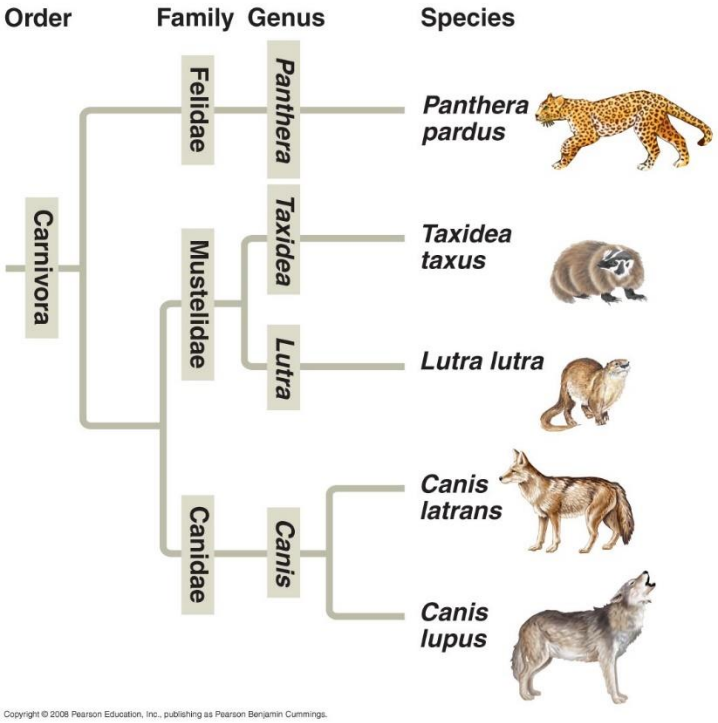
۱ تعاریف ابتدایی

۱-۱ درخت زندگی(Phylogenetic Tree)

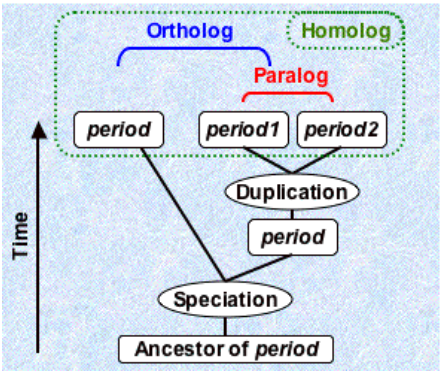
درخت زندگی: یک گراف به صورت درخت است که در آن روابط تکاملی(Evolutionary) بین موجودات(Species) مختلف نمایش داده می شود.



شکل ۱

دو ژن که از یک جد مشترک آمده باشند را Homolog گوییم. دو ژن Homolog که در دو موجود مختلف باشند را Ortholog نامیم. دو ژن Homolog در یک موجود را Paralog گوییم. پیدایش ژن‌های Paralog به دلیل آن است که ممکن است در فرآیند تکاملی یک قسمت از DNA، duزیرتیتر اول

جدول ۱- این یک جدول است.



شکل ۲

نگارنده: حامد واسعی
ایمیل: vasei.hamed@gmail.com

۲ به دست آوردن درخت

برای کشیدن درخت رندگی دو روش وجود دارد.

۱- برپایه فاصله (distance based)

۲- بر پایه کاراکتر (character based)

در این جلسه به ساخت درخت بر پایه فاصله می‌پردازیم. در این روش فرض می‌شود فاصله‌ی بین موجودات مختلف را در اختیار داریم و می‌خواهیم درختی وزن دار بسازیم، که فاصله‌ی بین دو موجود، برابر با وزن مسیر پیموده شده بر روی درخت باشد.

۲-۱ متریک additive و شرایط چهار نقطه‌ای

تعریف: یک متریک (X,d) را additive گوییم اگر درختی مانند T وجود داشته باشد که در آن یالها طول نامنفی و گره‌های X را در برگیرد و برای هر x,y درون X داشته باشیم

$$d(x,y) = d_T(x,y)$$

که $d_T(x,y)$ بهمعنای فاصله بر روی درخت است.

