Capítulo 3

Modelos de Evolução.

À medida em que se foi entendendo como funciona a variação de moléculas como o DNA e os peptídeos, também cresceu não só o interesse, como também a necessidade de se modelar matematicamente o processo no qual uma molécula ancestral se transformou naquela que observamos na atualidade. Dois modelos bastante simples, bem como a motivação para os respectivos estudos, são mostrados em suas formas puras nas Seções 3.1 e 3.2, considerando apenas eventos de troca entre os quatro tipos básicos de nucleotídeos, apresentados na Figura 3.3. Na Seção 3.3 são apresentadas formas de aumentar a capacidade de modelos de forma a adaptá-los a novas teorias.

3.1 O Modelo de Evolução de Jukes-Cantor

O modelo de evolução mais simples existente é o modelo proposto por T. H. Jukes e C. Cantor no ano de 1969. Neste modelo, a probabilidade α de um nucleotídeo mudar para outro diferente num intervalo de tempo Δt é proporcional ao tempo transcorrido. O tamanho do intervalo Δt varia de acordo com a necessidade. Há estudos que trabalham com intervalos de poucos anos, enquanto outros pode chegar a usar intervalos da ordem de milhares de anos. Se o tamanho de Δt é arbitrário, faz sentido supor que a probabilidade de uma determinada posição de uma seqüência sofrer uma mutação é maior num intervalo de tempo maior. Neste modelo, α é diretamente proporcional ao tamanho de Δt. A relação entre α e tamanho do intervalo de tempo é dada por $\alpha = u\Delta t$, onde u é uma constante.

Seja t_0 o momento que consideramos como sendo o começo de um processo evolutivo qualquer, a probabilidade de uma posição em uma seqüência conter o mesmo tipo de nucleotídeo no intante t_0 e num instante posterior t qualquer é dada pela função $X(t)$ e a probabilidade da mesma posição conter um nucleotídeo diferente no instante t é dada pela função $Y(t)$.

É possível obter as duas funções definido-se recursivamente tanto $X(t + \Delta)$ quanto

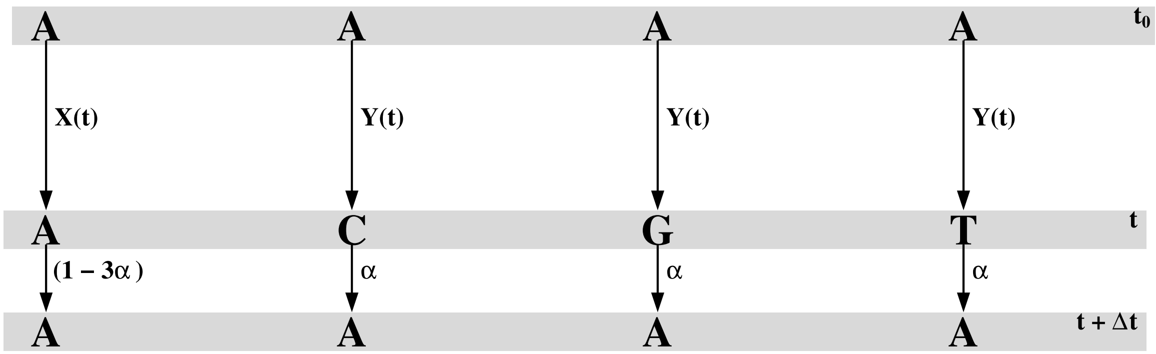

Figura 3.1: *Um exemplo ilustrando os quatro cenários possíveis para o caso em que o nucleotídeo em determinada posição na seqüência original não se diferencia do nucleotídeo na mesma posição da seqüência observada no tempo $t+\Delta t$ no modelo de Jukes e Cantor.*

$Y(t+\Delta)$ em função de $X(t)$ e $Y(t)$. As Seções seguintes se dedicam ao estudo dos dois casos separadamente, unindo-os no final para, através de um sistema de equações diferenciais, chegar às funções $X(t)$ e $Y(t)$.

