UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TRABALHO DE GRADUAÇÃO

Propriedades e Algoritmos para Recombinações entre espécies numa Rede Filogenética

Vinício Tavares de Melo Costa da Silva
Orientadora: Katia Silva Guimarães

Recife, 11 de março de 2005

Resumo

A recombinação é um tipo de mutação genética que acontece com bastante freqüência no DNA de todas as espécies existentes, ela é de grande importância na pesquisa de cura de doenças, melhoria genética de espécies, entre outros fins. No entanto, o estudo com recombinações começou recentemente, com pouquíssimos trabalhos publicados na área.

Este trabalho irá apresentar um método para calcular um limite inferior e um limite superior para o número máximo de recombinações que pode acontecer numa rede filogenética. Os cálculos serão feitos a partir de um conjunto de entradas formado por seqüências binárias que representam algumas características polimórficas de um gene ou espécie.

Agradecimentos

À minha família por todo o apoio, disciplina e incentivo dados nesses anos de faculdade e em toda a minha vida.

À minha amiga e namorada Kaline pelo incentivo.

Ao meu amigo Daniel Patinhas pela motivação.

Aos meus amigos da AVCIn pelos momentos de diversão.

À minha orientadora Katia por me fazer gostar da bioinformática e me encaminhar neste trabalho.

Aos professores do CIn por todo conhecimento e lições de vida passados.

Aos meus colegas de trabalho do LaViTE por não cobrarem muito de mim nesta reta final.

Índice

Introdução

A recombinação é um tipo de mutação genética extremamente importante, pois se acredita que ela seja a chave para localizar genes que influenciam em doenças herdadas geneticamente.

A importância dessa área de estudos está principalmente na possibilidade de mapear a origem de trechos do genoma associados a certas características ou sintomas desenvolvidos na vida do indivíduo. Através de um grupo de cobaias sem relações de parentesco, pode-se dizer que partes dos seus genomas vieram de um ancestral comum. E a partir daí, pode-se entender e determinar a história daquele trecho do genoma e fazer o mapeamento dessa parte da seqüência genética e possíveis evoluções que ela tenha sofrido em outras espécies. Por exemplo, se tiver um determinado organismo resistente a uma certa bactéria, através da análise da história de seu DNA pode-se descobrir como ele adquiriu essa resistência.

O número de trabalhos publicados na área, que engloba o comportamento das recombinações e algoritmos e propriedades das mesmas através do uso de redes filogenéticas, é bastante reduzido e ainda não explorou uma série de problemas e propriedades das recombinações.

A idéia deste trabalho é exatamente atacar uma dessas áreas ainda não exploradas do estudo das recombinações em redes filogenéticas. Um problema já resolvido é o de determinar se existe uma *galled tree* a partir de um conjunto de entradas formado por seqüências do mesmo tamanho e montá-la caso exista [1]. Um outro problema, esse com solução heurística, é

o da determinação de um limite inferior para o número mínimo de recombinações presentes numa rede filogenética [4].

Os artigos [1] e [4] serviram de base para este trabalho, que tem como principal objetivo desenvolver um método heurístico para que seja calculado um intervalo numérico em que o número máximo de recombinações de uma rede filogenética está contido. Esse intervalo será composto de um limite inferior e um limite superior. Para o cálculo do intervalo é imaginada a hipótese de que sempre que uma mutação puder ser recombinação, ela será. Ou seja, se uma certa mutação puder ser uma transferência de bases nucléicas ou uma recombinação, este trabalho assumirá que ela vai ser uma recombinação, isso é o número máximo de recombinações que pode existir.

Este estudo tem ainda como objetivo secundário descrever algumas propriedades das redes filogenéticas observadas durante as pesquisas. Isso contribuirá para um melhor entendimento da lógica de comportamento das redes filogenéticas e pode servir como referência para futuros trabalhos, já que não foram encontradas fontes que relatassem essas propriedades, elas serão descritas ao longo do trabalho na forma de explicações bem detalhadas e figuras de recombinações.

Essa análise aprimorada das redes filogenéticas com recombinações e de suas propriedades servirá para ajudar no desenvolvimento de algoritmos mais completos, para melhor manusear essas redes. A conseqüência disso é um melhor entendimento das relações entre as diferentes espécies, e num estudo mais dirigido, as mutações sofridas nos mesmos genes presentes em indivíduos diferentes.

Contexto

Até meados do século XVIII todos achavam que as espécies sempre foram como existiam naquela época, representando uma teoria criacionista. No final do século XVIII e início do século XIX, começaram a surgir os primeiros questionamentos sobre a origem das espécies. Carlos Lineu e Comte de Buffon, classificadores de espécies da época, foram os pioneiros nessa área. Mais tarde, Jean Baptiste Lamarck e Erasmus Darwin, avô de Charles Darwin, publicaram os primeiros trabalhos defendendo que as espécies estavam em constante evolução para melhor se adaptarem aos meios. Posteriormente, Charles Darwin e Alfred Wallace fizeram um estudo que durou mais de 20 anos, onde coletaram diversas amostras de uma infinidade de espécies por todo o mundo e classificaram-nas quanto às suas proximidade e origens, gerando o famoso livro de Darwin, A Origem das Espécies, que foi um o maior avanço da filogenética na época. A partir desse ponto, começou a surgir a necessidade de classificar as variações entre as espécies bem como a proximidade entre elas, foi criada então uma estrutura conhecida como árvore filogenética.

Uma árvore filogenética é uma representação gráfica em forma de árvore, com vértices e arestas, que vão se ramificando, e todos os indivíduos nela presentes partem de um ancestral comum, posicionado em seu topo, a raiz. Cada aresta da árvore representa uma mutação na espécie, ou, a derivação de um ancestral comum em mais de um organismo. Cada vértice representa um ponto na escala evolutiva, e cada folha uma espécie diferente.

A proximidade entre as folhas indica que aquelas espécies têm uma certa proximidade evolutiva (Ver Figura 1).

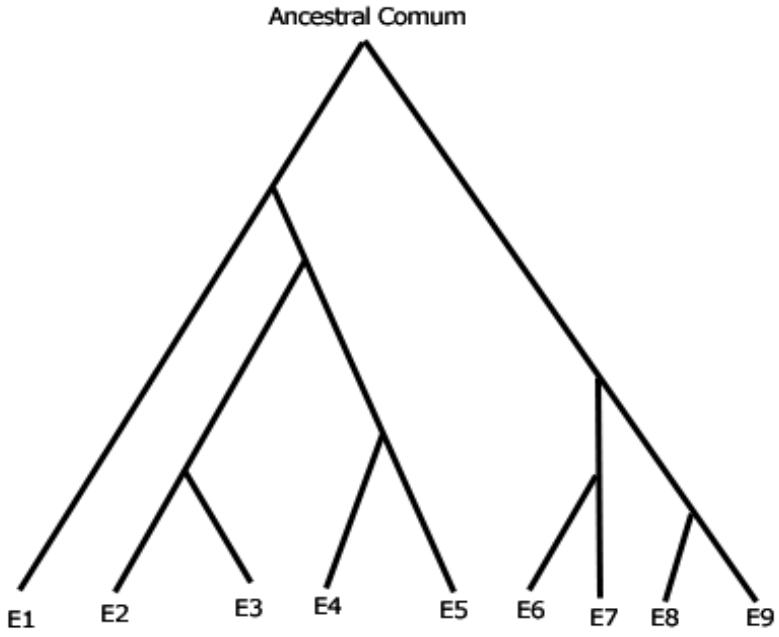

Figura 1: Árvore filogenética

As nove espécies acima (E1 a E9) derivam do mesmo ancestral comum. As espécies E6 e E7 são muito mais próximas entre si na escala evolutiva que a E7 da E4.