تعریف: یک متریک (X,d) در شرایط چهارنقطه ای صدق می کند اگر برای هر چهار نقطه، بزرگترین وزن تطابق یکتا نباشد یعنی برای هر چهار نقطه i,j,k,l و

$$A = d(i,j) + d(k,l), \quad B = d(i,k) + d(j,l), \quad C = d(i,l) + d(j,k)$$

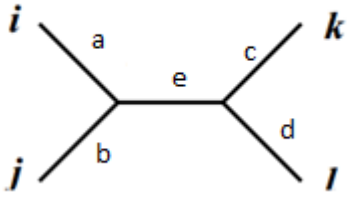
داشته باشیم:

$$A \leq \max(B,C), B \leq \max(A,C), C \leq \max(A,B)$$

قضیه: یک متریک additive است اگر و تنها اگر شرایط چهارنقطه‌ای برقرار باشد.

اثبات:

- ابتدا ثابت می کنیم اگر یک متریک M ، additive باشد آنگاه در شرط چهار نقطه صدق می کند:
چون متریک additive است پس یک درحت زندگی برای موجودات وجود دارد. چهار موجود مختلف i,j,k,l را در نظر بگیرید. کلی ترین حالت ممکن برای این چهار موجود به صورت زیر است(البته چون برچسب های i,j,k,l مهم نیستند تنها یک شکل در نظر می گیریم):



به راحتی دیده می شود که:

$$M_{ij} + M_{kl} = a + b + c + d \leq a + b + c + d + 2e = M_{il} + M_{jk} = M_{ik} + M_{jl}$$

- حال باید ثابت کنیم که شرط چهارنقطه ای additive بودن را نتیجه می دهد:

برای این منظور از استقرا استفاده می کنیم. پایه استقرا را n=4 در نظر می گیریم. فرض کنید ۴ نقطه I,j,k,l داریم که در روابط

$$M_{ij} + M_{kl} \leq M_{il} + M_{jk} = M_{ik} + M_{jl}$$

صدق می کنند. داریم:

$$2e = M_{il} + M_{jk} - \left(M_{ij} + M_{kl}\right) \geq 0$$
$$2a = M_{ij} + M_{ik} - M_{jk} \geq 0$$
$$2b = M_{ij} + M_{jk} - M_{ik} \geq 0$$
$$2c = M_{kl} + M_{ik} - M_{il} \geq 0$$
$$2d = M_{kl} + M_{li} - M_{ik} \geq 0$$

پس مقادیر مثبت a,b,c,d,e همگی به دست آمده و درختی مانند شکل بالا تشکیل می‌شود. حال فرض می کنیم درخت برای کمتر از n موجود قابل ساختن باشد. ثابت می کنیم برای n موجود نیز می توان این درخت را کشید.

(از اینجا به بعد برای خوانا تر بودن به جای M_{ij} می‌نویسیم $d(i,j)$)

اگر S مجموعه موجودات شامل n موجود باشد، از بین تمامی سه تایی های مرتب از اعضای S، سه تایی مرتب (p,q,r) را انتخاب می کنیم که برای آن مقدار زیر بیشینه باشد:

$$d(p,r) + d(q,r) - d(p,q)$$

با این حساب داریم

$$\forall x \in S - \{p,q,r\}$$
$$d(x,r) + d(q,r) - d(x,q) \leq d(p,r) + d(q,r) - d(p,q) \rightarrow$$
$$d(x,r) + d(p,q) \leq d(x,q) + d(p,r)$$

به صورت مشابه داریم:

$$d(x,r) + d(p,q) \leq d(x,p) + d(q,r)$$

حال با استفاده از شرط چهار نقطه برای x,p,q,r و نتیجه گیری از دو نامساوی بالا می توانیم بنویسیم:

$$d(x,q) + d(p,r) = d(x,p) + d(q,r)$$

با توجه به بدیهی بودن رابطه بالا برای x=r این رابطه $\forall x \in S - \{p,q\}$ درست است. با نوشتن این رابطه برای $x,y \in S - \{p,q\}$ و گرفتن تفاضل می‌توان نوشت:

$$d(y,p) + d(x,q) = d(x,p) + d(y,q) \qquad (1)$$

حال موجود جدید t را در نظر بگیرید.رابطه d را برای آن به صورت زیر گسترش می دهیم:

$$\begin{cases} d(t,p) = \dfrac{d(p,q) - d(p,r) - d(q,r)}{2} \\ d(t,x) = d(x,p) - d(t,p) \qquad x \neq p \end{cases}$$