3.1.1 Os Nucleotídeos Não Diferem

Embora os nucleotídeos não sejam diferentes em t_0 e $t+\Delta t$, como mostra a Figura 3.1, devemos considerar quatro cenários:

Cenário 1: O mesmo tipo de nucleotídeo pode ser observado nos três instantes: t_0, t e $t+\Delta t$. A probabilidade deste cenário é dada pela função $P_1(t+\Delta t) = (1-3\alpha)X(t)$.

Cenários 2, 3 e 4: Estes quatro cenários são semelhantes, com exceção do tipo de nucleotídeo que ocupa a posição na seqüência no instante t. Em todos eles o tipo de nucleotídeo encontrado na seqüência na posição em questão é o mesmo tanto no instante t_0 quanto no instante $t+\Delta t$, porém o tipo de nucleotídeo observado em t é diferente. Nestes casos, a probabilidade do cenário é dada por $P_2(t+\Delta t) = P_3(t+\Delta t) = P_4(t+\Delta t) = \alpha Y(t)$.

A probabilidade de um nucleotídeo numa determinada posição ser igual ao nucleotídeo original é dada pela soma das probabilidades dos cenários. Assim:

$$X(t+\Delta t) = (1-3\alpha)X(t) + 3\alpha Y(t)$$

Sabemos, porém, que α é determinado por $\alpha = u\Delta t$. Assim:

$$\begin{aligned} X(t+\Delta t) &= X(t) - 3u\Delta t X(t) + 3u\Delta t Y(t) \\ \frac{X(t+\Delta t) - X(t)}{\Delta t} &= -3uX(t) + 3uY(t) \end{aligned}$$

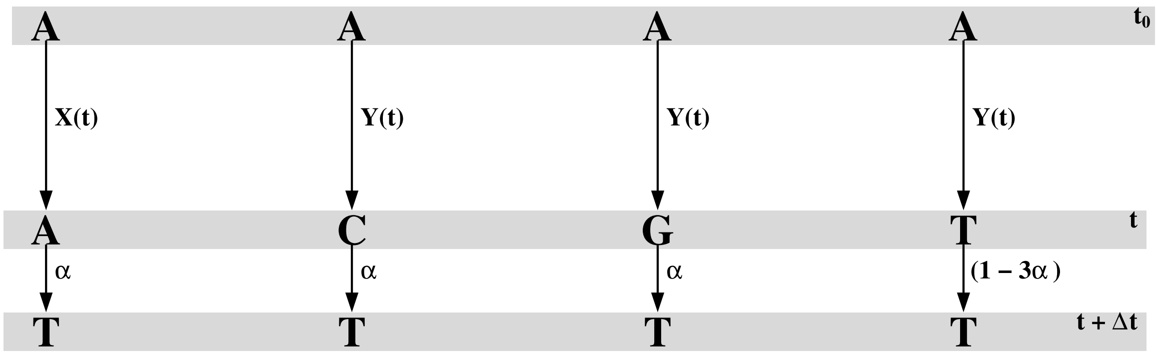

Figura 3.2: *Um exemplo ilustrando os quatro cenários possíveis para o caso em que o nucleotídeo em determinada posição na seqüência original se diferencia do nucleotídeo na mesma posição da seqüência observada no tempo $t + \Delta t$ no modelo de Jukes e Cantor.*

Quando Δt tende a zero, temos a seguinte equação diferencial:

$$\frac{dX(t)}{dt} = -3uX(t) + 3uY(t)$$

3.1.2 Os Nucleotídeos Diferem

Da mesma forma que no caso anterior, embora saibamos aqui que o nucleotídeo presente na posição em questão no tempo $t + \Delta t$ difira do nucleotídeo original, temos novamente quatro cenários, que são apresentados esquematicamente na Figura 3.2:

Cenário 1: Os nucleotídeos observados na posição em questão nos instantes t_0 e t são do mesmo tipo, de modo que uma mutação é observada entre t e $T + \Delta t$. Assim, a probabilidade deste cenário é dada pela função $P_1(t + \Delta t) = \alpha X(t)$.

Cenários 2 e 3: Nos cenários 2 e 3, os nucleotídeos observados nos três instantes são todos diferentes. O que difere os dois cenários é somente o nucleotídeo observade no instante t. Nestes casos, a probabilidade dos cenários é dada pelas funções $P_2(t + \Delta t) = P_3(t + \Delta t) = \alpha Y(t)$.

