INTRODUÇÃO À EVOLUÇÃO MOLECULAR: O MODELO DE JUKES-CANTOR

Aluno: Adriana Cruz Martins
Orientador: Sérgio Bernardo Volchan

1. Introdução

A biologia molecular tem tido enorme impacto em diversos ramos da biologia e medicina. Em particular, abriu novos horizontes para o estudo da evolução e da filogenia. De fato, um dos princípios fundamentais da *teoria da evolução de Darwin* é o de que todos os organismos atuais descendem de um ancestral comum. Portanto, a descoberta da existência de grande variedade (polimorfismos) à nível molecular (proteínas e ácidos nucléicos) entre indivíduos, tanto da mesma espécie quanto de espécies diferentes, levantou a possibilidade de se estudar o parentesco evolucionário dos organismos através da comparação de "caracteres moleculares", estendendo dessa forma as técnicas de estudo tradicionalmente usadas em paleontologia (morfologia, fisiologia, comportamento, etc).

Do ponto de vista molecular, as variações entre indivíduos estão ligadas a alterações na estrutura do DNA, tipicamente (mas não somente) por mutações, tais como substituições de nucleotídeos em sítios individuais da molécula. Surge então naturalmente a hipótese de que quanto maior o acúmulo de diferenças entre trechos homólogos do DNA de indivíduos de duas espécies, menor seria o "parentesco" entre tais espécies. Isto significa que a partir da observação destas seqüências em organismos atuais seria possível estimar o momento em que houve a separação do ancestral comum. Esta idéia tem revolucionado a área da sistemática/taxonomia, revelando relações insuspeitadas entre organismos e permitindo a construção de árvores filogenéticas ("árvores da vida") mais confiáveis, precisas e completas [1].

Porém, este método de reconstrução não é algo simples de ser realizado e constitui um dos maiores desafios do estudo da chamada filogenética molecular. Há vários fatores que complicam a análise, particularmente o fato de a taxa de mutação não ser geralmente constante; e ainda que em alguns caso ela seja, deve-se considerar a ocorrência de mutações "silenciosas" e "repetidas", que não são diretamente observáveis. Diante desta dificuldade, foram desenvolvidas diversas ferramentas estatísticas assim como modelos probabilísticos que permitem determinar, sob certas condições e com certa margem de erro, a distância evolucionária. Esta é uma vasta área de pesquisa atual, extremamente ativa e multidisciplinar, envolvendo biologia, genética, bioinformática, estatística, teoria da probabilidade, otimização, etc.

Neste projeto, estudamos um dos modelos mais simples de evolução molecular: o *modelo de Jukes-Cantor*. Apesar de conter certas hipóteses um tanto quanto irrealistas (tais como a independência entre os sítios do DNA e que as substituições de nucleotídeos ocorrem com mesma probabilidade) este modelo admite uma análise matemática relativamente direta e ainda é muito utilizado como uma primeira aproximação [6]. Os objetivos principais deste projeto foram o de entender os princípios básicos da evolução molecular e a aplicação de algumas técnicas matemáticas em seu estudo, particularmente noções de teoria da probabilidade. Nossa primeira tarefa foi a de nos familiarizar com os principais conceitos biológicos relacionados à evolução molecular. Em seguida estudamos conceitos de teoria da probabilidade necessários para compreender o modelo de Jukes-Cantor, no qual nos concentramos. Por outro lado, não foi possível abordar o estudo de construção de árvores filogenéticas propriamente dito, pois exigiria a abordagem de técnicas estatísticas sofisticadas que estão além do escopo do projeto.

2. A Teoria neutra da evolução molecular e o modelo de Wright-Fisher

A formulação inicial da teoria da seleção natural de Charles Darwin foi baseada exclusivamente em observações feitas no nível fenotípico, isto é, de características macroscópicas (morfológicas, fisiológicas e comportamentais) dos organismos. Desconhecia-se a origem das variações assim como os mecanismos da hereditariedade. Enquanto não havia um conhecimento efetivo referente à existência e a natureza dos genes e ao seu papel na evolução, acreditava-se que existiam essencialmente dois tipos de modificações responsáveis pela evolução das espécies – modificações vantajosas ou prejudiciais – e somente um tipo de mecanismo responsável pela determinação do destino destas modificações (fixação ou não) – a seleção natural.

Posteriormente, com os avanços da genética e da biologia molecular descobriu-se que as variações surgem devido à certas alterações estruturais no material genético (o DNA), tipicamente as *mutações*. Com a descoberta, nos anos 1960, de uma insuspeitada variação entre indivíduos ao nível molecular, surgiu a hipótese de que muitas destas variações não sofreriam ação da seleção natural, isto é seriam *neutras*.

O conceito de mutação neutra se aplica a todas as mutações que não são necessariamente responsáveis pelo aparecimento de características adaptativas (não possuem impacto significativo na habilidade dos organismos sobreviverem ou se reproduzirem) e que, portanto, não possuem sua fixação (numa população) determinada pela seleção natural. Um exemplo deste tipo de mutação são as alterações "silenciosas", que acarretam a substituição de certos aminoácidos de uma proteína (estrutura primária), mas não afetam a conformação (estruturas secundária e terciária) e, portanto, a função da proteína correspondente. Em 1968, o geneticista japonês Motoo Kimura propôs que, à nível molecular, mutações neutras seriam mais freqüentes que os demais tipos de mutação e que sua fixação ocorreria por efeitos puramente estatísticos ou aleatórios, a chamada "deriva gênica". Com isso, introduziu-se outro mecanismo de evolução: a fixação de mutações neutras por "deriva gênica".

