A Hipótese do Gráfico de Dependência - Como ela é Inferida

Por **Andrew Jones** *(adaptação)*

Como Brian Miller e Cornelius Hunter já observaram, um extraordinário artigo científico acaba de ser publicado por Winston Ewert. Escrevendo na revista *BIO-Complexity*, Ewert propõe que a vida é melhor explicada não pela hipótese de Darwin de uma árvore de ancestralidade, mas por uma hipótese moderna inspirada no design de um gráfico de dependência. O modelo do gráfico de dependência foi explicado aqui. O artigo original está aqui.

Eu quero explicar a matemática e a filosofia que nos permitem determinar objetivamente qual modelo tem o poder explicativo superior.

Da simplicidade à probabilidade

Queremos escolher o modelo que explica os dados mais simples.

O que é simplicidade? É a ausência de complexidade, a ausência de partes ou a ausência de opções. Outra maneira de ver isso é que há muito mais maneiras de se fazer um modelo complexo do que um modelo simples. Portanto, uma maneira de medir a complexidade de um modelo é contar quantas maneiras existem para construir um modelo como este. Então, assumindo que não sabemos os detalhes de como os processos naturais gerariam o padrão específico (ou como um agente inteligente o escolheria), assumimos que cada padrão em particular tem a mesma probabilidade de ser gerado (ou escolhido). Em outras palavras, a probabilidade de qualquer modelo em particular é o inverso do número de modelos semelhantes. Em geral, modelos simples apresentam menor número de possibilidades e são intuitivamente considerados mais prováveis. Por outro lado, modelos complexos apresentam número de possibilidades muito maior, e, portanto, são considerados menos prováveis. Ainda uma outra maneira de olhar para isso é que a complexidade é um custo que queremos evitar: modelos simples são **parcimoniosos**.

Essa correspondência entre complexidade e improbabilidade é muito útil, e podemos usar os dois conceitos de forma intercambiável para ajudar a organizar nosso pensamento. E se você tiver vários modelos complexos, mas semelhantes, que explicam bem os dados? Você agrupa-os em um modelo mais provável, que é um modelo mais simples com menos elementos específicos (*ad hocs*).

No Modelo Bayesiano, o melhor modelo é aquele que torna os dados mais prováveis. Não faz sentido ter um modelo simples se ele não explicar os dados. Da mesma forma, não faz sentido ter

um modelo mais complexo do que os dados que ele precisa explicar. Isso seria um sobreajuste. A complexidade geral é a probabilidade do modelo ser combinado com a probabilidade dos dados de um certo modelo. No artigo de Ewert, existem dois modelos abrangentes que queremos distinguir: a **árvore de ancestralidade** e o **gráfico de dependência**, mas há uma miríade de possíveis submodelos, cada um contribuindo para a probabilidade geral do modelo abrangente.

Valores médios e Priores Bayesianos

Infelizmente, ambos os modelos (a **árvore da vida** e o **gráfico de dependência**) são extremamente complexos, com um número muito grande de parâmetros ajustáveis. Isso pode parecer tornar a questão indecidível: muitas vezes argumentamos que a árvore da vida é um ajuste terrível para os dados, exigindo numerosos “epiciclos” ad-hoc para fazer os dados se encaixarem (veja mais sobre isso). Podemos argumentar ainda que um gráfico de dependência específico é mais adequado aos dados. Mas um defensor da ancestralidade comum pode responder razoavelmente que nossa teoria também não é parcimoniosa; se você adicionar módulos suficientes, você poderia explicar literalmente qualquer coisa, até mesmo dados aleatórios. Parece que decidir entre os dois modelos nunca poderia ser uma decisão racional; parece que sempre envolverá muita intuição ou até mesmo fé. Felizmente, no entanto, existem maneiras de domar a complexidade o suficiente para obter respostas objetivas e significativas.