Cenário 4: Neste cenário, uma mutação é observada entre os instantes t_0 e t. Os nucleotídeos observados entre t e $t + \Delta t$ são do mesmo tipo. A probabilidade deste cenário é dada por $P_4 = (1 - 3\alpha)Y(t)$.

Como no caso em que os nucleotídeos observados em t_0 e $t + \Delta t$ não diferem, aqui, probabilidade de um nucleotídeo numa determinada posição ser diferente do nucleotídeo original é dada pela soma das probabilidades dos quatro cenários apresentados acima.

$$Y(t + \Delta t) = \alpha X(t) + 2\alpha Y(t) + (1 - 3\alpha)Y(t)$$

Podemos novamente usar a relação $\alpha = u\Delta t$:

$$\begin{aligned}
Y(t+\Delta t) &= u\Delta t X(t) + 2u\Delta t Y(t) + (1-3u\Delta t)Y(t) \\
Y(t+\Delta t) - Y(t) &= u\Delta t X(t) - u\Delta t Y(t) \\
\frac{Y(t+\Delta t) - Y(t)}{\Delta t} &= uX(t) - uY(t)
\end{aligned}$$

Quando Δt tende a zero, temos a seguinte equação diferencial:

$$\tfrac{dY(t)}{dt} = uX(t) - uY(t)$$

3.1.3 Encontrando as Equações

Como resultado das duas seções anteriores, temos duas equações diferenciais que descrevem as derivadas de primeira ordem de $X(t)$ e $Y(t)$ em termos das próprias funções $X(t)$ e $Y(t)$. Estas duas equações dão origem ao sistema de equações diferenciais de primeira ordem apresentado abaixo:

$$\begin{cases}
\frac{dX(t)}{dt} = -3uX(t) + 3uY(t) \\
\frac{dY(t)}{dt} = uX(t) - uY(t)
\end{cases}$$

Uma solução para este sistema de equações obedecendo às condições iniciais $X(t_0) = 1$ e $Y(t_0) = 0$, para $t_0 = 0$, pode ser encontrado no livro de Graur e Li [6, Capítulo 3].

$$\begin{aligned}
X(t) &= \frac{1}{4} + \frac{3}{4}e^{-4ut} \\
Y(t) &= \frac{1}{4} - \frac{1}{4}e^{-4ut}
\end{aligned}$$

Como o modelo de Jukes Cantor prevê apenas duas possibilidades para cada posição na seqüência: permanecer inalterada ou sofrer uma mutação, a soma das probabilidades de uma posição permanecer inalterada com as probabilidades das três mutações possíveis deve ser iqual a um. De fato, é fácil observar que:

$$X(t) + 3Y(t) = 1$$

3.2 O Modelo de 2 Parâmetros de Kimura

O modelo de Jukes e Cantor é bastante simples e, por conter apenas um parâmetro, também é o mais fácil de ser aplicado, todavia, uma das primeiras constatações que se pode fazer a respeito de mutações em seqüências de nucleotídeos ou aminoácidos é que

