantisi .tar.gz ise, bu ziplenmis dosyayi bulundugu klasore bash komutlari ile giderek acin:

$ tar -xvzf indirdiginiz_dosya_ismi.tar.gz

tab yapmayi unutmayin!

T-coffee'yi install etmek icin: http://tcoffee.readthedocs.io/en/latest/tcoffee_installation.html.

Onemli: T coffee'yi bash uzerinde herhangi bir klasorde iken (herhangi bir konumdan) calistirmak icin program klasorunuzu linux veya bio-linux'unuzun "search path"ine kopyalamaniz gerekiyor:
https://oxaric.wordpress.com/2008/12/04/make-a-bash-script-globally-executabl/

Benim kullandigim Linux uzerinde search path /usr/bin/ konumu. “sudo copy” yaparak search path'e attim t_coffee'yi:
$ sudo cp -R /home/manager/Downloads/T-COFFEE_installer_Version_11.00.8cbe486_linux_x64/bin/t_coffee /usr/bin/

T-coffee versiyonunuzu ve herhangi bir konumdan calisiyor oldugunu kontrol edin (download ettiginiz klasorden baska bir konuma gecerek deneyebilirsiniz):

$ t_coffee -version

T-coffee seceneklerini okuyun (manual'i dikkatlice okudugunuzu varsayiyorum). http://www.tcoffee.org/Projects/tcoffee/#DOCUMENTATION

$ man t_coffee

Simdi istediginiz yontemle alignment yapin – biz en basit yontemi uyguladik, ornegin '-mode' secenegi ile 'accurate' komutu vererek daha iyi hesaba dayanan farkli islemler yapabilirsiniz:

$ t_coffee adh4_primates_fasta.txt

3 farkli formatta cikti ("output") dosyasi verecek: newick, aln, html.

'adh4_primates_fasta.txt.aln' ciktisini kullanacagiz.

Dosya sonucuna bakin:
$ cat adh4_primates_fasta.aln

2&3. ADIM: Filogeni olusturmak, substitution model'i secmek, ve tahmini atasal dizileri cikarmak:

Filogeni icin:
https://academic.oup.com/mbe/article/32/1/268/2925592

Atasal dizi tahmini ('-asr') secenegi hakkinda bilgi icin:
http://www.iqtree.org/doc/iqtree-doc.pdf

ONEMLI NOT: iqtree surumu iqtree-1.6.1-Linux veya ustu olmali - daha onceki surumlerde -asr secenegi (ancestral sequence reconstruction=asr) yok.

Dogru iqtree surumune sahip oldugunuzu kontrol edin:

$ iqtree -v

Komutlari/secenekleri calisin:

$ iqtree --help

ornegin MODEL-FINDER ve SUBSTITUTION MODEL seceneklerini okuyun.

-o ile outgroup secenegini inceleyin. Makalede gorunen outgroup (dis-grup) 'E_Tree.ADH.4_Tupaia_belangeri_' isimli tur. ADH4 protein dizisileri arasinda en uzak diziyi iceren ADH4 protein dizisi bu ture ait olabilir (unutmayin, tur-agaci ile tek bir gen dizisine dayanan filogenetik agaclar uyumsuz olabilir!). Asagida bu turu outgroup secenegi ile tanitiyoruz. IQTREE eger outgroup vermezseniz kendisi outgroup seciyor. Outgroup'u kendiniz manual olarak tanitmaniz lazim komutunuzda. Bu onemli bir ayrinti!
Not: iqtree outgroup tanimada sorun yasiyordu.

-m TEST seceneginin ne anlama geldigini inceleyin

-asr secenegine bakin

-nt seceneginin ne anlama geldigine bakin

Model secimini (-m secenegi ile), filogenetik agac olusturma, ve atasal dizilerin tahmini (tek bir komut ile - derste isleri hizlandirmak icin bu yolu sectik):

$ iqtree -s adh4_primates_fasta.aln -o E_Tree.ADH.4_Tupaia_belangeri_ -m TEST -asr

Komut isleme koyulacak. Burada ekrana verilenleri anlamak incelemek cok onemli! Bilimsel yayin icin ornegin Model Finder'in yaptigi islemleri, farkli modellere verdigi skorlari (orn. likelihood score:-LnL vb.) anlamaniz lazim. Derste substitution/evrim modelleri ile filogenetik agaclari "gercege" nasil yaklastirdigimizi konustuk. Hangi modeller en ustte (en olasi) cikiyor? Ne anlama geliyor secilen model (orn. 'Dayhoff' en olasi model olarak hesaplanmis olabilir, dayhoff subs. model'i veya sizin calistirdiginizda cikan en olasi modeli arastirin, ne anlama geldigine bakin.