A principal estratégia para lidar com a complexidade é somar (ou integração matemática) totalmente as possibilidades. Ewert lida com muitos dos parâmetros por integração: estes incluem a probabilidade de borda *b* (a *conectividade* esperada) dos nós e as diferentes propensões para adicionar ? ou perder ? genes em cada um dos *n* nós. Isso pode parecer estranho, mas é um raciocínio probabilístico padrão. Se a distribuição de probabilidade de Y (por exemplo, o número real de perdas genéticas) depende de X (por exemplo, ?), mas você não sabe X, ainda é possível calcular a probabilidade de Y *se* você tem a distribuição de probabilidade de X. Em muitos casos, nem sabemos qual seria a verdadeira distribuição de X. Em tais casos, Ewert assume que toda possibilidade tem uma probabilidade igual (uma distribuição plana), porque isso deve introduzir o mínimo de viés. Isso também pode parecer estranho, mas é bastante comum no raciocínio *bayesiano*. Embora a distribuição prévia de X seja tecnicamente uma *escolha*, e sim que essa escolha tenha alguma influência no resultado, a maneira como o raciocínio Bayesiano funciona é que quanto mais dados você adiciona, menor é a influência nos resultados. O importante é escolher um valor que não seja *tendencioso*; que permita que os dados falem. O raciocínio *bayesiano* é importante na medida em que nos dá alguma esperança de escapar da tirania do viés da confirmação e dos pressupostos dogmáticos.

A ideia é que queremos ter certeza de que as muitas coisas que não sabemos não nos impedem de fazer inferências razoáveis ??usando o que sabemos.

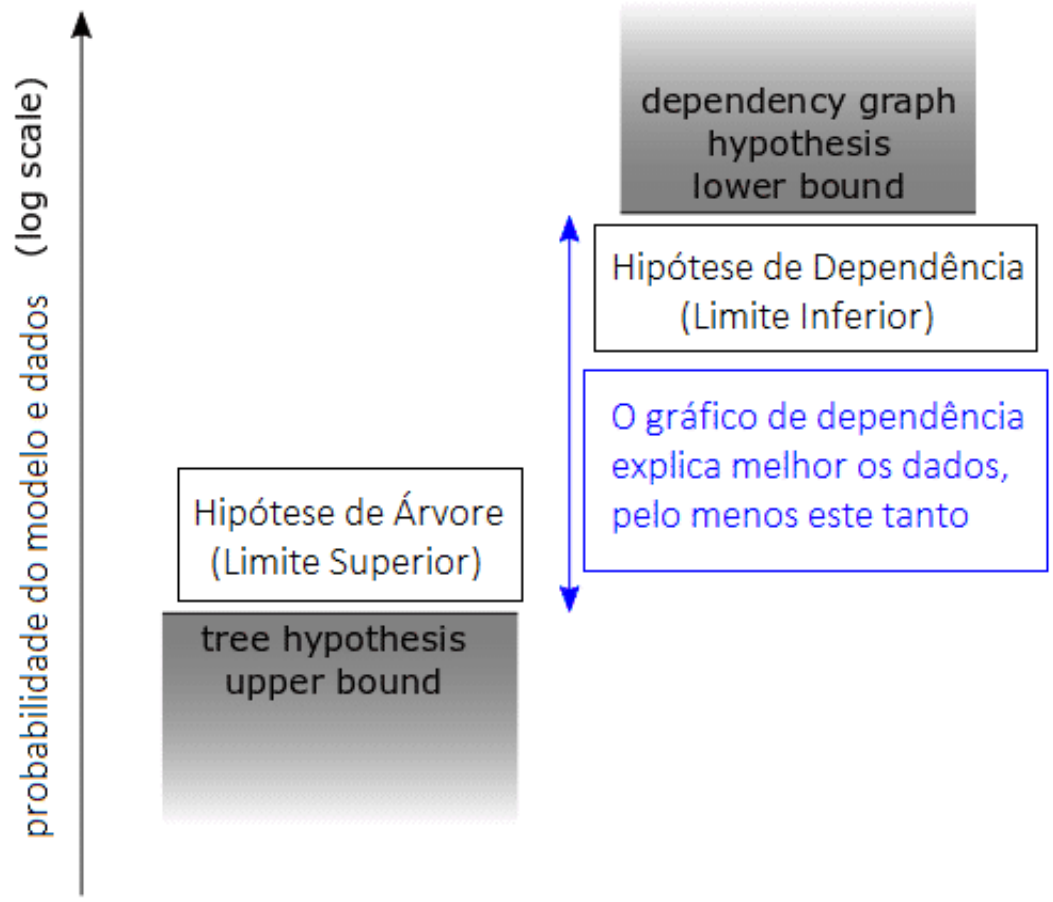
Calcular limites

Embora algumas das variáveis ??possam ser integradas matematicamente, outras não podem. Este é um problema para ambos os modelos, uma vez que (em teoria) devemos somar todas as probabilidades de todas as possíveis atribuições de genes para **nós** e deleções de **nós**. Mas é um problema específico para o gráfico de dependência: há diferentes números possíveis de **nós** no gráfico de dependência e uma miríade de maneiras possíveis de conectar cada número de **nós**. Para obter uma probabilidade precisa para o modelo de gráfico de dependência abrangente, precisaríamos somar cada uma dessas configurações possíveis. Na prática, essas somas são impossíveis de fazer; há muitos demais.