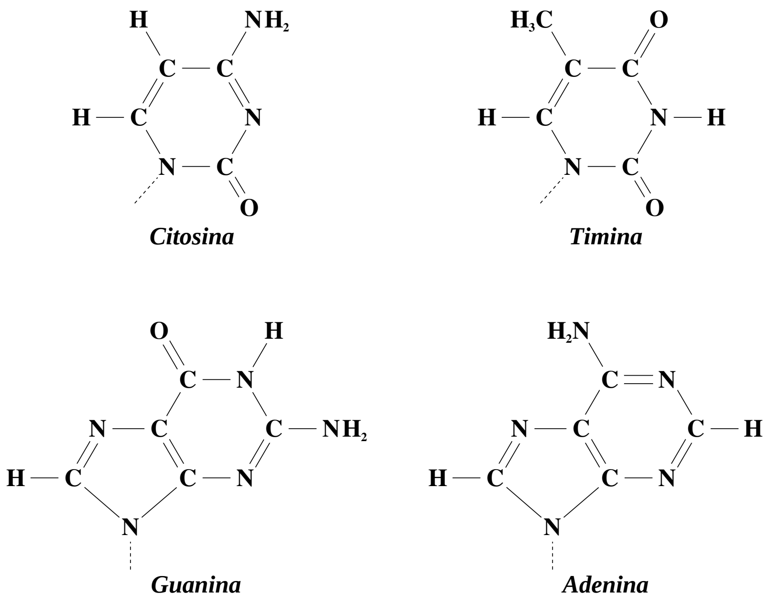

Figura 3.3: *As quatro bases nitrogenadas. Nucleotídeos são compostos por um açúcar (uma ribose, no caso do RNA, ou uma desoxirribose, no caso do DNA) unido a um grupo fosfato e uma base nitrogenada. O fosfato e o açúcar têm a função de unir os nucleotídeos formando a estrutura da fita de DNA (ou RNA), enquanto a função de "armazenamento" do código genético propriamente dito fica a cargo das bases nitrogenadas, que, por semelhanças estruturais claras nesta figura, são divididas em dois grupos: as pirimidinas (Citosina e Timina) e as purinas (Adenina e Guanina). A linha tracejada indica o ponto de ligação com o açúcar.*

certos tipos de mutações parecem ser mais freqüentes que outros. Em seqüências de aminácidos, esta característica é bem marcante no sentido em que aminoácidos com carga elétrica positiva dificilmente são trocados por aminoácidos com carga elétrica negativa. Da mesma forma, aminoácidos sem carga elétrica dificilmente são trocados por aminoácidos carregados. Em resumo, todos os tipos de substituições são possíveis, mas alguns são notavelmente mais freqüentes que outros.

Os quatro tipos de nucleotídeos que formam as moléculas de DNA podem ser divididos em dois grupos de nucleotídeos com estruturas químicas semelhantes: o das *purinas*, formado pelos nucleotídeos *Adenina* (A) e *Guanina* (G), e o das *pirimidinas*, formado pelos nucleotídeos *Timina* (T) e *Citosina* (C) (Veja a Figura 3.3 para maiores detalhes). Foi observado que, em seqüências de nucleotídeos, um certo tipo de mutação, denominado *transição*, onde ocorre a troca de um nucleotídeo qualquer pelo outro do mesmo grupo, são muito mais favoráveis que o outro tipo de mutação, denominada *transversão*, em que o novo nucleotídeo pertence a um grupo diferente do original.

Em 1980, Kimura propôs um modelo probabilístico para substituições em seqüências de nucleotídeos que leva em consideração diferenças nas taxas de transições e transversões por posição na seqüência de DNA. O modelo de dois parâmetros de Kimura estabelece que a probabilidade de ocorrer uma transição num intervalo de tempo fixo Δt, em uma determinada posição de uma seqüência de nucleotídeos é α. Assim como no modelo de Jukes e Cantor, no modelo de Kimura α também é proporcional ao tamanho de Δt e a proporcionalidade entre α e Δt é estabelecida pela expressão $\alpha = u\Delta t$. O segundo parâmetro do modelo de Kimura determina a probabilidade de ocorrer uma transversão num intervalo de tempo Δt. Essa probabilidade tem valor β e tem a proporcionalidade em relação ao tamanho do intervalo determinada pela expressão $\beta = v\Delta t$. Tanto u quanto v são constantes.

A presença de dois parâmetros no modelo de Kimura claramente torna um pouco mais complicada a determinação das equações do modelo, uma vez que no modelo de Kimura, não basta determinar apenas as equações para o casos em que a posição na seqüência permanece inalterada e para o caso em que ela se altera. As alterações agora podem ser de dois tipos e a probabilidade de cada um deles é diferente. Assim, o caso para o qual a posição da seqüência final difere da inicial é dividido em dois: o que a diferença se dá por uma transição e o que a diferença se dá por uma transversão.

Ao final desta sessão, chegaremos a três equações:

$X(t)$ determina a probabilidade de o nucleotídeo numa posição de uma seqüência no instante t ser do mesmo tipo que o nucleotídeo encontrado nesta mesma posição no instante t_0.