Esta nova idéia evolucionária sugeria que as mutações responsáveis pelo surgimento de características adaptativas vantajosas possuiriam pouca contribuição para a variabilidade genética das populações por serem extremamente raras e se fixarem muito rapidamente (pela seleção natural). Além disso, Kimura excluiu, em sua teoria neutra, as mutações prejudiciais de suas considerações já que estas não contribuiriam nem para a variabilidade genética nem para a evolução molecular, uma vez que são rapidamente eliminadas por meio da chamada "seleção negativa".

É importante ressaltar que esta chamada *teoria neutra da evolução*, ainda que tenha causado muita controvérsia, não nega a existência da seleção natural nem sua importância para a evolução. No entanto, ao contrário de Darwin, que não dispunha dos conhecimentos de biologia molecular, a teoria neutra lida essencialmente com variações à nível molecular.

Com o auxílio de modelos matemáticos apropriados e da noção de "relógio molecular (que discutiremos posteriormente), a teoria neutra de Kimura constitui uma das principais ferramentas para a compreensão da evolução molecular. Um modelo probabilístico relativamente simples de deriva neutra é discutido em seguida.

2.1. O modelo de Wright-Fisher

O *modelo de deriva gênica* foi introduzido pela primeira vez na década de 1920 por Sewall Wright e Ronald Fisher no contexto da genética de populações e em tempo discreto. O modelo em tempo contínuo foi retomado por Kimura na década de 1960, utilizando técnicas da teoria de processos estocásticos (difusões), no contexto da teoria neutra da evolução molecular. O modelo de Wright-Fisher ilustra o processo evolucionário de mudança na freqüência dos alelos em uma população que ocorre de forma inteiramente aleatória devido aos efeitos de amostragem em uma população finita.

Na versão mais simples, o modelo descreve a evolução de um lócus com apenas dois alelos (e.g. A e B) em uma população com número fixo N de indivíduos haplóides em gerações não-superpostas n (n=0, 1, 2, ...), sujeita a cruzamento aleatório na ausência de qualquer tipo de mutação. A seguir, descrevemos a obtenção de alguns resultados deste modelo e suas propriedades.

Primeiramente, consideramos X_n como a variável aleatória que representa o número de alelos do tipo A na geração n. A população na geração n+1 é gerada a partir da geração n pela amostragem binomial (ver Apêndice 1) de N alelos de um conjunto de genes ("gene pool") no qual a fração inicial de alelos A é suposta ser $\pi_i = \frac{i}{N}$ ($\pi_i = \frac{i}{2N}$, no caso diplóide).

Logo, dado que $X_n = i$ e considerando a amostragem binomial dos alelos, a probabilidade (condicional) de que $X_{n+1} = j$ é dada por:

$$p_{ij} \equiv \mathrm{P}(X_{n+1} = j \mid X_n = i) = \binom{N}{j} \pi_i^j (1-\pi_i)^{N-j},\ 0 \le i, j \le N.$$

A seqüência de variáveis aleatórias $\{X_n\}_{n\ge 0}$ é um exemplo de um *processo estocástico* em tempo discreto chamado *Cadeia de Markov Homogênea* com matriz de transição p_{ij} e espaço de estados S= {0,1,...,N}. Um processo Markoviano satisfaz a identidade:

$$\mathrm{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \mathrm{P}(X_{n+1} = i_{n+1} \mid X_n = i_n),$$

ou seja, a probabilidade condicional de que o sistema esteja em um dado estado após n passos, dados todos os passos anteriores, é a mesma que a probabilidade condicional conhecendo-se apenas o estado no passo imediatamente anterior (propriedade de "memória curta"). Já a homogeneidade temporal do processo significa que:

$$\mathrm{P}(X_{n+1} = j \mid X_n = i) = \mathrm{P}(X_1 = j \mid X_0 = i).$$

Listamos a seguir algumas propriedades básicas do modelo.

1) $\mathrm{E}[X_n \mid X_{n-1}] = X_{n-1}$, pois a partir da esperança da distribuição binomial, temos que

$$\mathrm{E}[X_n \mid X_{n-1} = i] = N\pi_i = N\left(\frac{i}{N}\right) = i \ \Leftrightarrow\ i = \sum_{j=0}^{N} j p_{ij}\ .$$

A partir disto, podemos concluir também que $\mathrm{E}[X_n \mid X_{n-1}] = X_{n-1} = ... = \mathrm{E}[X_0]$, isto é, o "processo é constante em média".

2) Os estados i=*0* e i=N são "absorventes": uma vez atingidos, não se alteram pois representam, respectivamente, a ausência do alelo A na população e a presença exclusiva do alelo A na população. Isto é, neste caso, $p_{ij} = 0$, para $(\forall j \neq i)$, i= 0, ..., N, e $p_{00} = p_{NN} = 1$. Note que como o número de estados é finito, eventualmente um dos estados absorventes é necessariamente atingido, em tempo finito, i.e., ocorre fixação de um dos alelos (ver figura 1).