Após as primeiras experiências com o cruzamento de diferentes variações de uma espécie, que foram realizadas pelo monge austríaco Gregor Mendel no final do século XIX, surgiu uma nova área de estudo, a genética. Quase um século depois, com a descoberta do DNA por Francis Crick e James Watson, esse trabalho de classificação das espécies começou a se tornar uma ciência mais exata. Então, surgiu definitivamente a filogenética, sendo possível fazer uma classificação mais precisa das espécies, assim como uma estimativa de suas proximidades e quanto tempo levaram para evoluir e passar de um ponto evolutivo a outro.

Devido à grande quantidade de informação genômica originada com as recentes descobertas, as árvores filogenéticas passaram a não satisfazer mais as necessidades, pois não suportavam certas propriedades estruturais nem conseguiam modelar as novas descobertas genéticas. Fenômenos como recombinação, mutação recorrente e reversa, transferência de genes horizontal, conversão de genes e elementos genéticos móveis, só podiam ser representados e entendidos com uma estrutura mais complexa, como um grafo. Foi então que surgiram as redes filogenéticas, como generalizações para as árvores que vinham sendo usadas.

Uma das últimas propriedades evolutivas descobertas é a recombinação, que é uma evolução decorrente do cruzamento de dois indivíduos gerando um novo organismo, que possui um trecho do DNA de cada um dos dois indivíduos que o geraram. Numa rede filogenética, um nó recombinante é um nó para o qual duas ou mais arestas se dirigem (Ver Figura 2).

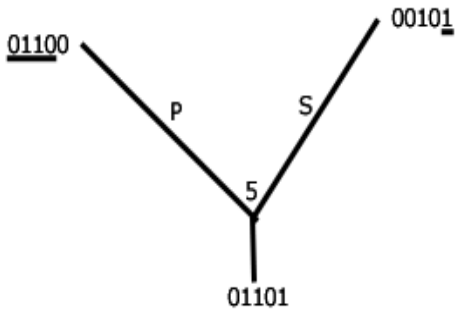

Figura 2: Ponto de recombinação

Duas arestas vindas de dois nós distintos da rede filogenética se juntam, uma dá o prefixo (aresta P) e a outra o sufixo (aresta S) para ser formada uma nova seqüência no ponto de recombinação 5.

Uma rede com filogenia perfeita seria topologicamente direcionada, ou seja, uma árvore, sem a presença de nós de recombinação.

As redes filogenéticas agregam cinco características principais:

- São grafos direcionados acíclicos que têm um nó raiz para onde nenhuma aresta se dirige,
- Têm mutações associadas às suas arestas,
- Uma seqüência associada a cada nó não recombinante,
- A especificação de uma seqüência "prefixo" e uma seqüência "sufixo" a cada nó recombinante e,
- A associação de um ponto e uma seqüência de recombinação para cada nó recombinante.

Na biologia, as redes filogenéticas são conhecidas como a possível história da evolução das seqüências presentes na rede, na premissa que existe um único ancestral comum a todas as seqüências.

Na rede filogenética, as seqüências são representadas como cadeias de caracteres binários, onde cada caractere representa um SNP (*Single Nucleotide Polymorphism*), por convenção, a raiz é uma cadeia de zeros, e a medida que vão aparecendo "1"s nos pontos de evolução, significa que o ancestral comum sofreu uma mutação naquele ponto, ou seja, o que estava ali foi mudado.

Numa rede filogenética com um ponto de recombinação X, se traçarmos caminhos partindo de X em direção à raiz (ancestral comum) e esses caminhos tiverem a possibilidade de se cruzar, temos então um ciclo de recombinação (Ver Figura 3).

Se existe um ciclo de recombinação que não compartilha nós com nenhum outro, ele é chamado de "*gall*". Uma rede filogenética é chamada de *galled tree* se todos os seus ciclos de recombinação forem *galls*.

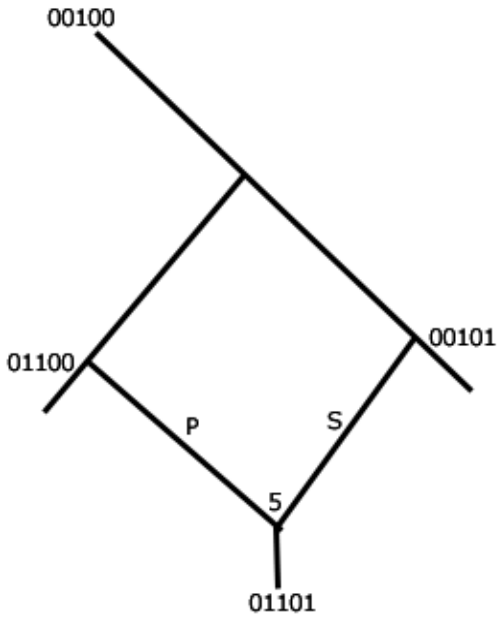

Figura 3: Ciclo de recombinação

Os dois vértices de que derivam as arestas P e S do ponto de recombinação 5 foram originados do mesmo vértice (00100), formando assim um ciclo de recombinação.

Metodologia

Num primeiro momento decidiu-se separar as redes filogenéticas em três grupos para facilitar o estudo de suas propriedades. São eles: Redes filogenéticas sem ciclos de recombinação sobrepostos; Redes filogenéticas com uma sobreposição de ciclos de recombinação; e Redes filogenéticas com múltiplas sobreposições de ciclos de recombinação.

Após essa divisão, serão analisados vários exemplos de cada um dos três casos para ver o número máximo de ciclos de recombinação. Será desenvolvida uma ferramenta que constrói *Galled Trees*, ou seja, redes filogenéticas sem ciclos de recombinação sobrepostos. Essa ferramenta será útil porque utiliza algoritmos que identificam conflitos entre as colunas das seqüências, nós recombinantes e ciclos de recombinação.

A ferramenta para a construção de *Galled Trees* será então adaptada para analisar o número de recombinações necessárias para formar as seqüências dadas como entrada. Porém, o algoritmo utilizado [1] para a montagem das árvores usa apenas uma forma de arrumação dos galhos, dentre as inúmeras possíveis. A idéia é adaptar o algoritmo e a ferramenta para fazer uma estimativa do número máximo de recombinações para *Galled Trees*, será criado um método para calcular um limite superior e inferior para o número máximo de recombinações.

A etapa seguinte será estender essa ferramenta para calcular o número máximo de recombinações em redes filogenéticas com um ciclo de recombinação sobreposto, o que se supõe ser mais um caso base do estudo.

A próxima etapa é adaptar a ferramenta para calcular também o número máximo de recombinações quando a rede filogenética tiver múltiplos ciclos sobrepostos. Assim, teremos uma ferramenta que faz esse cálculo para todos os casos de redes filogenéticas.

A ferramenta desenvolvida lerá as seqüências de entrada e dará como resultado o intervalo numérico em que está contido o número máximo de recombinações que pode existir para formar as seqüências a partir do ancestral comum. Esse intervalo será composto de um limite inferior e um limite superior.

Construção de *Galled Trees*

O objetivo desta primeira etapa é construir ou determinar a impossibilidade de construção de *galled trees*. Esta análise será feita a partir de um conjunto de seqüências binárias de mesmo tamanho, que representam etapas evolutivas de uma rede filogenética, dadas como entrada. Esta etapa estudará apenas os casos em que não existem ciclos de recombinação sobrepostos na estrutura (Ver Figura 4). O algoritmo utilizado será o de Gusfield[1] com pequenas adaptações feitas para facilitar a implementação e o entendimento do mesmo.