$Y(t)$ determina a probabilidade de os tipos dos nucleotídeos encontrados numa mesma posição de uma seqüência nos instantes t_0 e t serem diferentes e pertencerem ao mesmo grupo, ou seja, a probabilidade de os nucleotídeos se diferenciarem por um evento de *transição*.

$Z(t)$ determina a probabilidade de os tipos dos nucleotídeos encontrados numa mesma posição de uma seqüência nos instantes t_0 e t serem diferentes e pertencerem a grupos diferentes, ou seja, a probabilidade de os nucleotídeos se diferenciarem por um evento de *transversão*.

Analisando cada uma das situações e seus respectivos cenários, chegaremos a três equações diferenciais que relacionam cada uma das derivadas de primeira ordem em relação a t das equações com as próprias equações.

3.2.1 Os Nucleotídeos Não Diferem.

A primeira situação analisada será aquela em que um nucleotídeo em determinada posição na seqüência final não difere do nucleotídeo na mesma posição da seqüência inicial. Quatro

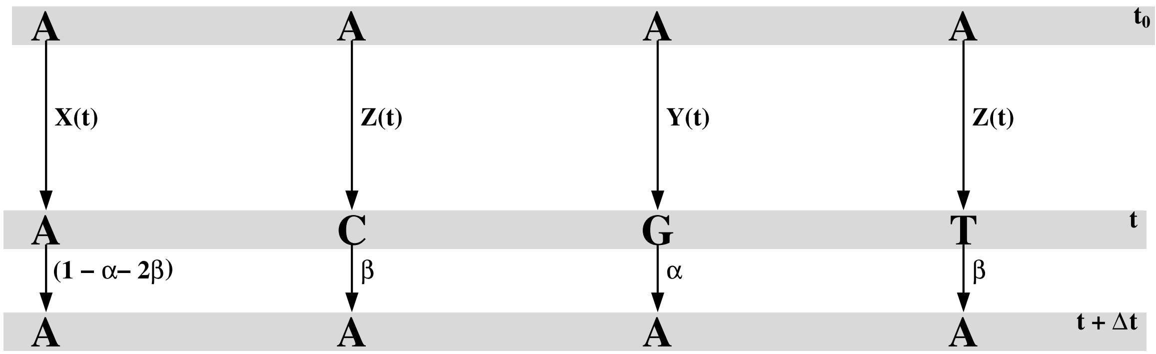

Figura 3.4: *Um exemplo ilustrando os quatro cenários possíveis para o caso em que o nucleotídeo em determinada posição na seqüência original não se diferencia do nucleotídeo na mesma posição da seqüência observada no tempo* $t + \Delta t$ *no modelo de dois parâmetros de Kimura.*

cenários diferentes podem levar a esta situação:

Cenário 1: Os nucleotídeos observados nos três instantes (t_0, t e $t + \Delta t$) são do mesmo tipo. A probabilidade deste cenário, é dada por $P_1(t + \Delta t) = (1 - \alpha - 2\beta)X(t)$.

Cenário 2: A posição, apesar de ter iniciado com o mesmo tipo de nucleotídeo encontrado na seqüência final, era ocupada por um nucleotídeo de um tipo que não pertence ao mesmo grupo que o nucleotídeo observado no instante t. A probabilidade deste cenário, é dada por $P_2(t + \Delta t) = \beta Z(t)$.

Cenário 3: A posição, apesar de apresentar o mesmo tipo de nucleotídeo da seqüência original, apresentava o outro nucleotídeo do mesmo grupo no instante t, voltando a apresentar um nucleotídeo do mesmo tipo do inicial em $t + \Delta t$. A probabilidade deste cenário, é dada por $P_3(t + \Delta t) = \alpha Y(t)$.

Cenário 4: O cenário 4 é exatamente idêntico ao cenário 2, desta vez considerando o outro tipo de nucleotídeo do grupo ao qual o nucletídeo inicial não pertence. Assim: $P_4(t + \Delta t) = P_2(t + \Delta t) = \beta Z(t)$.