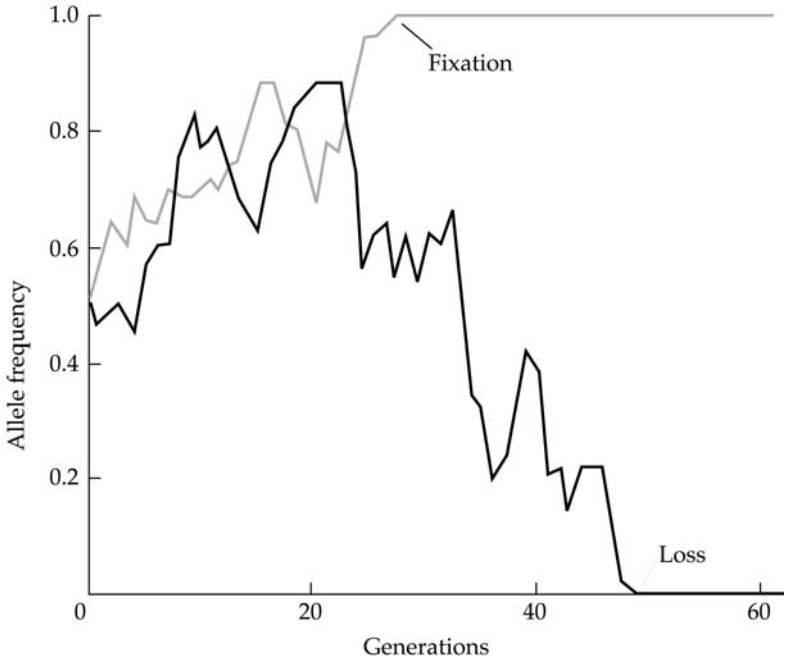


Figura 1: O gráfico ilustra a variação na freqüência de dois alelos ao longo das gerações. Verificamos que, em tempo finito, eventualmente ocorre a fixação de um dos alelos e o desaparecimento do outro alelo na população.

3) Seja a_i a probabilidade de fixação do alelo A, dado que $\mathrm{X}_0 = i$, então: $a_i = \pi_i = \dfrac{i}{N}$.

Portanto, se num instante inicial, surge um novo alelo (e que pode ser interpretado como o surgimento de uma mutação naquele instante) ele se fixa com probabilidade 1/*N*. Esta propriedade é extremamente importante e vale a pena verificá-la com mais detalhe.

Demonstração:

Seja *A* o evento de fixação e $a_i \equiv \mathrm{P}(A \mid \mathrm{X}_0 = i)$. Vamos condicionar na variável X_1:

Como $\sum_{j=0}^{N} \{\mathrm{X}_1 = j\} = \Omega$, onde "$\sum_{j=0}^{N}$" representa a união disjunta dos eventos $\{\mathrm{X}_1 = j\}$, $j = 0,..., N$ e Ω é o espaço amostral, temos:

$$\mathrm{A} = \mathrm{A} \cap \Omega = \sum_{j=0}^{N} \mathrm{A} \cap \{\mathrm{X}_1 = j\} \Rightarrow a_i = \mathrm{P}(\mathrm{A} \mid \mathrm{X}_0 = i) \equiv \frac{\mathrm{P}(\mathrm{A} \cap \{\mathrm{X}_0 = i\})}{\mathrm{P}(\mathrm{X}_0 = i)}$$

$$= \sum_{j=0}^{N} \mathrm{P}(\mathrm{A}, \mathrm{X}_1 = j \mid \mathrm{X}_0 = i)$$

Mas temos:

$$a_i = \sum_{j=0}^{N} \mathrm{P}(A, \mathrm{X}_1 = j \mid \mathrm{X}_0 = i) = \sum_{j=0}^{N} \mathrm{P}(A \mid \mathrm{X}_1 = j, \mathrm{X}_0 = i) \cdot \mathrm{P}(\mathrm{X}_1 = j \mid \mathrm{X}_0 = i) *$$

$$= \sum_{j=0}^{N} \mathrm{P}(A \mid \mathrm{X}_1 = j) \cdot p_{ij} \qquad \text{(Considerando a propriedade de Markov)}$$

$$= \sum_{j=0}^{N} \mathrm{P}(A \mid \mathrm{X}_0 = j) \cdot p_{ij} \qquad \text{(Considerando a homogeneidade)}$$

* Aqui, usamos que para quaisquer eventos A, B e C:

$$\mathrm{P}(A \cap B \mid C) = \mathrm{P}(A \mid B \cap C) \cdot \mathrm{P}(B \mid C)$$

$= \sum_{j=0}^{N} a_j \cdot p_{ij}$

Isto é, $a_i = \sum_{j=0}^{N} a_j \cdot p_{ij}$ e $a_0 = 0, a_N = 1$.

Lembrando que $i = \sum_{j=0}^{N} jp_{ij}$, temos que o vetor $\vec{v} = (0, 1, ..., N)^T$ é solução da equação:

$\rho \cdot \vec{v} = \vec{v} \Rightarrow c \cdot \rho \cdot \vec{v} = \rho(c \cdot \vec{v}) = c \cdot \vec{v}$,

para qualquer constante real *c*, onde $\rho = [p_{ij}]$ é a matriz de transição. Logo $\vec{a} = (a_0, a_1, ..., a_N)^T = c(0, 1, ..., N)^T$, isto é, $a_i = c \cdot i$. Logo,

$a_N = 1 \Rightarrow c = \frac{1}{N} \quad \Rightarrow \quad a_i = \frac{i}{N} = \pi_i$, C.Q.D.

2.2. O relógio molecular

Em uma população de *N* indivíduos diplóides. Com o auxílio do modelo de deriva gênica de Wright-Fisher, pode-se obter a *taxa de fixação*, *k,* de um novo alelo nesta população:

$$k = (\text{nº médio de mutações por geração}).(\text{fração de mutações que se fixam})$$

O número médio de mutações por geração é determinado pelo produto entre número de gametas produzido por geração, *2N*, e a taxa de ocorrência de uma mutação por geração, *u*. Considerando que a fração de mutações que se fixam, de acordo com o modelo de Wright-Fisher, é 1/*2N*, obtemos:

$$k = (2 \cdot N \cdot u) \cdot \left(\frac{1}{2 \cdot N}\right) = u,$$

isto é, $k = u$.