O primeiro passo desta etapa consiste em criar uma matriz que represente a rede filogenética. Essa matriz tem em cada uma de suas linhas uma seqüência do conjunto de entrada. No processo de criação da matriz serão eliminadas as linhas repetidas, caso que costuma ocorrer com freqüência e que pode influenciar nos resultados do algoritmo.

Depois de feita esta arrumação, pode-se facilmente deduzir que cada coluna da matriz significa o estado que um certo trecho daquelas seqüências estava no ponto evolutivo representado pela linha. Vale salientar que cada seqüência pode estar presente em níveis diferentes de evolução, o que explica o fato de termos vários estágios evolutivos num mesmo conjunto de entradas.

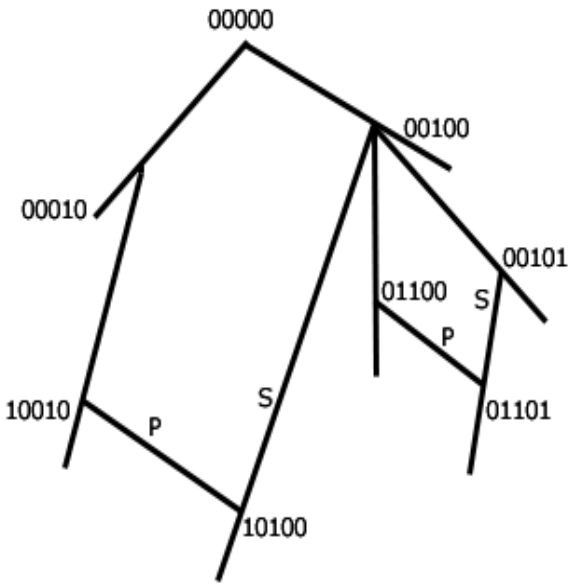

Figura 4: Ciclos de recombinação não sobrepostos

Rede filogenética com dois ciclos de recombinação não sobrepostos.

Essa representação nos mostra com clareza quais pontos evolutivos causaram mudanças naquele trecho da seqüência.

O ancestral comum não aparece nessa matriz, já que por ser uma linha formada apenas por 0s (zeros), ele não influenciaria nos resultados. As outras seqüências são formadas por 0s (zeros) e 1s (uns), onde o 0 (zero) representa um lócus inalterado, igual ao do ancestral comum e o 1 (um) representa um lócus que sofreu alguma mudança ao longo de sua evolução.

O passo seguinte é verificar se existem ciclos de recombinação, conseqüentemente, nós recombinantes na estrutura. Essa verificação é feita comparando-se duas a duas cada coluna da matriz com todas as demais. A comparação é feita utilizando o seguinte procedimento, cada elemento da primeira coluna que está sendo comparada é colocado lado a lado com o elemento presente na mesma linha da segunda coluna, formando um par. Se ao juntar todos os pares decorrentes da comparação das duas colunas em

questão, estiverem presentes os pares 1,0; 0,1 e 1,1, temos um conflito entre essas duas colunas, o que caracteriza a existência de um ciclo de recombinação (Ver Figura 5).

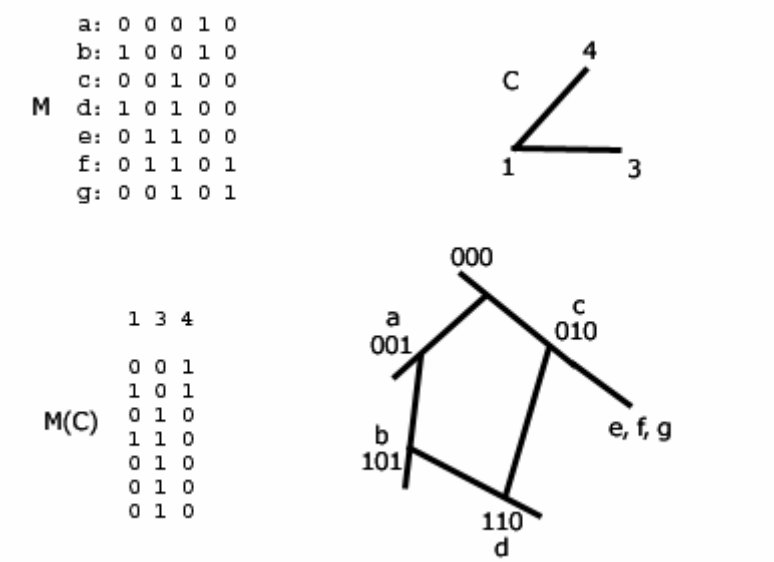

Figura 5: Matriz M e grupo de conflito C com sua matriz

Podemos observar na matriz M que a coluna 1 tem conflito com as colunas 3 e 4, formando assim um grupo de conflito C. A matriz M(C) é a matriz M restrita às colunas pertencentes ao grupo de conflito C.

A explicação para isso é simples, um par 1,0 significa que na primeira coluna existe uma variação da seqüência original daquele trecho do genoma (número 1), enquanto na segunda coluna, o lócus está preservado (número 0). Um par 0,1 significa o contrário, a primeira coluna contém um trecho original, enquanto a segunda coluna tem uma variação. E, um par 1,1 significa que ambas as colunas possuem variações da seqüência original. Em outras palavras, a presença desses três pares significa que ocorreu uma recombinação entre as seqüências das linhas em que os pares são 1,0 e 0,1, e que essa recombinação originou a seqüência da linha em que o par é 1,1,

assim, podemos concluir que a linha em que o par é 1,1 é um nó de recombinação.

O terceiro passo do processo é criar grupos conflitantes. Por exemplo, a coluna 1 entra em conflito com as colunas 2 e a 4, enquanto a coluna 2 entra em conflito com as colunas 1 e 5, temos então um grupo conflitante que contem as colunas 1, 2, 4 e 5. Esses pequenos grupos serão chamados de *galls*, e representam um ciclo de recombinação.

O quarto passo é criar matrizes de conflito a partir dos grupos conflitantes. Uma matriz de conflito é semelhante à matriz que representa a rede filogenética, a única diferença é que ao invés de conter todas as colunas, a matriz de conflito só tem as colunas pertencentes a um único grupo. Cada matriz representará um *gall*.

O quinto passo é a remoção dos nós de recombinação das matrizes de conflito para que seja testado se os ciclos de recombinação em questão são sobrepostos ou não. Os nós de recombinação são os que formaram os pares 1,1. Eles devem ser removidos um a um, nunca ao mesmo tempo (Ver Figura 6). Ao tirar um nó, deve-se testar se a matriz de conflitos tem filogenia perfeita, ou seja, se não existem colunas conflitantes. Esse procedimento deve ser feito para cada um dos nós de recombinação de cada um dos grupos de conflito, sempre um por vez. O algoritmo usado para isso é o mesmo usado anteriormente, a comparação de cada coluna com as demais checando a existência dos pares 1,0; 0,1 e 1,1 ao mesmo tempo. Se após a remoção do nó, a matriz de conflitos continuar com colunas entrando em conflito, significa que essa rede possui no mínimo uma sobreposição de ciclos de recombinação, o que impede a construção de uma *galled tree*. O algoritmo pára aqui se a impossibilidade de construção for detectada, já que os passos seguintes explicam como determinar o parentesco entre os *galls* e

uni-los, tarefa que não é possível se os ciclos de recombinação tiverem sobreposições.