که نتیجه می‌دهد:

$$d(p,x) = d(p,t) + d(t,x) \qquad (2a)$$

علاوه بر این طبق معادله (1) می توانیم بنویسیم:

$$\begin{aligned} d(q,x) &= d(p,x) + \quad d(q,r) - d(p,r) \\ &= d(t,x) + d(t,p) + d(q,r) - d(p,r) \\ &= d(t,x) + \quad d(q,t) \qquad (2b) \end{aligned}$$

ادعا آن است که $S + \{t\}$ و متریک گسترش یافته d در شرط چهار نقطه صدق می‌کند. برای اثبات این مطلب کافیست ثابت کنیم اگر t کی از چهارنقطه باشد شرط چهارنقطه برقرار است:

دو حالت زیر را در نظر بگیرید:

- چهار نقطه شامل t,p,x,y باشد:

$$d(t,x) + d(p,y) = d(x,p) + d(p,y) - d(t,p)$$
$$d(t,y) + d(p,x) = d(x,p) + d(p,y) - d(t,p)$$

پس داریم:

$$d(t,x) + d(p,y) = d(t,y) + d(p,x)$$

پس کافیست ثابت کنیم که

$$d(t,p) + d(x,y) \leq d(t,x) + d(p,y)$$

می دانیم:

$$d(t,x) + d(p,y) = d(q,x) + d(p,y) - d(q,t)$$

پس باید ثابت کنیم:

$$d(q,t) + d(p,t) + d(x,y) \leq d(q,x) + d(p,y)$$

اما می دانیم:

$$d(q,t) + d(p,t) = d(p,q)$$

پس باید ثابت کنیم:

$$d(p,q) + d(x,y) \leq d(q,x) + d(p,y)$$

که این موضوع نیز با توجه به رابطه (1) و نیز رابطه چهر نقطه برای مجموعه S برقرار است.

- چهارنقطه شامل t,x,y,z باشد و شامل p نباشد:

$$d(t,x) + d(y,z) = d(x,p) + d(y,z) - d(t,p)$$
$$d(t,x) + d(y,z) = d(y,p) + d(x,z) - d(t,p)$$
$$d(t,x) + d(y,z) = d(z,p) + d(x,y) - d(t,p)$$

که چون p,x,y,z در شرایط چهار نقطه ای صدق می کنند در شرایط چهار نقطه ای صدق می کند.

پس ثابت شد مجموعه جدید با d شرایط چهار نقطه ای را ارضا می کند پس d گسترش یافته شرایط متریک را نیز دارد. اگر مجموعه $S - \{p,q\} + \{t\}$ را در نظر بگیریم این مجموعه حداکثر n-1 عضو دارد(ممکن است کمتر از n-1 عضو داشته باشد هنگامی که نقطه ای در S باشد که فاصله اش از هر نقطه با فاصله t از آن نقطه برابر باشد که آن دو موجود را باید یک موجود در نظر بگیریم) که در شرایط چهار نقطه ای صدق می کنند پس سبق فرض استقرا می توانیم یک درخت برای آن بسیازیم. حال باید نقاط p,q را به گونه ای در درخت قرار دهیم که فاصله آنها روی درخت از تمام نقاط S برابر با فاصله ی آنها طبق d باشد. این کار با مجاور کردن p,q با t در درخت ساخته شده به ترتیب با فواصل d(p,t),d(q,t) ممکن است و طبق روابط (2a) و (2b)شرط خواسته شده برآورده می شود.

