Artigo de Winston Ewert Demonstra a Superioridade do Modelo do Design

Por **Cornelius Hunter**

Você sabia que Marte está retrocedendo? Nas últimas semanas e por várias semanas, Marte está em fase de movimento retrógrado. Se você traçar sua posição a cada noite contra as estrelas no fundo, você o verá pausar, inverter a direção, pausar novamente e, em seguida, continuar em sua direção normal.

E você sabia ainda que o movimento retrógrado ajudou a causar uma revolução? Dois milênios atrás, a física aristotélica ditava que a Terra estava no centro do universo. O modelo heliocêntrico de Aristarco, que colocou o Sol no centro, caiu em desgraça. Mas o que o geocentrismo de Aristóteles não conseguiu explicar foi o movimento retrógrado. Se os planetas estão girando em torno da Terra, então por que eles às vezes param e invertem a direção? Esse problema recaiu sobre Ptolomeu e as lições aprendidas ainda são importantes hoje em dia.

Ptolomeu explicou anomalias como o movimento retrógrado com mecanismos adicionais (*ad hocs*), como os epiciclos, mantendo o movimento circular que, como todos sabiam, deve ser a base de todo movimento no cosmos. Com menos de cem epiciclos, ele foi capaz de modelar e prever com precisão os movimentos do cosmos. Mas essa precisão teve um custo - um modelo altamente complicado.

Um modelo melhor

Na Idade Média, Guilherme de Occam salientou que as teorias científicas deveriam buscar simplicidade ou parcimônia. Este pode ter sido um dos fatores que levou Copérnico a ressuscitar o modelo heliocêntrico de Aristarco. Copérnico preservou o movimento circular necessário, mas ao mudar para um modelo centrado no sol, ele conseguiu reduzir bastante o número de mecanismos adicionais, como os epiciclos.

Ambos os modelos de Ptolomeu e Copérnico previram com precisão o movimento celestial. Mas Copérnico foi mais parcimonioso. Um modelo melhor foi encontrado.

Kepler propôs elipses e mostrou que o modelo heliocêntrico poderia se tornar ainda mais simples. Não foi bem aceito, porém, porque, como todos sabiam, os corpos celestes viajam em círculos. Quão tolo pensar que eles viajariam por caminhos elípticos. O próximo passo em direção a uma maior parcimônia teria que esperar por nomes como Newton, que mostrou que as elipses de

Kepler eram ditadas por sua nova física altamente parcimoniosa. Newton descreveu uma lei gravitacional simples e universal. A força gravitacional de Newton produziria uma aceleração que poderia manter o movimento orbital no cosmos.

Acurácia e parcimônia

O ponto aqui é que a acurácia de uma teoria científica, por si só, significa muito pouco. Deve ser considerada juntamente com parcimônia. Esta lição é importante hoje nesta era dos grandes conjuntos de dados. Os analistas sabem que um modelo sempre pode ser mais preciso adicionando mais termos. Mas esses termos adicionais são significativos ou são meramente "epiciclos"? Parece ser bom reduzir o erro de modelagem para zero adicionando termos, mas quando usados ??para fazer previsões futuras, esses modelos têm um desempenho pior.

Existe uma penalidade muito real ao adicionar termos e violar a Navalha de Occam, e hoje algoritmos avançados estão disponíveis para estimar o balanço entre a acurácia e a parcimônia do modelo.

Isso nos leva à descendência comum, uma teoria popular para modelar as relações entre as espécies. Como já discutimos muitas vezes, a descendência comum não consegue modelar a espécie e muitos mecanismos adicionais - epiciclos biológicos - são necessários para ajustar os dados.

E assim como a cosmologia viu uma corrente de modelos cada vez melhores, os modelos biológicos também podem melhorar. Nesta semana, um modelo muito importante foi proposto em um novo artigo, já mencionado aqui por Brian Miller. É de autoria de Winston Ewert, na revista BIO-Complexity.

Três tipos de dados

Inspirada em software de computador, a abordagem de Ewert modela as espécies como módulos de compartilhamento que são relacionados por um gráfico de dependência. Esse modelo útil na ciência da computação também funciona bem na modelagem da espécie. Para avaliar essa hipótese, a Ewert usa três tipos de dados e avalia o quanto eles são prováveis ??(considerando a parcimônia e a acurácia do ajuste) usando três modelos.

Os três tipos de dados de Ewert são: (i) exemplo de software de computador, (ii) dados de espécies simuladas gerados a partir de algoritmos de computador de descendência comum/evolutiva e (iii) dados de espécies reais.

Os três modelos de Ewert são: (i) um modelo nulo que não implica relações entre qualquer espécie, (ii) um modelo de descendência evolutiva/comum e (iii) um modelo de gráfico de dependência.

Os resultados de Ewert são a Revolução Copernicana do momento. Primeiro, para os dados do software de exemplo, não é de se surpreender que o modelo nulo tenha um desempenho ruim. O software de computador é altamente organizado e há relacionamentos entre diferentes programas de computador e como eles se baseiam em bibliotecas de software fundamentais. Mas, comparando os modelos de descendência comum e gráfico de dependência, este último se adéqua muito melhor na modelagem de “espécies” de software. Em outras palavras, o design e desenvolvimento de software de computador é muito melhor descrito e modelado por um gráfico de dependência do que por uma árvore de descendência comum.