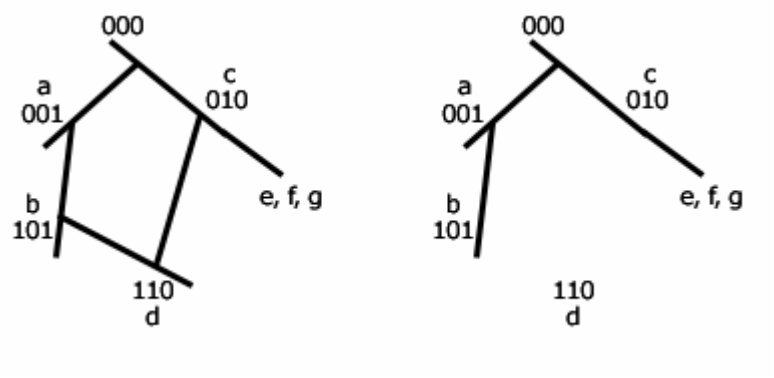

Figura 6: Remoção de um nó recombinante

O nó "d" foi removido do ciclo de recombinação, criando uma estrutura em que se houver a recombinação de dois de seus nós, "b" e "c", o nó removido será formado.

Se após a remoção dos nós recombinantes, todas as matrizes de conflitos ficarem com filogenia perfeita, partimos para o sexto passo, que é a fase de determinar o parentesco e logo depois conectar os *galls* (Ver Figura 7). Para isso, são criadas matrizes, ou seja, matrizes de passagem, cada uma delas é equivalente a uma matriz de conflitos, e representa o caminho evolutivo das seqüências, por quais *galls* passa a evolução de uma determinada seqüência a partir do ancestral comum. Com essas matrizes dá para saber qual *gall* é antecessor de qual. As matrizes *pass-through* na verdade são arrays, em que o número de elementos é o número total de seqüências. Cada posição recebe o valor 0 (zero) ou 1 (um), o 0 (zero) significa que os trechos do genoma da seqüência que fazem parte da matriz de conflitos em questão não sofreram alterações naquele ponto, enquanto o 1 significa que pelo menos um dos trechos do genoma sofreu alterações. A construção dessa matriz é simples, cada linha da matriz de conflitos equivale

a uma posição da matriz *pass-through*, então, deve-se analisar se na linha em questão existe algum número 1 (um), caso exista, é colocado um 1 (um) na posição equivalente da matriz *pass-through*, se não, é colocado um 0 (zero).

Após esse procedimento é feita uma análise das matrizes *pass-through* para determinar qual *gall* é antecessor de qual na escala evolutiva. Com as informações de por quais *galls* a evolução das seqüências passou, contidas nas matrizes *pass-through*, podemos ter uma idéia do caminho seguido pela evolução daquelas seqüências e assim determinar se um *gall* antecede ou descende de outro. Por exemplo, se por uma matriz de passagem passa o caminho evolutivo das seqüências d, e, f, g e h, e por outra passa apenas o caminho das seqüências f, g e h, podemos afirmar que o *gall* representado pela primeira matriz é antecessor do outro *gall*, já que, seguindo aquele caminho, a evolução das seqüências d e e parou na primeira matriz, enquanto f, g e h continuaram evoluindo a partir daquele ponto. Um *gall* é antecessor direto de outro se existe uma aresta que liga um ponto do primeiro *gall* a um ponto do segundo gall diretamente. Baseado nesse modelo, o mesmo procedimento deverá ser utilizado com todas as colunas da matriz *pass-through*, e comparando uma a uma e encaixando-as para que as árvores sejam formadas (Ver Figura 8).

A partir desse ponto, temos um esqueleto da árvore, com todos os *galls*, provindos de matrizes cujas colunas têm conflitos, encaixados de acordo com o parentesco entre eles, para concluir o processo basta colocar as folhas e componentes sem conflitos. Esse é o sétimo e último passo do processo. Ele adicionará as seqüências cujas colunas não têm conflitos e colocará as seqüências com colunas conflitantes em suas folhas específicas, possivelmente adicionando vértices da árvore fora de alguns *galls*.

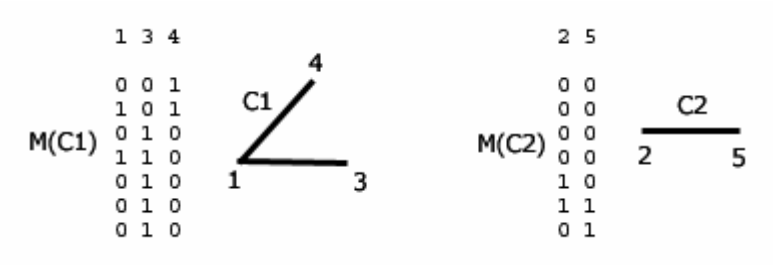

Figura 7: Passagem pelos componentes conflitantes

Como se pode ver, o caminho para todas as folhas passam (*pass-through*) pelo componente C1, mas apenas os caminhos para "e", "f" e "g" passam pelo componente C2.

A primeira ação a ser feita é dividir as seqüências em dois grupos, as que têm valor 1 (um) em colunas conflitantes e as que não têm valor 1 (um) em nenhuma coluna conflitante, o segundo grupo forma galhos de filogenia perfeita na árvore, ou seja, que não participam de nenhum ciclo de recombinação, esses devem estar sempre no topo, em cima de qualquer *gall*. Caso não existam seqüências no segundo grupo, o topo da árvore será formado apenas pela raiz, o ancestral comum. Terminado esse processo temos que posicionar as seqüências do primeiro grupo, para isso, devemos antes de tudo identificar em que parte da árvore elas devem ficar. Esse procedimento é feito analisando-se quais *galls* estão no caminho evolutivo das colunas da seqüência em que o valor é 1 (um), e então ir descendo até o *gall* mais baixo da árvore, que também tem valor 1 (um) para as colunas em questão, é lá que deve ser colocada a folha que representa a seqüência. Ela será ligada a um nó do ciclo de recombinação em que apenas uma mutação é necessária para formar a seqüência em questão.

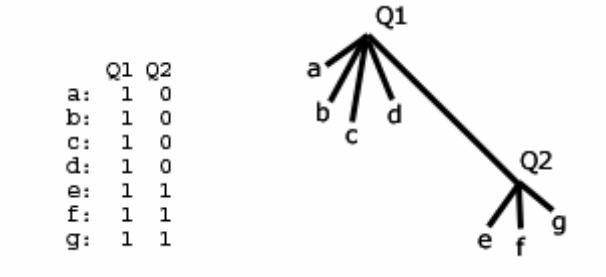

Figura 8: Matriz Pass-through

À esquerda, matriz *pass-through* e à direita o esqueleto da *galled-tree* com filogenia perfeita, montado a partir da matriz *pass-through.*

Após esses sete passos temos a construção da *galled tree* a partir de um conjunto de seqüências de entrada concluída, ou, se for o caso, a determinação da impossibilidade de sua construção por motivo de sobreposição de ciclos de recombinação.

Adaptação e Extensão do Algoritmo

Até aqui foi criada uma ferramenta para construir ou determinar a impossibilidade de construção de *Galled Trees*, ela utiliza um algoritmo que identifica as recombinações numa rede filogenética e consegue diferenciar ciclos de recombinação sem sobreposições de ciclos de recombinação com sobreposições.

O objetivo desta etapa do trabalho é adaptar o algoritmo utilizado para que ele identifique os nós recombinantes e estime o número máximo de recombinações não só para redes filogenéticas sem ciclos de recombinação sobrepostos, mas para qualquer tipo de rede filogenética.

O primeiro passo desta etapa é adaptar o algoritmo de construção de *Galled Trees* para contar o número de recombinações nas redes filogenéticas sem ciclos de recombinação sobrepostos. Para isso deve-se então estudar o algoritmo utilizado.

O algoritmo da ferramenta constrói *Galled Trees* a partir de redes filogenéticas sem ciclos de recombinação sobrepostos, no entanto, não se precisa construir *Galled Trees*, apenas identificar os conflitos entre as colunas para achar nós de recombinação.