اگر شرایط چهار نقطه‌ای برقرا نباشد، نمی‌توانیم یک درخت Additive بسازیم پس به دنبال درختی خواهیم گشت که

$$SSE(T) = \sum_{i \neq j} w_{ij}\big(d(i,j) - d_T(i,j)\big)^2$$

برای آن کمینه باشد. ساخت چنین درختی در حالت کلی NP-Complete است لذا برای ساخت درخت، از روش‌های heuristic استفاده می‌کنیم.

۲-۲ الگوریتم Neighbor Joining

در الگوریتم Neighbor-Joining هدف آن است که درخت را تقریب بزنیم و در صورتی که واقعا additive بود به درخت اصلی برسیم. ایده اصلی آن است که موجوداتی که به یکدیگر نزدیک هستند و از بقیه دورتر هستند را به یکدیگر وصل کنیم.

الگوریتم **Neighbor-Joining(NJ)**:

آغاز:

۱- N را قرار بده برگ های درخت و هر موجود را به یکی از آنها وصل کن.

۲- قرار بده n=L

تکرار:

۱- جفت i,j که دارای کمترین d_{ij} می‌باشد را یافته و گره جدید k را با فواصل زیر اضافه کن:

$$d_{km} = \frac{1}{2}\big(d_{im} + d_{jm} - d_{ij}\big) \quad \forall m \in L, m \neq i,j$$

$$d_{ik} = \frac{1}{2}\big(d_{ij} + r_i - r_j\big)$$

$$d_{jk} = d_{ij} - d_{ik}$$

که در آن:

$$r_x = \frac{1}{|L| - 2}\sum_{k \in L} d_{xk}$$

۲- K را به i,j وصل کن.

۳- i,j را حذف کرده و k را به درخت اضافه کن.

پایان

۱- اگر L تنها دو گره داشت آن‌گاه گره‌های باقیمانده را با فواصل مربوطه به هم وصل کن.

۲-۳ تعداد درخت‌های ممکن

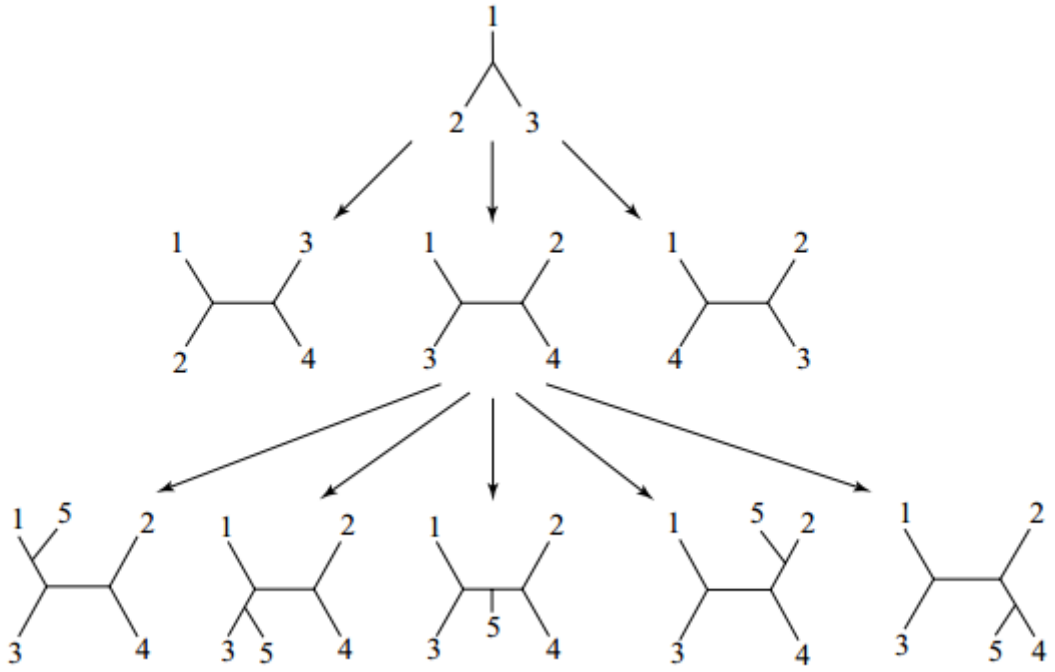
قضیه: تعداد درخت‌هایی دودویی بدون ریشه برای n موجود برابر !!(2n-5) و تعداد درخت‌های ریشه دار برای n موجود برابر !!(2n-3) است که در آن

$$(2k-1)!! = (2k-1) \times (2k-3) \times \ldots \times (3) \times (1)$$

تعداد درخت های دودویی بی ریشه :