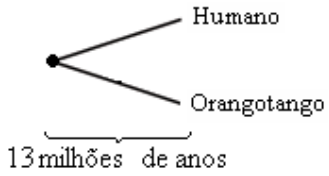
Este é um dos resultados mais importantes da teoria neutra da evolução. De acordo com este modelo a taxa na qual as mutações neutras ocorrem é igual à taxa de fixação de um novo alelo. Ora, um alelo pode ser pensado como um trecho de DNA, digamos de *m* nucleotídeos. Supondo que a taxa de fixação e a taxa de mutação por sítio sejam, respectivamente, u_s e k_s, idênticas para todos os sítios, e que os sítios são independentes, segue do resultado acima que:

$$u = m \cdot u_s = m \cdot k_s \Rightarrow u_s = k_s.$$

Ou seja, sob a teoria neutra e no contexto do modelo de Wright-Fisher, a taxa de mutação de nucleotídeos por sítio é igual à sua taxa de fixação por sítio. Esta é a base da noção do "relógio molecular".

O termo "relógio molecular" foi introduzido em 1965, por Emile Zuckerkandl e Linus C. Pauling, para ilustrar esta acumulação de substituições de monômeros de macromoléculas de importância biológica (o caso originalmente por eles estudado era o de substituições de aminoácidos em proteínas), *supostamente a uma taxa constante*, o que permitiria estimar o "parentesco" entre organismos pela comparação de diferenças observáveis de seqüências homólogas (ver figura 2). Neste sentido, pode-se dizer que as moléculas são capazes de determinar seu "tempo evolucionário" através do acúmulo de substituições (divergência), funcionando assim como um verdadeiro "documento histórico" da evolução. É importante

observarmos, entretanto, que a hipótese da ocorrência de substituições a uma taxa constante é uma aproximação, e que na verdade esta taxa pode flutuar muito de gene para gene, de espécie para espécie, etc; o que complica substancialmente a análise.



Número de substituições = x

Número por linhagem = $\frac{x}{2}$

Número por linhagem

por milhão de anos = $\frac{x}{2 \times 13}$

Figura 2: Calculando um relógio molecular humano
O número observado de diferenças é determinado para um par de genes homólogos de humano e orangotango, aqui, este número é chamado de 'x'. O número de substituições por linhagem é x/2 e o número por milhões de anos é x/26. Neste caso, a partir do tempo de divergência entre os dois organismos (tempo de separação de um ancestral comum) foi possível determinar o número de diferenças acumuladas x. Poderíamos também realizar o cálculo inverso, obtendo o tempo de separação a partir da observação de x. Em todo caso, é preciso um modelo matemático que faça a correção entre as diferenças *observadas* e as substituições que *realmente* ocorreram desde a separação entre as espécies. O modelo de Jukes-Cantor faz exatamente isso.

3. O modelo de Jukes-Cantor

3.1. A distribuição de probabilidade de Poisson

Os principais eventos responsáveis pela divergência entre seqüências do DNA são as *mutações* e *fixações.* De acordo com a teoria neutra da evolução molecular, a taxa de fixação das mutações é igual à taxa com a qual as mutações neutras surgem, portanto, podemos analisar estes dois eventos de forma conjunta. No caso mais simples, mutações correspondem à troca de um nucleotídeo por outro (ou substituição) em um sítio específico de uma molécula de DNA. Apesar de seu caráter aleatório individual, mutações têm efeitos previsíveis, no sentido de que podem ser estimados através de médias estatísticas obtidas a partir da aplicação de modelos probabilísticos adequados.

No modelo de Jukes-Cantor supõe-se que o acúmulo de substituições de nucleotídeos durante a evolução molecular é um processo que pode ser descrito pela *distribuição de probabilidade de Poisson* de parâmetro λ cuja fórmula é expressa por $\mathrm{P}(k,\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$, $k = 0,1,2,..., \lambda > 0$ que, por sua vez, pode ser obtida a partir da distribuição binomial de probabilidade. Existem várias formas de obter este resultado e descrevemos uma delas a seguir [3].

A distribuição binomial $\mathrm{B}(k;n,p) = \binom{n}{k} p^k q^{n-k}$ pode ser escrita como:

$$\frac{n(n-1)(n-2)\cdots(n-i+1)}{i!} p^i q^{n-i}$$

Podemos multiplicar e dividir por n^i e obter:

$$\frac{1(1-\frac{1}{n})(1-\frac{2}{n})\cdots(1-\frac{i-1}{n})}{i!}(np)^i(1-\frac{np}{n})^{n-i}.$$

Fazendo com que $n \to \infty$ de tal forma que $np = \lambda$ permaneça constante, cada termo do produto $(1-\frac{1}{n})...[1-\frac{(i-1)}{n}]$ tenderá a 1, e $(np)^i$ se reduzirá a λ^i. Também:

$$\left(1-\frac{np}{n}\right)^{n-i} = \left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-i} \to e^{-\lambda}(1) = e^{-\lambda}.$$

Portanto, no limite como $n \to \infty$ com $np = \lambda$ (e como $p = \frac{\lambda}{n} \to 0$), temos:

$$\binom{n}{i}p^i q^{n-i} \to \frac{\lambda^i e^{-\lambda}}{i!} \text{ e } \mathrm{P}(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}.$$

Este resultado fornece a distribuição de Poisson com parâmetro λ como limite da distribuição binomial, para o caso de *n* grande e *p* pequeno, que é o caso de mutações de nucleotídeos. Observamos que a média da distribuição de Poisson é exatamente λ (ver Apêndice 2).