Como não se precisa da estrutura das *Galled Trees*, que são redes filogenéticas sem ciclos de recombinação sobrepostos, pode-se deduzir que a

restrição do algoritmo de só funcionar com redes filogenéticas sem sobreposição deve cair.

A primeira adaptação feita no algoritmo é remover os procedimentos que apaga os nós recombinantes para testar a filogenia da árvore. Esse procedimento é o que testa se a árvore tem ciclos de recombinação sobrepostos, para cumprir o objetivo do trabalho não há a necessidade de remover esses nós e nem fazer o teste de filogenia.

A próxima mudança é remover os procedimentos que identificam as matrizes *pass-through*, essas matrizes ajudam a determinar o parentesco entre os nós da rede para saber em que ponto eles estarão posicionados na hora de montar a árvore.

A alteração seguinte é remover os procedimentos que fazem a montagem dos galhos e o posicionamento das folhas da árvore. Feito isso, teremos uma ferramenta que dado um conjunto de seqüências de entrada, identifica as colunas conflitantes e os ciclos de recombinação.

Número Máximo de Recombinações

A ferramenta construída até este ponto do trabalho lê um conjunto de seqüências e dá como resposta as colunas conflitantes e os ciclos de recombinação. Essas informações serão de grande importância na hora de determinar os limites do intervalo em que o número máximo de recombinações está presente e também no processo de verificação dos algoritmos desenvolvidos para esse fim.

O objetivo desta seção é fazer os cálculos do limite inferior e do limite superior do intervalo em que número máximo de recombinações possível está contido.

Limite Inferior

O limite inferior é o valor mínimo que o número máximo de recombinações poderá assumir.

Para se chegar nesse valor, será usado o algoritmo para a construção de *Galled Trees* com as alterações descritas acima. O algoritmo agora sofrerá algumas adaptações para calcular o limite inferior do número máximo de recombinações.

Para isso, deve-se antes pensar em qual seria o limite inferior para o número máximo de recombinações e como calculá-lo a partir da informação gerada até aqui.

O primeiro passo para fazer o cálculo do limite inferior é definir um valor que nunca poderá ser maior que o número máximo de recombinações, e, a partir desse valor, deve-se trabalhar para aumentá-lo o máximo possível a fim que se aproxime do número máximo de recombinações.

O número que será usado como valor inicial para fazer o cálculo do limite inferior será o número de recombinações que o algoritmo utilizado para a construção de *Galled Trees* achou. Esse algoritmo faz uma das inúmeras arrumações possíveis, logo, o número de recombinações achado por ele nunca será maior que o número máximo de recombinações e estará sempre com uma certa proximidade do valor que queremos, isso porque o valor que esse algoritmo calcula é aproximadamente a média do número de recombinações possível, ou seja, um valor que fica entre o número mínimo e o máximo possível. No entanto, apesar desse valor ser um bom limite inferior para os casos em que o número máximo de recombinações é próximo à média de recombinações possível, ele não é um bom limite

inferior para os grandes conjuntos com inúmeras possibilidades de arrumações e recombinações.

Devemos tentar aproximar o nosso limite inferior primário do número máximo de recombinações. Para isso, vamos pensar na estrutura da rede filogenética, e no que já foi implementado até aqui. Podemos então deduzir que um bom limite inferior seria o número de seqüências que só podem ser formadas fazendo-se o uso da recombinação. Essa hipótese leva em conta que para cada seqüência que precisa de uma recombinação para ser formada, foi feita apenas uma recombinação. Esse valor é mais alto que o limite inferior primário, pois o algoritmo utilizado até aqui contava uma recombinação por grupo de recombinação. A partir de agora ele irá contar uma recombinação por seqüência conflitante. Por exemplo, se um grupo tinha dois nós recombinantes, o algoritmo só contava uma recombinação, a partir de agora, serão contadas duas. Esta nova forma de calcular se aproxima bastante do número máximo de recombinações.

Com o algoritmo para a construção de *Galled Trees* pode-se facilmente identificar quais nós só podem ser formados apenas através de recombinações.

Para isso, deve-se primeiro identificar quais colunas entram em conflito entre si, após essa identificação, deve-se ver para cada par de colunas que entraram em conflito, quais são os nós de recombinação, ou seja, os nós que formam o par 1-1. O passo seguinte é varrer todos os pares de colunas conflitantes e identificar quais seqüências formam pares 1-1, essas seqüências são as que só podem ser formadas através da recombinação. A quantidade desse tipo de seqüências é o limite inferior do número máximo de recombinações (Ver Figura 9). Lembrando que o limite inferior considera

que para cada seqüência que só pode ser formada através de recombinação, é feita apenas uma única recombinação.

Este limite inferior em grande parte dos casos é muito próximo do número máximo de recombinações, pois na maioria das vezes é necessária apenas uma recombinação para formar as seqüências, e nos casos em que são necessárias mais de uma recombinação, este número tende a ser próximo de 1.

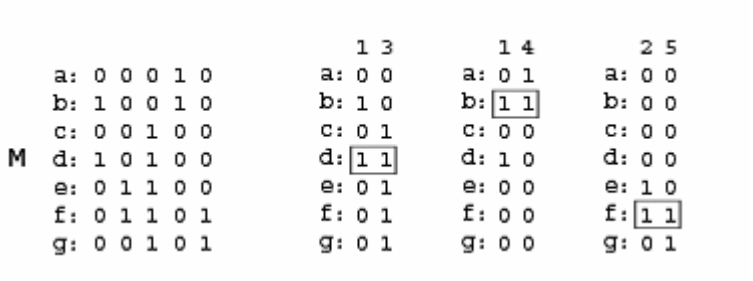

Figura 9: Seqüências que só podem ser formadas com recombinação

Como se pode ver, as seqüências b, d e f são as que formam os pares 1-1 nos conflitos entre as colunas, logo, elas só podem ser formadas através de recombinação. Sendo assim, o limite inferior do número máximo de recombinações para este exemplo será 3, que é o número de seqüências que formam o par 1-1 em colunas conflitantes.

Limite Superior

O limite superior é o valor máximo que o número máximo de recombinações poderá assumir.

Assim como o limite inferior, o seu cálculo será feito através de aproximações. Primeiro será imaginado um valor que com certeza é maior que o número máximo de recombinações e depois esse valor deverá ser diminuído para que se aproxime do valor desejado.

O valor inicial pensado para o limite superior foi o número total de mutações sofrido pelo conjunto de seqüências, levando-se em conta que todas as mutações sofridas pelas seqüências foram recombinações. No entanto, apesar desse valor parecer ser muito bom, já que estamos considerando que sempre que houve a possibilidade de ocorrer uma recombinação, ela aconteceu, ele não se aproxima muito do número máximo de recombinações, isso porque existem muitos casos em que ocorrem mutações sem a possibilidade de haver cruzamento de seqüências.

Devemos então tirar do cálculo os casos em que com certeza a mutação que ocorreu não foi uma recombinação. Os únicos casos em que se pode afirmar com certeza que a mutação ocorrida não foi uma recombinação são os casos em que a seqüência tem um único algarismo 1, isso porque para ter uma recombinação a seqüência precisa herdar no mínimo um algarismo 1 de cada seqüência que a originou.