برای به دست آوردن این عدد از استقرا استفاده می کنیم. این عدد برای n>2 موجود برابر است با حاصل ضرب اعداد فرد از 1 تا 2n-5 و هر درخت برای n موجود شامل n برگ و 2n-3 یال است.

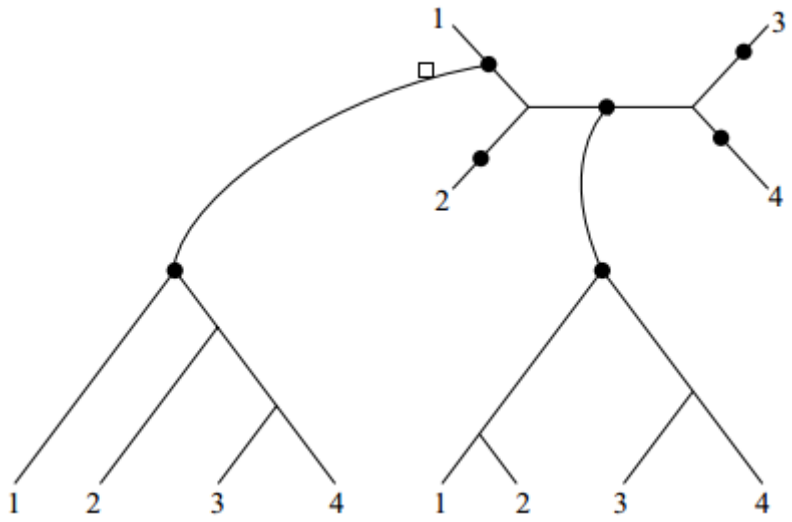
پایه استقرا n=3 می باشد. حکم برای پایه به وضوح برقرار است.



با فرض حکم برای k، جکم را برای k+1 ثابت می کنیم. برای ساخت یک درخت دودویی برای k+1 موجود باید یک برگ به یک درخت دودویی با k برگ اضافه کنیم به گونه ای که این برگ جدید با یک یال به وسط یکی از یال های درخت با k برگ متصل باشد. برای این کار 2k-3=2(k+1)-5 حالت ممکن داریم که البته تمام این درخت های ممکن از یکدیگر متمایز هستند چون برگ ها برچسب دارند و در هر کدام از این حالات فاصله برچسب ها از برگ اضافه شده با یکدیگر متمایز است و چون تعداد درخت های k برگی حاصل ضرب اعداد فرد از 1 تا 2k-5 است با ضرب در 2(k+1)-5 حکم حاصل می شود. علاوه بر این با این کار دو یال به بالهای درخت k برگی که طبق فرض تعدادشان 2k-3 می باشد اضافه می شود و تعدادشان 2(k+1)-3 می شود.

تعداد درخت های دودویی ریشه دار:

برای به دست آوردن این تعداد کافی است یک درخت بدون ریشه برای n موجود را در نظر بگیریم. این درخت را به 2n-3 روش متفاوت می توان ریشه دار کرد چون با در نظر گرفتن ریشه بر روی هر کدام از یال های درخت بدون ریشه به یک درخت ریشه دار دودویی متمایز برای n موجود می رسیم(این که همه ی درخت های ریشه دار را به این روش می شماریم واضح است چون ار هر درخت ریشه دار می توانیم با در نظر نگرفتن ریشه به عنوان یک راس به یک درخت بدون ریشه برسیم) پس تعداد درخت های با ریشه برابر است با حاصل ضرب اعداد فرد از 1 تا 2n-3.


1
3
2
4
1
2
3
4
1
2
3
4