Somando-se todas as probabilidades, chegamos à seguinte equação:

$$X(t + \Delta t) = (1 - \alpha - 2\beta)X(t) + \alpha Y(t) + 2\beta Z(t)$$

Lembrando das equações que relacionam α e β com o tamanho do intervalo de tempo considerado, temos:

$$\begin{aligned}
X(t + \Delta t) &= (1 - u\Delta t - 2v\Delta t)X(t) + u\Delta t Y(t) + 2v\Delta t Z(t) \\
X(t + \Delta t) &= X(t) - (u + 2v)X(t)\Delta t + uY(t)\Delta t + 2vZ(t)\Delta t \\
\frac{X(t + \Delta t) - X(t)}{\Delta t} &= -(u + v)X(t) + uY(t) + 2vZ(t)
\end{aligned}$$

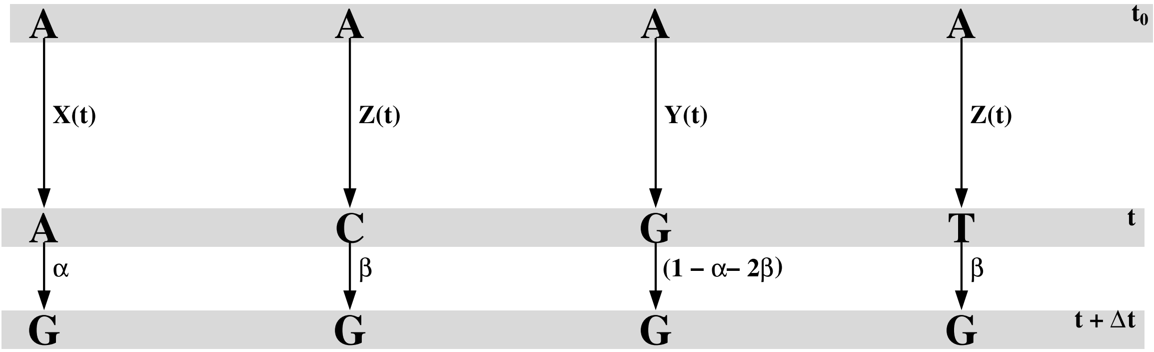

Figura 3.5: *Um exemplo ilustrando os quatro cenários possíveis para o caso em que o nucleotídeo em determinada posição na seqüência original se diferencia do nucleotídeo na mesma posição da seqüência observada no tempo $t + \Delta t$ por um evento de transição no modelo de dois parâmetros de Kimura.*

Quando o tamanho de Δt tende a zero, temos

$$\frac{dX(t)}{dt} = -(u + 2v)X(t) + uY(t) + 2vZ(t)$$

3.2.2 Os Nucleotídeos Diferem Por Um Evento de Transição

Também para este caso temos que analisar quatro cenários diferentes.

Cenário 1: No instante t, o nucleotídeo observado em determinada posição era do mesmo tipo do nucleotídeo observado na mesma posição no intante t_0. Entre t e Δt notamos uma diferença por um evento de transição. A probabilidade deste cenário é dada por $P_1(t + \Delta t) = \alpha X(t)$.

Cenário 2: Nos instantes t_0 e t, os nucleotídeos numa mesma posição se diferenciam por um evento de transversão, mas outro evento de transversão é observado entre t e $t + \Delta t$, desta vez para o outro nucleotídeo do mesmo grupo do nucleotídeo originalmente na posição em questão. Neste caso, a probabilidade do cenário é dada por $P_2(t + \Delta t) = \beta Z(t)$.

Cenário 3: No cenário 3, a transversão é observada já no instante t, sendo que os tipos dos nucleotídeos na posição em questão nos tempos t e $t + \Delta t$ não diferem. Assim: $P_3(t + \Delta t) = (1 - \alpha - 2\beta)Y(t)$.

Cenário 4: Este último cenário é exatamente idêntico ao 2, desta vez considerando o outro tipo de nucleotídeo do grupo ao qual o nucletídeo inicial não pertence como nucleotídeo intermediário. Assim: $P_4(t + \Delta t) = P_2(t + \Delta t) = \beta Z(t)$.