O número de mutações sofridos por uma seqüência é o número de algarismos 1 presentes nela, mas qual seria o número máximo de recombinações necessárias para formar uma única seqüência se todas as mutações fossem do tipo recombinação? Esse valor seria o número de algarismos 1 presente na seqüência subtraído de 1. A explicação desse detalhe é simples, vamos supor que exista uma seqüência com cinco algarismos 1 e três 0s, numa primeira evolução, o ancestral comum, seqüência de oito 0s, sofreu uma mutação que não pode ter sido uma recombinação, porque ele não pode recombinar com nenhuma outra seqüência. A partir disso, a nossa seqüência de estudo já está com um algarismo 1 e sete 0s, agora sim ela pode recombinar, supondo que todas as outras mutações foram recombinações, essa seqüência terá passado por quatro recombinações para atingir sua forma atual (Ver Figura 10).

Sendo assim, o limite superior será o número de algarismos 1 das seqüências que só podem ser formadas com recombinação, ou seja, as seqüências que formam pares 1-1 nos conflitos de colunas, subtraído do número de seqüências que só podem ser formadas com recombinação, que é o limite inferior.

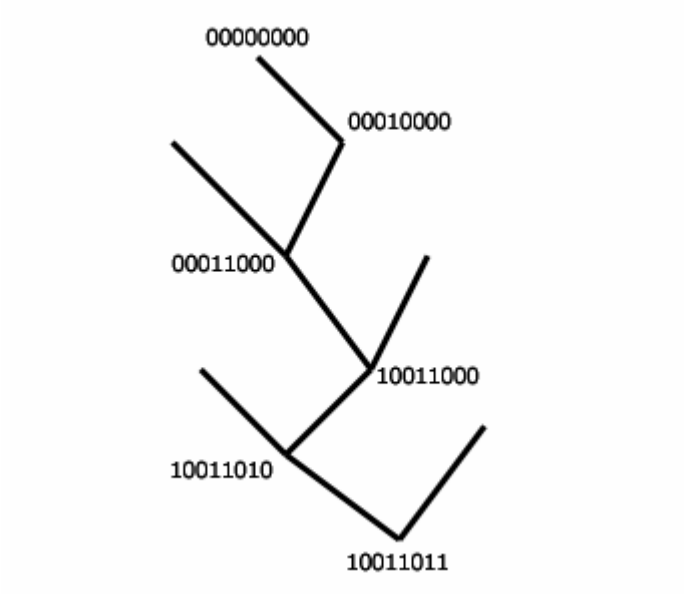

Figura 10: Recombinações que poderiam para formar uma seqüência

Pela figura podemos entender porque o número máximo de recombinações que pode formar uma seqüência é o número de algarismos 1 da seqüência subtraído de 1, porque o ancestral comum não tem com quem recombinar, logo, ele nunca participará de uma recombinação.

Esse valor do limite superior é muito bom e se aproxima bastante do número máximo de recombinações, principalmente para os grandes conjuntos de entradas, para os pequenos, o limite inferior continuará sendo mais próximo. Ao contrário do que se pode parecer, esse valor não será sempre o número máximo de recombinações, isso porque existem casos em que as recombinações necessárias para formar uma determinada seqüência

do conjunto de entradas formaram outras seqüências do conjunto no caminho. (Ver Figura 11).

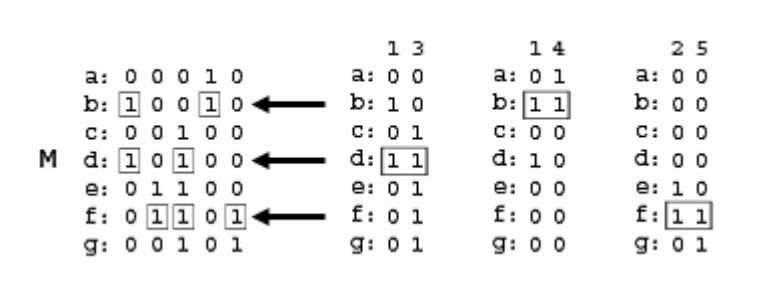

Figura 11: Cálculo do Limite Superior

Como se pode ver, as seqüências b, d e f são as que formam os pares 1-1 nos conflitos entre as colunas, contando-se o número de algarismos 1 presentes nelas, temos 7, subtraindo o limite inferior, ou seja, o número de seqüências que formam pares 1-1, que é 3, temos como limite superior para este exemplo 4.

Uma e Múltiplas Sobreposições de Ciclos de Recombinação

O algoritmo construído até aqui lê o conjunto de seqüências de entrada e partindo desses dados dá como resposta o limite inferior e o limite superior do número máximo de recombinações.

O próximo passo é checar se o que restringia o algoritmo a trabalhar apenas com redes filogenéticas sem ciclos de recombinação sobrepostos eram os procedimentos responsáveis pela construção da *Galled Tree* que foram removidos.

Podemos deduzir que sim, o único fator que restringia o algoritmo a trabalhar apenas com redes filogenéticas sem ciclos de recombinação sobrepostos eram os procedimentos removidos. A justificativa para isso é que *Galled Trees* são rede filogenéticas que assumem características de

árvores, pois tem ciclos de recombinação no meio de seus galhos, tornando o ciclo uma parte do galho, mantendo assim um pouco da estrutura e propriedades das árvores, logo, uma sobreposição nessa estrutura quebraria completamente a condição de ser *Galled Tree* (Ver Figura 12), mas não quebraria o conceito de recombinação, que é o que interessa a este trabalho. Uma recombinação será sempre uma recombinação, tendo ciclos sobrepostos ou não. Dessa forma, podemos concluir que o algoritmo construído até aqui não tem motivos para não funcionar com redes filogenéticas com ciclos sobrepostos.

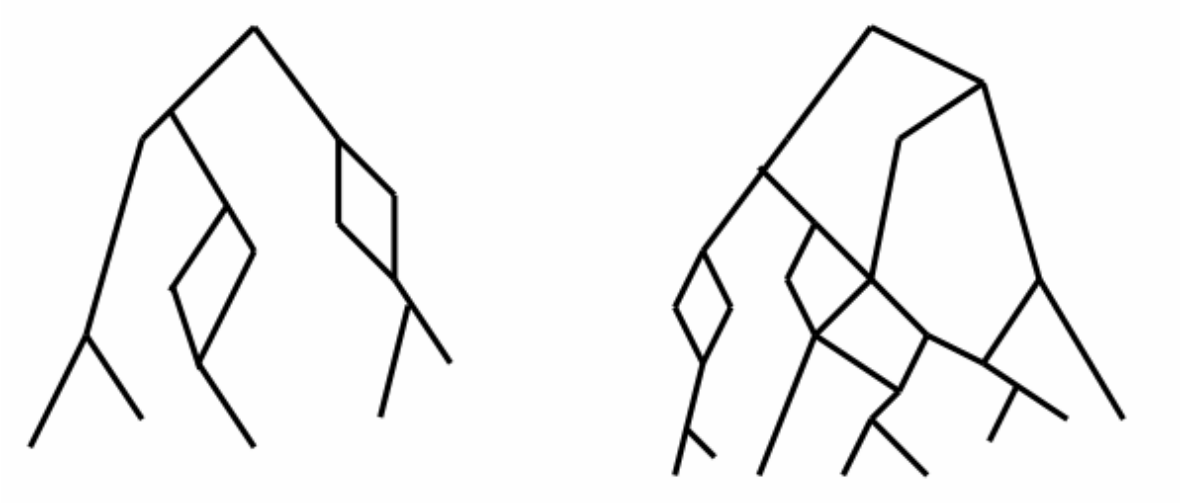

Figura 12: Galled Tree e Rede Filogenética com ciclos sobrepostos

Como se pode ver, a *Galled Tree*, estrutura da esquerda, é praticamente uma árvore, logo, para a sua construção, é importantíssimo que as seqüências do conjunto de entradas não formem ciclos de recombinação sobrepostos, como os da estrutura da direita, se não, seria impossível a construção de uma estrutura com propriedades de árvore.