Aplicada à evolução molecular, a distribuição de Poisson fornece a probabilidade de que 0,1,2,3,... substituições ocorram em um segmento de DNA de um determinado tamanho em um intervalo de tempo definido. O número médio esperado de substituições observadas em um intervalo fixo de tempo é dado por $2t\mu$, onde μ é a taxa de substituição (número médio de substituições por sítio de seqüência, por unidade de tempo) e *t* é o tempo decorrido desde o momento da divergência entre as duas seqüências de DNA comparadas. Como cada uma das duas seqüências acumulou substituições independentemente durante um intervalo de tempo *t*, juntas elas tiveram um tempo correspondente a $t + t = 2t$ para divergir (ver figura 2). Portanto, a fórmula de Poisson para a evolução molecular pode ser expressa por:

$$\mathrm{P}(k;2t\mu) = \frac{(2t\mu)^k}{k!}e^{-2t\mu}.$$

Aqui, $\mathrm{P}(k;2t\mu)$ é a probabilidade de que um número k = 0,1,2,3,... de substituições ocorra em um sítio do DNA em um intervalo de tempo *t* quando a taxa de substituição é μ.

3.2. A fórmula de Jukes-Cantor

A aplicação da fórmula de distribuição de probabilidade de Poisson no estudo da evolução molecular é, no entanto, limitada, pois, freqüentemente, não sabemos nem a taxa de substituição μ nem o tempo *t* de divergência entre as duas seqüências. Para solucionar este problema, foram desenvolvidos métodos de obtenção do número médio $2t\mu$ de substituições independentemente das variáveis *t* e μ.

Teoricamente, poderíamos obter o número médio de substituições através da observação do número de posições que diferem entre as duas seqüências homólogas alinhadas. No entanto, esta proporção de diferenças, x/L (número médio de substituições diferentes observadas, x, dividido pelo número total de nucleotídeos no trecho comparado, *L*) não leva em consideração os eventos que não são observados como, por exemplo, as "substituições múltiplas" e "recorrentes". Portanto, para obter o número médio de substituições, precisamos considerar os eventos "escondidos" e, assim, converter/corrigir a proporção de diferenças observada em uma distância evolucionária efetiva (número total de substituições realmente ocorridas por sítio desde a separação das espécies).

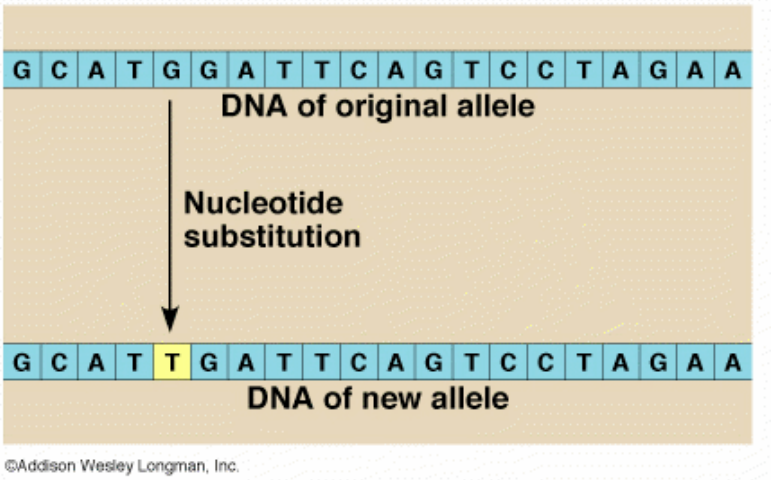


Figura 3: O diagrama ilustra a ocorrência de uma substituição de nucleotídeo em um sítio da molécula de DNA.

Para realizar esta conversão, precisaríamos considerar todas as mudanças que um nucleotídeo específico e os nucleotídeos de um determinado sítio podem sofrer. Em seguida, deveríamos calcular a probabilidade de mudanças individuais, assumindo o processo de substituição como sendo um processo de Poisson, e estimar o número de mudanças que não são reveladas na comparação das duas seqüências. Este procedimento aparentemente complicado pode ser condensado por uma fórmula matemática, a *fórmula de Jukes-Cantor*.

O primeiro e mais simples modelo desenvolvido com o objetivo de obter esta distância evolucionária entre seqüências de DNA foi descrito em 1969 por Thomas H. Jukes e Charles R. Cantor. Este modelo é baseado na suposição de que as *transições* (troca entre bases de mesmo tipo: purinas ou pirimidinas) ocorrem com a mesma probabilidade que as demais substituições -*transversões*- (ver figura 4) e a obtenção de sua fórmula geral é descrita a seguir [4].