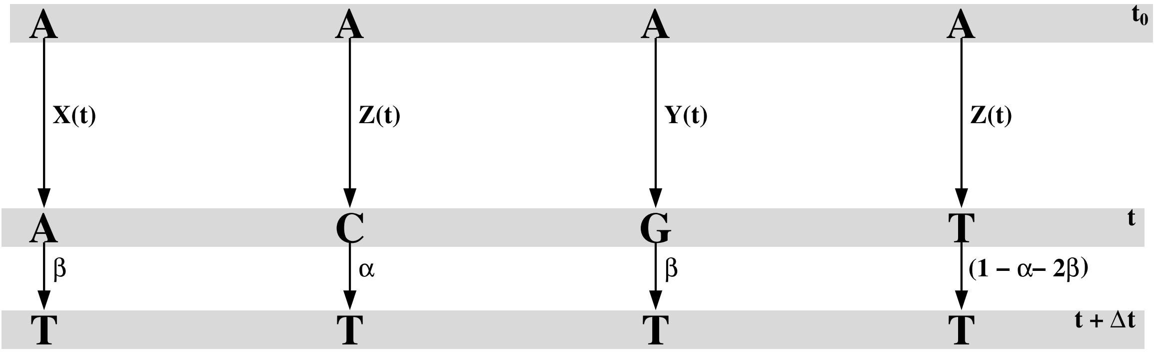

Figura 3.6: *Um exemplo ilustrando os quatro cenários possíveis para o caso em que o nucleotídeo em determinada posição na seqüência original se diferencia do nucleotídeo na mesma posição da seqüência observada no tempo $t + \Delta t$ por um evento de transversão no modelo de dois parâmetros de Kimura.*

A equação procurada é mais uma vez obtida pela soma dos quatro cenários:

$$Y(t + \Delta t) = \alpha X(t) + (1 - \alpha - 2\beta)Y(t) + 2\beta Z(t)$$

Usamos novamente as equações que relacionam α e β com um intervalo de tempo e obtemos:

$$\begin{aligned}
Y(t + \Delta t) &= u\Delta t X(t) + (1 - u\Delta t - 2v\Delta t)Y(t) + 2v\Delta t Z(t) \\
Y(t + \Delta t) &= uX(t)\Delta t + Y(t) - (u + 2v)Y(t)\Delta t + 2vZ(t)\Delta t \\
\frac{Y(t + \Delta t) - Y(t)}{\Delta t} &= uX(t) - (u + 2v)Y(t) + 2vZ(t)
\end{aligned}$$

Mais uma vez, analisamos o caso em que Δt tende a zero.

$$\tfrac{dY(t)}{dt} = uX(t) - (u + 2v)Y(t) + 2vZ(t)$$

3.2.3 Os Nucleotídeos Diferem Por Um Evento de Transversão

Este último caso também possui quatro cenários diferentes a serem analisados. A Figura 3.6 fornece um exemplo de cada um destes cenários.

Cenário 1: Os nucleotídeos observados nos instantes t_0 e t são de um mesmo tipo. O evento de transversão só é observado entre os instantes t e $t + \Delta t$. A probabilidade deste cenário é dada pela função $P_1(t + \Delta t) = \beta X(t)$.

Cenário 2: Após um intervalo de tempo t, os tipos dos nucleotídeos numa mesma posição se diferenciavam por um evento de transversão, mas um evento de transição é observado entre t e $t + \Delta t$. Neste caso, a probabilidade do cenário é dada por $P_2(t + \Delta t) = \alpha Z(t)$.

Cenário 3: Uma transição é observada no tempo t, e a transversão só é observada mesmo entre t e $t + \Delta t$. Deste modo, podemos descrever este cenário através da equação: $P_3(t + \Delta t) = \beta Y(t)$.

Cenário 4: Após um intervalo de tempo t, os nucleotídeos numa mesma posição se diferenciavam por um evento de transversão. Desta vez, porém, não se observa diferença alguma ocorrida na posição em questão entre t e $t + \Delta t$. Neste caso, a probabilidade do cenário é dada por $P_4(t + \Delta t) = (1 - \alpha - 2\beta)Z(t)$.