No entanto, mesmo quando não é possível calcular uma probabilidade precisa, muitas vezes é possível calcular ***limites***. Esta é uma maneira antiga de provar as coisas sem fazer cálculos difíceis, especialmente em matemática.

Qual é a probabilidade de um bule de chá se unir em átomos em órbita ao redor do sol? Ninguém sabe. Mas, por exemplo, você poderia dizer que é pelo menos tão improvável quanto uma coleção de 100g de átomos que supera a entropia termodinâmica. Este é um limite superior na probabilidade, e é algo que você *poderia* calcular com bastante facilidade (e, como se vê, é extremamente improvável; efetivamente impossível).

Para mostrar que o grafo de dependência é uma explicação melhor, Ewert estima um limite *superior* na probabilidade dos dados pela hipótese de árvore ancestral, e calcula um limite *inferior* na probabilidade dos dados dados a hipótese do gráfico de dependência, e então veja a diferença.



Complexidade do gráfico

O modelo de árvore tem uma grande vantagem, pois é mais simples que o modelo de gráfico de dependência. Se houver espécies S, o modelo de árvore possui espécies ancestrais (S-1). Em contraste, o gráfico de dependência pode ter até $N_{max} = (2^S - 1 - s)$ módulos opcionais. À medida que S aumenta, essa diferença cresce exponencialmente. Por exemplo, onde existem apenas 4 espécies (como nas figuras abaixo), são até 3 espécies ancestrais, mas até 11 módulos. Já onde existem 10 espécies, são 9 espécies ancestrais, mas até 1013 módulos possíveis! Isso é muita complexidade extra.

Em segundo lugar, as conexões adicionam complexidade. Uma árvore de N **nós** tem exatamente (N-1) conexões, mas um gráfico de dependência de N **nós** pode ter até N×(N-1)/2 conexões possíveis. Cada uma dessas conexões pode existir ou não, dando um grande número de possibilidades. A probabilidade de qualquer combinação particular de conexões é muito pequena, o que se traduz em uma probabilidade muito pequena para qualquer gráfico particular.

Este gráfico de dependência não é real; é apenas minha impressão artística para mostrar o aumento potencial em complexidade.

Para o modelo de árvore, Ewert supõe que os biólogos fizeram um bom trabalho sob as restrições de seu modelo e já encontraram um bom ajuste de qualidade, então ele pega a árvore ancestral hipotética desenvolvida por biólogos especialistas encontrados no NCBI. Como estamos calculando limites, não probabilidades precisas, e porque a árvore é muito mais simples, o cálculo pode ser simplificado dando à árvore uma probabilidade condicional de 1 (em palavras: se a hipótese da árvore for verdadeira, então essa árvore é a árvore correta). Em contraste, todo grafo de dependência é tratado como altamente improvável; há uma grande penalidade apenas por ser um gráfico de dependência.

Se tudo fosse igual, seria muito mais parcimonioso escolher a árvore ao invés do gráfico de dependência. Então, por que alguém optaria pelo gráfico de dependência? Por causa dos dados.

Complexidade dos dados dado o gráfico

Os dois modelos recebem os mesmos dados para explicar. Os dados que Ewert escolheu foram a distribuição de famílias de genes em espécies. Essas famílias de genes também foram tiradas de categorias e dados criados por biólogos especialistas, encontrados em nove bancos de dados públicos diferentes.

O algoritmo de Ewert então atribui essas famílias de genes a espécies ancestrais na árvore, ou módulos no gráfico de dependência, para otimizar a probabilidade de cada um. Ewert segue a hipótese de parcimônia baseada na Lei de Dollo, o que significa que cada família de genes aparece apenas uma vez na história da vida, ou em não mais do que um módulo projetado.

O algoritmo inicialmente atribui cada família de genes ao suposto ancestral comum (de acordo com a hipótese da árvore) e, em seguida, tenta melhorar isso. A árvore e essas atribuições de genes também são usadas como um primeiro palpite para o gráfico de dependência. Por padrão, os **nós** obtêm todos os genes dos **nós** superiores aos quais estão conectados. No caso da árvore, isso significa que espécies descendentes **herdam** genes de espécies ancestrais. No caso do gráfico de dependência, isso significa que os módulos importam genes de suas *dependências*. Para ajustar o gráfico, alguns genes podem precisar ser excluídos uma vez ou muitas vezes dos **nós** inferiores[1].