Ekrana FINALIZING TREE SEARCH altinda verilen raporlardaki log-likelihood verileri nedir (beyaz-tahta dersinde biraz konustuk), 'total number of iterations' ne anlama geliyor, total tree length farkli modeller ile agac cizerseniz nasil degisiyor, vb. ozelliklere bakmaniz, anlamaniz gerekiyor.

Peki elde edilen 4 dosya neler?
1. adh4_primates_fasta.aln.iqtree
2. adh4_primates_fasta.aln.treefile
3. adh4_primates_fasta.aln.mldist
4. adh4_primates_fasta.aln.state
5. adh4_primates_fasta.aln.log

less programi ile iqtree'nin verdigi en kapsamli filogeni dosyasini acin:

$ less adh4_primates_fasta.aln.iqtree

Burada istatistik verileri, filogenetik agac, outgroup konusunda bir bilgi, modeller, ve agacin "newick" format olarak cizimi de var.

Ornegin: "WARNING: 4 near-zero internal branches (<0.0027) should be treated with caution
Such branches are denoted by '**' in the figure below" - bu uyarinin neden verildigini agaca bakarak gorebilirsiniz.

 # 'q' ile cikabilirsiniz.
 # Derste "newick" formatini kopyala-yapistir ile phylogeny.fr'de agaci cizdirmistik.

adh4_primates_fasta.aln.iqtree dosyasinin uygulamanin hedefi baglaminda (atasal protein dizilerini bulmak) bize verdigi EN ONEMLI BILGI: .iqtree dosyasinda bash ekranina verilen agaca baktiginizda "node 1", "node 2" ile atasal node'lari goreceksiniz. Ornegin Homo sapiens, Pan paniscus, Gorilla gorilla, ve Pan troglodytes'lerin birlestigi node hangi rakam ile gosteriliyor? Iste bu rakam bu 4 turun ortak atasinda bulunan atasal ADH4 dizisini baska bir dosyadan bulup incelememizi saglayacak.

Son olarak olusturdugumuz atasal diziler icin .state dosyasina bakacagiz:

$ less adh4_primates_fasta.aln.state

less ile .state'i acinca mouse ile hareket edebilirsiniz dosyada – read-only olarak aciyorsunuz o nedenle dosya icerigini bozma sansiniz -iyi ki- yok!

dosyayi inceleyin, ne anlama geldigini kendiniz zaten anlayacaksiniz.

Sutun 1 Node#, sutun 2 proteinin bastan sona icerdigi tahmin edilen en olasi amino asit secimleri (347 amino asitlik bir proteindi ADH4, sanirim), sutun 4 ve sonrasi her olasi amino asit icin verilen olasilik skorlari olmali.

.state dosyasi her bir 'node'da bulunan atasal amino asid dizisini veriyor. Istediginiz bir atasal 'node'da bulunan atasal protein/DNA dizisini (orn. Node1) bash uygulama dersinde ogrendigimiz yontemlerle secip yeni bir dosyaya aktarabilirsiniz.

 # .state dosyasinda her bir amino asid icin ilgili 'node'da bulunan tahmini amino asid icin bir likelihood skoru veriliyor (1 ve daha dusuk skorlar). O skorlar size ileride makalenizde istatistik verilerini paylasirken kullanmaniz icin gerekli olacak.

 # Makalede bulunan en kritik fenotipik (enzim aktivitesi degisimine sebep olan) A294V mutasyonunu bulabildiniz mi? Inceleyin...

Not: Bu islem dersin suresine uygun olarak en basit yontemi ve sorunsuz (akrabalik olarak yakin dizileri) kullanarak yapildi. Tahmin edilen atasal amino asidlerin olasilik skorlarina baktiginizda zaten olasiliklarin genelde yuksek oldugunu goreceksiniz. Fakat tahmin edilen atasal sekanslarin/DNA'nin/protein'in atasal diziyi tamamen yansitmadigini, yuksel olasilikli bir tahmin dizisi oldugunu unutmamalisiniz (derste istatistik ile, dogruluk ile ilgili calismalardan bahsedildi, slaytlarda ilgili referanslari bulabilirsiniz, ki bu konuda calisacaksaniz o calismalari kavrayarak yola cikmaniz gerekiyor!). Sonuc olarak asil hedef ornegin 100 milyon yil onceki “gercek” atasal diziyi aynen bulmak degil (ki bu olasi degil) fakat en olasi diziyi bulup bu yuksek olasilikli (!) tahmini dizinin gercek diziden birkac fark icersede (farkli genotip) atasal FENOTIPI (atasal enzimin aktivitesini, atasal enzimin hucre icin konumunu vb) yansitiyor olmasi (bundan bahseden makalede yine slaytlarda atiflaniyor).

SON