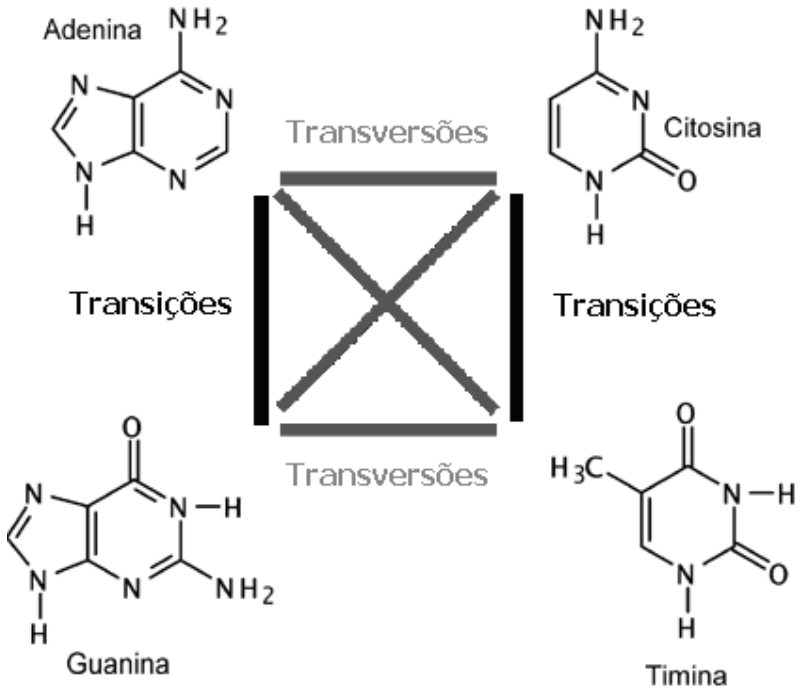


Figura 4: Existem dois tipos de mutações de substituições do DNA. Transições são trocas que ocorrem entre purinas (A ↔ G) ou entre pirimidinas (C ↔ T). Já as transversões correspondem às trocas entre purinas e pirimidinas. Apesar de haver o dobro de transversões possíveis, transições são mais freqüentes que transições.

Primeiramente, consideremos um sítio de nucleotídeo específico com uma determinada probabilidade de sofrer k=0,1,2,3,... substituições durante um *intervalo de tempo fixo*. De acordo com a distribuição de probabilidade de Poisson, a probabilidade de ocorrerem k substituições é $\mathrm{P}(k) = \frac{\lambda^k}{k!} e^{-\lambda}$. Como a média da distribuição de Poisson é λ, segue que λ é o número médio de substituições que levou às diferenças observadas ("fixadas") entre duas seqüencias homólogas no intervalo de tempo dado. Assumindo que, no início deste intervalo de tempo, o sítio estivesse sendo ocupado por um determinado nucleotídeo, por exemplo, A, podemos designar por $I(k)$ a probabilidade de que, após k substituições, no final do intervalo, o sítio seja ocupado novamente por um nucleotídeo A. De forma similar, podemos designar por $D(k)$ a probabilidade de que, após k substituições, o sítio seja ocupado por um nucleotídeo diferente: G, C ou T. Desta maneira, concluímos que $I(k) + D(k) = 1$ e, portanto, $D(k) = 1 - I(k)$.

Agora, podemos analisar o que ocorre quando a próxima (k+1) substituição ocorre. As probabilidades correspondentes seriam então $I(k+1)$ e $D(k+1)$. Se, após k substituições o sítio estivesse sendo ocupado por um A, então, após k+1 substituições, o nucleotídeo neste sítio não pode ser um A. Se, após k substituições o sítio estivesse sendo ocupado por um C, então, após uma substituição adicional, a probabilidade de substituição por um A é $1/3$, e o mesmo é verdade para o sítio ocupado por um G ou um T, após k substituições.

A partir disto, concluímos que, se após k substituições, independentemente de o sítio estar sendo ocupado por um C, T ou G, a probabilidade de que ele volte a ser ocupado por um A é $1/3$. Como a probabilidade de um nucleotídeo A ser substituído por C, T ou G, após k substituições é $D(k)$ e como, se ocorreu esta substituição, existe uma probabilidade de 1:3 de que o sítio volte a ser ocupado por um nucleotídeo A após uma nova substituição, então:

$$I(k+1) = \frac{1}{3} D(k).$$

Se, agora, substituirmos $D(k)$ por $1 - I(k)$, obtemos:

$$I(k+1) = \frac{1}{3}\left[1 - I(k)\right].$$

Podemos notar que, se, originalmente, o sítio estava sendo ocupado por um A e se nenhuma substituição ocorreu (k=0), o sítio permanece com o nucleotídeo A. Portanto, definimos $I(0) = 1$. Para $I(1)$, obtemos:

$$I(0+1) = \frac{1}{3}\left[1 - I(0)\right] = \frac{1}{3}\left[1 - 1\right] = 0.$$

Para obter $I(2)$, escrevemos $I(1+1) = 1/3(1-0)$ ou $I(2) = 1/3$. Repetindo este processo, podemos obter $I(k)$ e, conseqüentemente, $D(k)$ para todos os inteiros não-negativos. Quando k torna-se muito grande, a diferença entre $I(k)$ e $I(k+1)$ fica desprezível e, sob estas circunstancias, podemos substituir ambas as expressões por um símbolo comum b e reescrever a equação $I(k+1) = 1/3\left[1 - I(k)\right]$ como $b = 1/3\left(1 - b\right)$, ou seja, $b = 1/3 - 1/3b$, isto é, $1/3b + b = 1/3$ e $(4b)/3 = 1/3$. Após as simplificações adequadas, obtemos finalmente: $b = 1/4$.