Nota do tradutor [1]: há um texto abordando isso neste portal: Um Padrão Suspeito de Deleções.

Ewert supõe que há uma certa probabilidade de adicionar um gene a cada **nó**, mas como não sabemos o qual é, ele calcula a média de todas as probabilidades possíveis. A fórmula também é dada na **equação 4** do artigo. Ele também assume que as deleções gênicas acontecem com uma certa probabilidade em cada **nó**, mas, como não sabemos qual é a probabilidade, ele calcula a média de todas as probabilidades possíveis. A fórmula também é dada na **equação 6** do artigo. O

efeito é que a primeira adição/exclusão custa uma penalidade que depende do número de genes e, em seguida, as adições/exclusões subsequentes custam um pouco menos e assim por diante. A consequência é que pode ser mais provável adicionar um gene mais acima na árvore/gráfico ou apagá-lo mais abaixo no gráfico, mesmo que isso signifique excluí-lo mais vezes. Portanto, há também um algoritmo que tenta mover atribuições de famílias de genes para cima da árvore ou gráfico, e/ou exclusões na árvore ou gráfico, se isso daria uma probabilidade melhor. No caso dos gráficos de dependência, o algoritmo também cria ou remove módulos inteiros onde isso daria melhores probabilidades. Este último algoritmo é o que permite a descoberta de **gráficos de dependência que são radicalmente diferentes de uma árvore**.

Considere esses dois conjuntos imaginários de dados. Cada linha ou cor é uma família de genes e cada coluna é uma espécie. Um círculo preenchido indica que essa espécie tem um representante dessa família genética em seu genoma.

O primeiro conjunto de dados (à esquerda acima) se ajusta perfeitamente a uma árvore (abaixo à esquerda). O gráfico de dependência ideal (abaixo à direita) parece idêntico à árvore, mas devido às penalidades extras por ser um gráfico de dependência, ele perde: a melhor explicação para esses dados é uma árvore.

Para o segundo conjunto de dados, ainda podemos encaixar uma árvore (abaixo à esquerda), mas é uma bagunça. Temos que deletar genes tantas vezes para que se encaixe que a hipótese se torna bastante improvável. Em contraste, o algoritmo de localização do gráfico de dependência encontra um conjunto de módulos sem exclusões. Como resultado, o gráfico de dependência (abaixo à direita) ganha para esse conjunto de dados. Na prática, com dados mais complicados, o melhor gráfico de dependência também pode envolver algumas exclusões. (Só porque um módulo é importado não significa necessariamente que tudo nele é usado, e também é possível que uma espécie projetada tenha perdido genes que já teve).

O método não precisa encontrar *o* melhor gráfico de dependência, porque sabemos que o melhor gráfico de dependência é pelo menos tão bom quanto o melhor que encontramos e, portanto, o gráfico de dependência é pelo menos muito melhor do que a hipótese da árvore[2].

Nota do tradutor [2]: Creio que um modelo ainda diferente corresponda a vida, uma sobreposição de árvores que pode ser concluída da interpretação de As Árvores Informacionais da Vida .

Aplicando o Método

É assim que o método de seleção de modelos funciona em princípio. No artigo, você pode ver que Ewert testou o método gerando gráficos a partir de software real, de várias simulações evolutivas e de dados aleatórios (uma hipótese nula onde os genes de cada espécie foram sorteados

aleatoriamente de um único conjunto comum). Ele descobre que identifica corretamente as árvores a partir de processos evolutivos (embora a evolução simulada) e identifica corretamente os gráficos de dependência do software compilado e identifica corretamente os dados aleatórios também.

A coisa realmente excitante é o que acontece quando ele olha para dados *biológicos*. Por mais diferente que a biologia possa ser do software, o estranho resultado é que a grande figura da biologia parece ser *objetivamente* muito mais semelhante ao software do que à evolução.

Um último comentário

Finalmente, uma razão pela qual eu amo este artigo é que ajuda a explicar por que a árvore da vida de Darwin foi aceita por muitas pessoas racionais ??até agora. A razão é: quando não há muitos dados, o modelo mais simples ganha por padrão. Mas agora que temos mais dados biológicos do que nunca, e agora que nossa própria compreensão de design/engenharia/tecnologia aumentou de maneiras que ninguém nunca poderia ter imaginado, podemos começar a ver que um modelo um pouco mais complexo pode ser uma explicação muito mais poderosa.

Original: Andrew Jones. The Dependency Graph Hypothesis — How It Is Inferred. July 23, 2018.