A verificação da corretude de tudo o que foi apresentado até aqui será mostrada na etapa de Validação deste trabalho.

Validação

A etapa de validação é certamente uma das mais importantes num trabalho como este, que propõe um método para calcular um valor nunca antes calculado.

A validação é feita em duas partes, a criação dos casos de teste e os testes propriamente ditos.

Criação de Casos de Teste

Num trabalho como este em que o único meio de validar os experimentos realizados é através de testes, a escolha dos casos de teste passa a ser uma etapa muito importante do desenvolvimento do trabalho como um todo.

O objetivo desta seção é mostrar os critérios que foram o usados para desenvolver os tipos de casos de teste e os exemplos que foram gerados de acordo com cada critério e a sua importância para a comprovação das teorias fundamentadas neste trabalho.

Dada a grande importância dos testes para este trabalho, foram criados casos de teste simples e casos de teste complexos, ambos se completam e juntos tornam o resultado mais preciso.

Para os casos simples foram criados exemplos pequenos, dirigidos às pequenas peculiaridades dos algoritmos isoladamente. Enquanto para os casos complexos foram desenvolvidos conjuntos de seqüências que explorassem ao máximo não só as peculiaridades dos algoritmos

isoladamente, como interagindo em conjunto, tudo isso em conjuntos de entradas muito maiores.

Casos Simples

Os casos simples foram escolhidos tendo-se em mente as pequenas peculiaridades funcionais da ferramenta, ou seja, casos extremos que definam claramente um estado, são eles:

a. Um grupo de seqüências em que não existe a possibilidade de ser feita nenhuma recombinação, para esse grupo o algoritmo deve retornar sempre o valor 0 (zero), tanto como limite inferior como superior, dentro desse grupo foram criados alguns subgrupos, são eles: grupo com entradas distintas e sem o ancestral comum, grupo de seqüências com entradas repetidas, grupo de seqüências com apenas uma entrada, e grupo de seqüências em que aparece o ancestral comum;
b. Um grupo de seqüências em que possa ser feita apenas uma recombinação, para esse grupo o algoritmo deve retornar os valores mais próximos da realidade, e assim como no caso anterior, ele foi dividido em alguns subgrupos, que são: grupo com entradas distintas e sem o ancestral comum, grupo de seqüências com entradas repetidas e grupo de seqüências em que o ancestral comum está presente;
c. Um grupo de seqüências com várias recombinações, porém, sem ciclos de recombinação sobrepostos, esse grupo não será dividido em subgrupos, pois os exemplos anteriores são suficientes para testar o funcionamento das condições de conjuntos de entradas com seqüências repetidas e conjuntos de entrada em que aparece o ancestral comum;

d. Um grupo de seqüências com várias recombinações e que existe uma única sobreposição entre seus ciclos de recombinação, assim como o grupo anterior, pelo mesmo motivo, este grupo não será dividido em subgrupos;
e. Um grupo de seqüências com várias recombinações que formam ciclos de recombinação com múltiplas sobreposições.

Casos Complexos

Foram criados conjuntos de entrada muito grandes em que aparecessem várias das particularidades apresentadas acima simultaneamente. Esses exemplos são os que mais se aproximam dos casos reais.

Entre os exemplos testados aqui, foi usado um caso real, os genes de 11 indivíduos de 5 populações diferentes da espécie *Drosophila melanogaster*, a mosca da fruta, referentes à enzima álcool desidrogenase (ADh), pesquisados em 1983 por Kreitman [6]. Ignorando-se as inserções e deleções, que não são relevantes a este estudo, existem 43 polimorfismos, que foram transformados em 11 seqüências binárias de 43 caracteres cada, lembrando que cada polimorfismo é representado por um caractere binário da seqüência, onde o 1 expressa o ponto em que a seqüência sofreu a mutação naquele momento. A seqüência considerada como "ancestral comum" é formada pelas bases nitrogenadas mais comuns para cada lócus. Por exemplo, para o lócus número 7 do ADh da *Drosophila melanogaster*, a base mais comum é uma Timina, mas em uma das espécies havia uma Guanina nesse lócus, logo, na seqüência binária que representa essa espécie aparecerá um número 1 na posição 7 ao invés do número 0 que apareceria se lá tivesse uma Timina (Ver Tabelas 1 e 2).

	CCGCAATATGGGCGCTACCCCCGGAATCTCCACTAGACAGCCT
M1	--------AT--------TT-ACA-TAAC--------------
M2	--C---------------TT-ACA-TAAC--------------
M3	------------------------------------A----T-A
M4	----------------GT-----------------A--TA---
M5	---AG---A-TC--AGGT-------------------C------
M6	--C------------G--------------T-T-CAC----T-
M7	--C------------G-------------GTCTCC-C------
M8	TGCAG---A-TCG--G-------------GTCTCC-CG-----
M9	TGCAG---A-TCG--G-------------GTCTCC-CG-----
M10	TGCAG---A-TCG--G-------------GTCTCC-CG-----
M11	TGCAGGGGA----T-G----A---G----GTCTCC-C------

Tabela 1: Gene responsável pela ADh da Drosophila melanogaster

Na tabela está o sequenciamento do gene responsável pelo ADh da *Drosophila melanogaster* em 11 moscas recolhidas em 5 locais e épocas diferentes. A primeira coluna trás o código das moscas, a primeira linha mostra as bases nucléicas mais comuns para cada lócus deste gene, ela foi considerada o "ancestral comum", nos lócus em que aparecem "-" é porque a base nucléica é a mesma do ancestral comum. As inserções e deleções foram excluídas da tabela.

00000000110000000011011101111000000000000000
00100000000000000011011101111000000000000000
00000000000000000000000000000000000010000101
00000000000000001100000000000000000010011000
00011000101100111100000000000000000001000000
00100000000000010000000000000001010111000010
00100000000000010000000000000011111101000000
11111000101110010000000000000011111101100000
11111000101110010000000000000011111101100000
11111000101110010000000000000011111101100000
11111111100001010000100010000111111101000000

Tabela 2: Tabela criada a partir das seqüências da Tabela 1

A partir das seqüências da Tabela 1 é gerada esta matriz de seqüências binárias, onde o 1 representa uma mudança nas bases nucléicas e o 0 indica que no gene em questão a base é a mesma. Essa matriz é a que é usada como conjunto de entradas para o algoritmo descrito neste trabalho.

Testes

Neste trabalho foi demonstrado um algoritmo para calcular o limite inferior e o superior do número máximo de recombinações que pode existir numa rede filogenética. Por ser um algoritmo novo, para que seja comprovada a sua corretude ele precisa ser validado, essa validação se dará através de testes com inúmeros casos.

O objetivo desta seção do trabalho é mostrar os testes que foram realizados para fazer a validação do algoritmo. Para cada caso de teste serão mostrados a matriz com as seqüências do conjunto, o limite inferior e o limite superior dado como resposta pelo algoritmo e o número máximo de recombinações. Também serão feitas algumas observações referentes aos casos testados.

Conjunto formado por uma entrada sem recombinações

00001

Limite Inferior: 0
Limite Superior: 0
Número Máximo de Recombinações: 0

O algoritmo respondeu como era esperado, retornando 0 tanto para o limite inferior como para o limite superior.

Conjunto com entradas repetidas e sem recombinações

0001 1000 0001 0100

Limite Inferior: 0
Limite Superior: 0
Número Máximo de Recombinações: 0

Como no caso anterior, o algoritmo respondeu como era esperado.

Conjunto com o ancestral comum e sem recombinações

0001
1000
0000
0100

Limite Inferior: 0

Limite Superior: 0

Número Máximo de Recombinações: 0

O algoritmo mais uma vez funcionou perfeitamente.