A última das equações que procuramos também é obtida pela soma dos quatro cenários:

$$Z(t + \Delta t) = \beta X(t) + \alpha Z(t) + (1 - \alpha - 2\beta)Z(t) + \beta Y(t)$$

e pode ser simplificada da seguinte maneira:

$$Z(t + \Delta t) = \beta X(t) + (1 - 2\beta)Z(t) + \beta Y(t)$$

Uma última vez usamos as equações que relacionam α e β com o tamanho do intervalo de tempo e obtemos:

$$\begin{aligned} Z(t + \Delta t) &= v\Delta t X(t) + (1 - 2v\Delta t)Z(t) + v\Delta t Y(t) \\ Z(t + \Delta t) &= vX(t)\Delta t + Z(t) - 2vZ(t)\Delta t + vY(t)\Delta t \\ \frac{Z(t + \Delta t) - Z(t)}{\Delta t} &= vX(t) + vY(t) - 2vZ(t) \end{aligned}$$

Assim como as outras equações, esta também pode ser analisada no caso em que Δt tende a zero.

$$\frac{dZ(t)}{dt} = vX(t) + vY(t) - 2vZ(t)$$

3.2.4 Encontrando as Equações

Até agora o que temos são três equações diferenciais que relacionam as derivadas das três funções que queremos descobrir com as próprias funções. Para chegar às funções desejadas, devemos encontrar uma solução para o sistema abaixo:

$$\begin{cases} \frac{dX(t)}{dt} = -(u + 2v)X(t) + uY(t) + 2vZ(t) \\ \frac{dY(t)}{dt} = uX(t) - (u + 2v)Y(t) + 2vZ(t) \\ \frac{dZ(t)}{dt} = vX(t) + vY(t) - 2vZ(t) \end{cases}$$

Graur e Li [6, Capítulo 3] fornecem a seguinte solução para este sistema de equações, com condições iniciais $X(t_0) = 1$, $Y(t_0) = Z(t_0) = 0$, para $t_0 = 0$:

$$\begin{aligned} X(t) &= \frac{1}{4} + \frac{1}{4}e^{-4vt} + \frac{1}{2}e^{-2(u+v)t} \\ Y(t) &= \frac{1}{4} + \frac{1}{4}e^{-4vt} - \frac{1}{2}e^{-2(u+v)t} \\ Z(t) &= \frac{1}{4} - \frac{1}{4}e^{-4vt} \end{aligned}$$

Como o modelo de dois parâmetros de Kimura prevê apenas três possibilidades para uma posição em uma seqüência: permasnecer inalterada, sofrer uma transição ou sofrer uma transversão, sendo que o último caso pode ocorrer em duas variações para qualquer nucleotídeo, podemos verificar facilmente que a expressão

$$X(t) + Y(t) + 2Z(t) = 1$$

é sempre verdadeira para as funções $X(t)$, $Y(t)$ e $Z(t)$ encontradas anteriormente.

3.3 Variação de Parâmetros ao Longo das Seqüências

Midelos de substituição como os de Jukes e Cantor e o de dois parâmetros de Kimura, apesar de bastante simples, são bem fundamentados e se aproximam do modo como uma posição numa seqüência varia. De fato, não seria estranho se a evolução de regiões não funcionais de moléculas de DNA evoluíssem de uma maneira muito parecida com a prevista pelo modelo de Kimura, por exemplo. Por outro lado, os trechos mais interessantes das moléculas de DNA, assim como as seqüências de aminoácidos que compõem as proteínas, têm funções que dependem do fato de certas posições serem ocupadas por determinado nucleotídeo, ou aminoácido, o que torna a mutação em algumas posições da seqüência menos freqüentes que em outras, ou mesmo proibidas.

Quando paramos de focar a evolução de uma posição específica para avaliarmos a evolução da seqüência como um todo, é preciso ter consciência que a velocidade com que posições da seqüência variam ao longo do tempo não é a mesma para todas as posições e esta variação na taa de evolução deve fazer parte de qualquer modelo que tente exlpicar a evolução de uma seqüência como um todo.

As seçóes seguintees resumem dois modelos bastante comuns para a variação da taxa de evolução entre as posições de uma seqüência.

3.3.1 Modelos RAS

O termo *RAS* é uma abreviação de *R*ate *A*cross *S*ites e se aplica a todo modelo que utiliza um modelo de substituição de posições em seqüências, como os apresentados an-