Em segundo lugar, para os dados de espécies simuladas gerados com um algoritmo de descendência comum, não é surpreendente que o modelo de descendência comum fosse muito superior ao gráfico de dependência. Isso seria verdade por definição e serve para validar a abordagem de Ewert. A descendência comum é o melhor modelo para os dados gerados por um processo esse processo de ancestralidade comum.

Terceiro, para os dados das espécies reais, o modelo do gráfico de dependência é astronomicamente superior comparado ao modelo de descendência comum.

Em que isso implica

Deixe-me repetir. Onde ele contava, a descendência comum falhou em comparação com o modelo de gráfico de dependência. Os outros tipos de dados serviram como verificações úteis, mas para os dados que importavam - os dados reais das espécies biológicas - os resultados não foram ambíguos.

Ewert acumulou um total de nove bancos de dados genéticos massivos. Em cada um deles, sem exceção, o modelo de gráfico de dependência superou a descendência comum.

Darwin nunca poderia ter sonhado com um teste em escala tão massiva. Darwin também nunca poderia ter sonhado com a magnitude do fracasso de sua teoria. Porque você vê, os resultados de Ewert não revelam dois modelos competitivos com um modelo desbancando o outro.

Não estamos falando de algumas diferenças de pontos decimais. Para um dos conjuntos de dados (HomoloGene), o modelo de gráfico de dependência foi superior à descendência comum por um fator de 10.064. A comparação dos dois modelos rendeu uma preferência pelo modelo de gráfico de dependência maior que dez mil!

Dez mil é um grande número. Mas fica pior, muito pior.

Ewert usou o modelo de seleção bayesiano que compara a probabilidade do conjunto de dados a partir de modelos hipotéticos. Em outras palavras, dado o modelo (gráfico de dependência ou descendência comum), qual é a probabilidade desse conjunto de dados específico? O modelo de

seleção bayesiano compara os dois modelos dividindo essas duas probabilidades condicionais. O chamado fator de Bayes é o quociente produzido por essa divisão.

O problema é que o modelo de descendência comum é tão incrivelmente inferior ao modelo de gráfico de dependência que o fator Bayes não pode ser digitado. Em outras palavras, a probabilidade do conjunto de dados, a partir do modelo do gráfico de dependência, é muito maior do que a probabilidade do conjunto de dados a partir do modelo de descendência comum, de modo que não podemos digitar o quociente de sua divisão.

Em vez disso, Ewert apresenta o *logaritmo* do número. Lembre-se de logaritmos? Lembra de como 2 realmente significa 100, 3 significa 1.000 e assim por diante?

Inacreditavelmente, o valor de 10.064 é o logaritmo (com base 2) do quociente! Em outras palavras, a probabilidade dos dados no modelo de gráfico de dependência serem muito maiores do que os dados no modelo de descendência comum, precisamos de logaritmos até mesmo para expressá-los. Se você tentou digitar o número simples, você teria que digitar um 1 seguido por mais de 3.000 zeros. Essa é a proporção de quão provável os dados estão entre esses dois modelos!

Usando um valor base de 2 no logaritmo, expressamos o fator Bayes em bits. Portanto, a probabilidade condicional para o modelo de gráfico de dependência tem uma vantagem de 10.064 sobre a de descendência comum.

10,064 bits está longe, longe do alcance em que se pode realmente considerar o modelo menor. Veja, por exemplo, na página da Wiki mesmo o fator Bayes, que explica que um fator Bayes de 3,3 bits fornece evidência “substancial” para um modelo, 5,0 bits fornecem “forte” evidência e 6,6 bits fornecem evidência “decisiva”.

Isto é ridículo. Considera-se que 6.6 bits fornecem evidência “decisiva” e, quando o caso do modelo de gráfico de dependência é comparado ao caso de ancestralidade comentado acima, obtemos 10.064 bits.

Mas fica pior

O problema com tudo isso é que o fator Bayes de 10.064 bits para o conjunto de dados no HomoloGene é o melhor caso para ancestralidade comum. Para os outros oito conjuntos de dados, os fatores de Bayes variam de **40.967** a **515.450**!

Em outras palavras, enquanto 6,6 bits seriam considerados para fornecer evidência “decisiva” para o modelo de gráfico de dependência, os dados reais biológicos, fornecem fatores Bayes de **10.064** até **515.450**.

Nós sabemos há muito tempo que a descendência comum fracassou[1]. No novo artigo de Ewert, agora temos resultados quantitativos detalhados que demonstram isso. E a Ewert fornece um novo modelo, com um ajuste muito superior aos dados.

Nota do tradutor [1]: Não foi falta de aviso.

...

Cornelius Hunter. New Paper by Winston Ewert Demonstrates Superiority of Design Model. 20 de julho de 2018.
(Acessar)