Escrevemos $I'(k) = I(k) - b$ (tal que $I(k) = I'(k) + b$) e $I'(k+1) = I(k+1) - b$. Subtraindo b de ambos os lados da equação $I(k+1) = 1/3\left[1 - I(k)\right]$, podemos escrever:

$$\begin{aligned}
I(k+1) - b &= \frac{1}{3}\left[1 - I(k)\right] - b = \frac{1}{3} - \frac{1}{3} I(k) - b \\
&= \frac{1}{3} - \frac{1}{3}\left[I'(k) + b\right] - b \qquad \text{(aqui, substituímos } I(k) \text{ por } I'(k) + b\text{)} \\
&= \frac{1}{3} - \frac{1}{3} I'(k) - \frac{1}{3} b - b = \frac{1}{3} - \frac{1}{3} I'(k) - \frac{4}{3} b \\
&= \frac{1}{3} - \frac{1}{3} I'(k) - \left(\frac{4}{3}\right)\left(\frac{1}{4}\right) \quad \text{(pois } b = 1/4\text{)} \\
&= \frac{1}{3} - \frac{1}{3} I'(k) - \frac{1}{3} = -\frac{1}{3} I'(k).
\end{aligned}$$

E, como $I'(k+1) = I(k+1) - b$, temos que:

$$I'(k+1) = -\frac{1}{3} I'(k).$$

Podemos, então, escrever:

$$I'(0) = I(0) - b = 1 - \frac{1}{4} = \frac{3}{4},$$

$$I'(1) = I'(0)\left(-\frac{1}{3}\right) = \left(\frac{3}{4}\right)\left(-\frac{1}{3}\right),$$

$$I'(2) = I'(1)\left(-\frac{1}{3}\right) = \left(-\frac{1}{3}\right)\left[\left(-\frac{1}{3}\right) I'(0)\right] = \left(-\frac{1}{3}\right)^2 I'(0).$$

Logo, $I'(k) = \left(-\frac{1}{3}\right)^k I'(0)$.

Adicionamos b a ambos os lados da última equação e escrevemos:

$$I'(k)+b=\left(-\frac{1}{3}\right)^k I'(0)+b\,.$$

Como $I'(k)+b=I(k)$, obtemos:

$$I(k)=\left(-\frac{1}{3}\right)^k I'(0)+b$$

E, como $b=1/4$ e $I'(0)=3/4$, obtemos:

$$I(k)=\frac{1}{4}+\frac{3}{4}\left(-\frac{1}{3}\right)^k.$$

Finalmente, uma vez que $D(k)=1-I(k)$, podemos escrever:

$$D(k)=1-\left[\frac{1}{4}+\frac{3}{4}\left(-\frac{1}{3}\right)^k\right]=\frac{3}{4}-\frac{3}{4}\left(-\frac{1}{3}\right)^k=\frac{3}{4}\left[1-\left(-\frac{1}{3}\right)^k\right].$$

Até este momento, consideramos substituições individuais uma por uma e obtivemos a probabilidade de diferenças em sítios individuais. Agora, ao invés de analisar valores individuais da variável aleatória 0, 1, 2,... e especificar a probabilidade de cada uma individualmente, devemos considerar a seqüência inteira e as diferenças de todos os sítios juntos. Consideramos que *k* pode assumir qualquer valor inteiro não-negativo com uma certa probabilidade. Para isso, precisamos analisar o somatório do produto entre a probabilidade de diferenças para os valores individuais da variável e a proporção de diferenças observadas entre as seqüências. Chamando o somatório de $\overline{D}$, podemos escrever:

Probabilidade de diferenças para valores individuais — Proporção de diferenças após *k* substituições

$$\overline{D}=\sum_{k=0}^{n}\overbrace{\frac{3}{4}\left[1-\left(\frac{1}{3}\right)^k\right]}\overbrace{\mathrm{P}(k)}$$

$$=\frac{3}{4}\left[1-\sum_{k=0}^{n}\left(-\frac{1}{3}\right)^k\mathrm{P}(k)\right]$$ (movemos o somatório para dentro dos parênteses)

$$=\frac{3}{4}\left[1-\sum_{k=0}^{n}\frac{\left\{\left(-\frac{1}{3}\right)\lambda\right\}^k}{k!}\,e^{-\lambda}\right]$$ (substituímos P(*k*) pela fórmula geral da distribuição de Poisson)

$$=\frac{3}{4}\left[1-e^{-\frac{\lambda}{3}}e^{-\lambda}\right]\,.$$ (pela definição de e^x onde $x=-\lambda/3$).

Logo:

$$\overline{D}=\frac{3}{4}\left[1-e^{-\frac{4}{3}\lambda}\right].$$

Desenvolvendo este resultado, podemos obter:

$$\frac{4}{3}\overline{D} = 1 - e^{-\frac{4}{3}\lambda} \Rightarrow e^{-\frac{4}{3}\lambda} = 1 - \frac{4}{3}\overline{D}.$$

Aplicando o logaritmo natural em ambos os lados, obtemos:

$$-\frac{4}{3}\lambda = \ln\left(1 - \frac{4}{3}\overline{D}\right) \Rightarrow \boxed{\lambda = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\overline{D}\right)}.$$

Esta é a *fórmula de Jukes-Cantor* para estimar λ, o número médio de substituições por sítio. Com isso, podemos achar $2t\mu$ e isto nos permite inferir μ e/ou t, já que $\overline{D} \approx \frac{x}{L}$; número médio de diferenças observado por sítio (ver figura 3) obtido da observação das diferenças entre as seqüências.

Diante do crescimento do estudo da evolução molecular, já foram desenvolvidos outros modelos probabilísticos mais complexos que levam em consideração, por exemplo, a variação na composição de nucleotídeos, a diferença na probabilidade de ocorrência de transversões e transições (sabe-se que transições são mais freqüentes que transversões), assim como outros fatores que podem influenciar a freqüência e a natureza das substituições de nucleotídeos. Dessa forma, tais modelos são capazes de fornecer uma correção mais precisa para as substituições não observadas.