Conjunto com entradas distintas e sem recombinações

0001
1000
0010
0100

Limite Inferior: 0

Limite Superior: 0

Número Máximo de Recombinações: 0

O algoritmo respondeu como era esperado.

Conjunto com entradas distintas e com apenas uma recombinação

0001
0100
1100
1000

Limite Inferior: 1

Limite Superior: 1

Número Máximo de Recombinações: 1

Tanto o limite inferior quanto o limite superior deram como resposta 1, que é o número máximo de recombinações.

Conjunto com algumas entradas repetidas e com apenas uma recombinação

00001
00100
01100
01100
01000

Limite Inferior: 1

Limite Superior: 1

Número Máximo de Recombinações: 1

Mais uma vez o algoritmo deu a resposta correta.

Conjunto com o ancestral comum e com apenas uma recombinação

000001
000100
000000
010100
010000

Limite Inferior: 1

Limite Superior: 1

Número Máximo de Recombinações: 1

O algoritmo mais uma vez acertou.

Conjunto com várias recombinações e sem sobreposições

10000
01000
10001
00010
00001
01010

Limite Inferior: 2

Limite Superior: 2

Número Máximo de Recombinações: 2

A precisão do algoritmo para os casos mais simples está muito grande.

Conjunto com várias recombinações e uma sobreposição

0010
0011
0100
0111

Limite Inferior: 2

Limite Superior: 3

Número Máximo de Recombinações: 2

Neste exemplo o número máximo de recombinações foi igual ao limite inferior e menor que o limite superior.

Conjunto com várias recombinações e mais de uma sobreposição

1011
0001
0101
0111
0011

Limite Inferior: 4

Limite Superior: 6

Número Máximo de Recombinações: 4

Mais uma vez o número de recombinações foi igual ao limite inferior.

Conjunto com várias recombinações e várias sobreposições

0101001
0111001
0100001
0111101
0100011
0011011
0011010
1111011
1111010

Limite Inferior: 9

Limite Superior: 26

Número Máximo de Recombinações: 23

Neste caso o número máximo de recombinações se afastou bastante do limite inferior e se aproximou do limite superior.

Conjunto com várias recombinações e várias sobreposições

000100
100100
001001
101000
011001
011011
001011

Limite Inferior: 6

Limite Superior: 10

Número Máximo de Recombinações: 9

O número de recombinações mais uma vez se aproximou bastante do limite superior.

Conjunto com várias recombinações e várias sobreposições

11010
11000
11100
01111
10111
00111

Limite Inferior: 6

Limite Superior: 13

Número Máximo de Recombinações: 7

O número máximo de recombinações se afastou do limite superior ficando muito próximo do limite inferior.

Conjunto com várias recombinações e várias sobreposições

1101
1100
1110
0111
1011
0011

Limite Inferior: 6

Limite Superior: 10

Número Máximo de Recombinações: 6

O número máximo de recombinações é o limite inferior.

ADh da Drosophila melanogaster

00000000110000000011011101111000000000000000
00100000000000000011011101111000000000000000
00000000000000000000000000000000000010000101
00000000000000001100000000000000000010011000
00011000101100111100000000000000000001000000
00100000000000010000000000000010101110000010
00100000000000010000000000000011111101000000
11111000101110010000000000000011111101100000
11111000101110010000000000000011111101100000
11111000101110010000000000000011111101100000
11111111100001010000100010000111111101000000

Limite Inferior: 9

Limite Superior: 85

Número Máximo de Recombinações: 83

Apesar do tamanho das seqüências o algoritmo funcionou perfeitamente, e o número máximo de recombinações se aproximou bastante do limite superior.

Conclusões

Após todos os testes e verificações feitos na seção de Validação, pode-se constatar que o método para cálculo dos limites inferior e superior do número máximo de recombinações, tendo como entrada um conjunto de seqüências binárias que representam as características polimórficas de um gene, espécie ou fragmento de DNA, é válido. O número máximo de recombinações nunca será menor que o limite inferior e nem maior que o limite superior. O método utilizado para cálculo faz uma boa aproximação tanto para casos em que o conjunto de entrada é formado por seqüências grandes como em exemplos de menor porte.

Pelos testes pudemos observar que para conjuntos de entradas em que as seqüências são pequenas o número máximo de recombinações se aproxima bastante do limite inferior, enquanto para conjuntos de entradas em que as seqüências são grandes, o número máximo de recombinações se fica próximo do limite superior.

O motivo para este fenômeno interessante é que costumam ocorrer poucas recombinações para formar as seqüências pequenas, enquanto nas seqüências grandes esse número tende a ser bem maior, já que são feitas mais mutações.

Dessa forma, como o limite inferior leva em conta que para cada seqüência que só pode ser formada através de recombinações é feita apenas uma recombinação, o número máximo de recombinações quando o conjunto de entradas é formado por seqüências pequenas se aproxima bastante dele,

já que as seqüências pequenas realmente passaram por poucas recombinações.

Por outro lado, o número máximo de recombinações nos conjuntos de entradas compostos por grandes seqüências se aproximará bastante do limite superior, já que este leva em conta que todas as mutações por que as seqüências que só podem ser formadas com no mínimo uma recombinação passaram foram recombinações, e na formação das grandes seqüências realmente acontecem muitas recombinações.

Foi constatado também que o fato da rede filogenética ter ciclos de recombinação com nenhuma, uma, ou várias sobreposições não influi no resultado do algoritmo.

Trabalhos Futuros

Este trabalho foi apenas o primeiro passo para outros estudos que podem explorar o número máximo de recombinações sofridas dado como entrada um conjunto de seqüências binárias representantes de características polimórficas de algum gene ou espécie.

Trabalhos futuros podem tentar usar o que foi exposto aqui para fazer uma aproximação ainda maior dos limites tomando como base o tamanho das seqüências do conjunto de entradas.

Assim como este trabalho tomou como base estudos para construção de *Galled Trees* e cálculo do número mínimo de recombinações presentes numa rede filogenética, este estudo poderá servir de base para outras pesquisas com redes filogenéticas e recombinações, uma área de extrema importância na ciência, que está começando a ser descoberta e desenvolvida recentemente e que tem pouquíssimos trabalhos publicados sobre ela.

Referências Bibliográficas

[1] D. Gusfield, S. Eddhu and C. Langley. Optimal, Efficient Reconstruction of Phylogenetic Networks with Constrained Recombination. Journal of Bioinformatics and Computational Biology, Vol. 2 no. 1, 2004.

[2] L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. Journal of Computational Biology, 2001.

[3] Song, Y.S. and J.J.Hein (2004) "Phylogenetics" by C. Semple and M.Steel (2002) Oxford University Press - In Press.

[4] Song, Y and J.J. Hein (2002) On the Minimum Number of Recombination Events in the Evolutionary History of DNA Sequences.

[5] Schierup, M. and J.Hein (2000): Consequences of Recombination on Traditional Phylogenetic Analysis. (Genetics 156.897-91).

[6] Kreitman, M.: Nucleotide Polymorphism at the Alcohol Dehydrogenase Locus of Drosophila Melanogaster. Nature 304, 412-417 (1983).

[7] D. Gusfield. Efficient algorithms for inferring evolutionary history. Networks, 1991.

[8] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. Mathematical Biosciences, 1990.

[9] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. Journal of Molecular Evolution, 1993.

[10] J. D. Kececioglu and D. Gusfield. Reconstructing a history of recombinations from a set of sequences. Discrete Applied Mathematics, 1998.

Assinaturas

Katia Silva Guimarães – Orientadora

Vinício Tavares de Melo Costa da Silva – Autor