4. Apêndice 1

Ao lançarmos uma moeda, por exemplo, temos dois resultados possíveis, caras *K* e coroas *C*, e estes são os elementos do espaço amostral Ω. Quando a moeda é lançada duas vezes, o espaço amostral apropriado Ω contém 4 elementos, *KK*, *KC*, *CK*, *CC*. Neste caso, podemos definir uma variável aleatória X como sendo o número de caras. Considerando uma moeda honesta, cada um dos eventos, *KK*, *KC*, *CK*, *CC*, ocorre com uma mesma probabilidade (=1/4). De uma maneira mais geral, quando "caras" são obtidas com uma probabilidade *q* e "coroas" são obtidas com probabilidade *p* (*p*+*q*=1), e se os resultados dos lançamentos são independentes, temos:

$$\begin{aligned} \mathrm{P}(\mathrm{X} = 0) &= (p)^2 = (1-q)^2 \\ \mathrm{P}(\mathrm{X} = 1) &= 2q(p) = 2q(1-q) \\ \mathrm{P}(\mathrm{X} = 2) &= (q)^2 \end{aligned}$$

De uma forma geral, se lançarmos a moeda *n* vezes, então $\binom{n}{i}$ pontos do espaço amostral Ω correspondem à exatamente *i* caras (logo, *n-i* coroas) e a função de distribuição de probabilidade neste caso é portanto:

$$\mathrm{B}(n,i,p) = \binom{n}{i} p^i q^{n-i}.$$

Esta fórmula, chamada *distribuição binomial*, fornece a probabilidade de *i* "sucessos" em *n* tentativas independentes de um "experimento" que tem probabilidade *p* de "sucesso" (e $q = 1 - p$ de "fracasso") em cada tentativa. Aqui, $\binom{n}{i}$ é o coeficiente binomial que pode ser reescrito na forma:

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}.$$

Podemos aplicar a distribuição binomial de probabilidade para qualquer experiência que tenha dois resultados possíveis, "sucesso" e "fracasso" (ou "caras" e "coroas", "alcançou" e "falhou", etc.) considerando uma seqüência independente de eventos em que cada resultado tem a mesma probabilidade de ocorrência (esta seqüência é chamada uma *seqüência de provas de Bernoulli*).

Utilizada como hipótese para a amostragem de genes na construção de uma nova geração, por exemplo, a distribuição binomial de probabilidade fornece a probabilidade de que um gene específico (eg. A) seja "escolhido" para formar o conjunto de genes da geração seguinte, n+1, a partir de um conjunto de genes com dois alelos (eg. A e B), na geração n (ver figura 5). Considerando a amostragem binomial neste caso, estamos supondo que os dois alelos possuem a mesma probabilidade de serem "escolhidos" para formar a geração seguinte e que estes eventos são independentes. Podemos observar que os eventos têm dois resultados possíveis: "escolher o alelo" ou "não escolher o alelo".

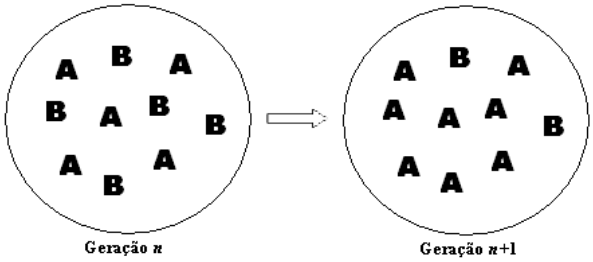


Figura 5: Formação de um novo conjunto de genes ("gene pool") através da amostragem de alelos de uma geração para a outra.

5. Apêndice 2

No estudo da probabilidade, a esperança, valor médio ou expectância de uma variável aleatória mede, grosso modo, como seus valores estão localizados. Mais geralmente, para uma variável aleatória X que admita somente valores discretos, com pesos $p(x)$, o seu k-ésimo momento é definido por:

$$\mathrm{E}(\mathrm{X}^k) = \sum_x x^k \, p(x),$$

sendo a esperança correspondendo ao caso k=1. No caso uma variável aleatória com distribuição de Poisson de parâmetro λ, temos:

$$\mathrm{E}[\mathrm{X}] = \sum_{k=0}^{\infty} k\mathrm{P}(\mathrm{X}=k) = \sum_{k=0}^{\infty} k \cdot e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

Como o termo $e^{-\lambda}$ não depende de k, podemos retirá-lo do somatório e obter:

$$\mathrm{E}[\mathrm{X}] = e^{-\lambda} \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!}.$$

Como o termo k=0 da distribuição de Poisson é igual a zero, temos:

$$\mathrm{E}[\mathrm{X}] = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1} \cdot \lambda}{(k-1)!}.$$

Se chamarmos $k-1=m$, temos:

$$\mathrm{E}[\mathrm{X}] = e^{-\lambda} \cdot \lambda \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{-\lambda} \cdot \lambda \cdot e^{\lambda} = \lambda$$

Concluímos, portanto, que a média da distribuição de probabilidade de Poisson de parâmetro λ é exatamente λ.

6. Bibliografia

1. LECOINTRE, G. & LE GUYADER, H. **The tree of life.** Harvard University Press, Cambrigde, Massachustts (2006).

2. BROWN, T. A. **Genomes**. 2.ed. Oxford: Wiley-Liss, 2002. 572p.

3. GRIMMETT, G. R. & STIRZAKER, D. R. **Probability and Random Processes**. 2.ed. Oxford: Oxford University Press, 1992. 600p.

4. KLEIN, J. & TAKAHATA, N. **Where do we come from? The molecular evidence for Human Descent**. 1.ed. Berlin: Springer, 2001. 462p.

5. TAVARÉ, S. & ZEITOUNI, O. **Lectures on Probability Theory and Statistics**. New York: Springer Verlag, 2004. 314p.

6. PATCHER, L. & STURMFELS, B. **The mathematics of phylogenomics**. Siam Review, Vol. 49, Nº 1, 2007. pp. 3-31.