Universidade de São Paulo
Instituto de Matemática e Estaística
Curso de Ciência da Computação

Samuel Gales Guimarães

SAGE Suite: Análise e desenvolvimento

São Paulo
2007

Samuel Gales Guimarães

SAGE Suite: Análise e desenvolvimento

Monografia apresentada ao Programa de Graduação em Ciência da Computação do Instituto de Matemática e Estaística da Universidade de São Paulo, como requisito parcial para a obtenção do título de BACHAREL em Ciência da Computação.

Supervisor: Alan Michell Durham

Doutor em Ciência da Computação

São Paulo

2007

Sumário

1 Introdução

Um dos maiores desafios da biologia moderna é descobrir o funcionamento de cada mecanismo constituinte de um ser vivo. Com isso torna-se possível entender cada etapa dos processos que regem a cada organismo, podendo assim analisar e manipular estes processos. Com este conhecimento podem-se entender males e criar curas e tratamentos mais específicos e eficientes, de acordo com cada problema.

Para entender o funcionamento dos seres vivos em escala celular deve-se entender como e quando o DNA, estrutura que codifica cada parte de um organismo, é transcrito e interpretado. Este processo é chamado de análise de expressão gênica. Com tal conhecimento é possível analisar como são sintetizadas as proteínas que controlam este funcionamento celular e assim saber como cada função é executada.

Atualmente existem diversos métodos empregados para a análise de expressão gênica. Dentre eles está a técnica de SAGE [8] (Serial Analysis of Gene Expression). Esta técnica permite a extração em larga escala de uma dada população de RNAs mensageiros. Com isto tem-se uma análise quantitativa e qualitativa destes. Com a geração de etiquetas para a identificação dos transcritos é possível constatar a manifestação do mesmo no dado organismo. Assim é possível fazer um trabalho conciso de forma rápida e econômica.

O Laboratório de Biologia Molecular de Coccídeas, localizado no Instituto de Ciências Biomédicas no departamento de Parasitologia, investiga aspectos da biologia molecular e genômica de protozoários dos gêneros Eiméria e Toxoplasma. O grupo de pesquisa do laboratório recentemente decidiu utilizar-se da técnica de LongSAGE [9], protocolo alternativo de SAGE, para analisar as diferenças entre as várias fazes de desenvolvimento do protozoário Eiméria Tenella, causador de uma coccidiose aviária. Como muitas técnicas empregadas, o SAGE necessita de uma análise computacional cuidadosa antes que os dados possam ser avaliados.

1.1 Estudo Preparatório

Durante os primeiros meses de trabalho do aluno no laboratório, foram estudados os conceitos básicos da biologia molecular, pois sem tais conhecimentos o trabalho na área de bioinformática seria inviável. Os estudos abordaram uma visão mais profunda da estrutura molecular do DNA, de seu funcionamento em um nível mais detalhado e os mecanismos básicos de transcrição e tradução destes em organismos procarióticos. Também foram vistas as formas mais básicas de estruturas secundárias do RNA e seus desdobramentos no funcionamento quanto à síntese de proteínas. O sistema de tradução e síntese de proteínas também foi visto em detalhes[1].

Após ter-se adquirido um pouco de familiaridade com o assunto, a área de bioinformática propriamente dita e as soluções computacionais foram abordadas, estudando as técnicas mais comuns e seus principais pontos e particularidades. Assim, com cerca de seis meses de estudo pode-se criar uma base sólida de conhecimento para o trabalho no laboratório com o contato direto com os biólogos[5].

1.2 A Técnica

O protocolo de SAGE gera uma um perfil da expressão gênica de uma dada população de RNA mensageiros através da criação de etiquetas que identificam os mesmo. Estas etiquetas consistem de seqüências de dez pares de base de uma seção única de cada RNA. Com isso garantimos que a chance de dois RNAs terem a mesma etiqueta (tag), sem considerar viés algum na distribuição das bases, é de uma em 1.048.576 (4^{10}). Considerando um genoma como o humano, com 25.000 RNAs expressos conhecidos. Apesar do tamanho paraecer satisfatório na unicidade das tags, ainda se tem dificuldades na identifiicação correta de cada RNA no genoma, devido ao grande número de colisões. Assim considerando erros de seqüenciamento e imprecisões da técnica este número se mostra relativamente pequeno. Por isso o grupo de pesquisa optou por utilizar um protocolo alternativo de SAGE, o LongSAGE.

Este protocolo, uma variação do SAGE tradicional, é muito semelhante, tendo como principal diferença o fato de que as tags geradas possuem 18 pares de base. Assim temos a probabilidade de colisão de dois RNAs de um em 68.719.476.736, número mais que satisfatório para qualquer expressão gênica conhecida atualmente. Este fato também

facilita a localização de cada ocorrência no genoma e uma maior flexibilidade quanto a erros que possam ocorrer.

As tags de SAGE/LongSAGE são criadas nas etapas a seguir, de acordo com o protocolo de LongSAGE[9]:

Isolamento do RNAm

Primeiramente o RNA mensageiro é isolado do resto do material gênico através de beads magnéticos. Um bead magnético é uma estrutura polarizada com cauda poli-T, uma pequena simples fita de DNA com uma seqüência de Timina. Como o RNA mensageiro, por razões de estabilidade, possui uma causa poli-A em sua extremidade, esta se liga à cauda poli-T do bead magnético. Usando um campo magnético podem-se atrair os beads que levam consigo as seqüências que estão unidas ao mesmo.

Uma vez que as seqüências estão isoladas, através de outra enzima, elas são completadas em cDNA, tendo assim fita dupla, como o DNA. Com isso podem ser utilizadas enzimas que necessitam da fita dupla para a identificação de sítios de corte.

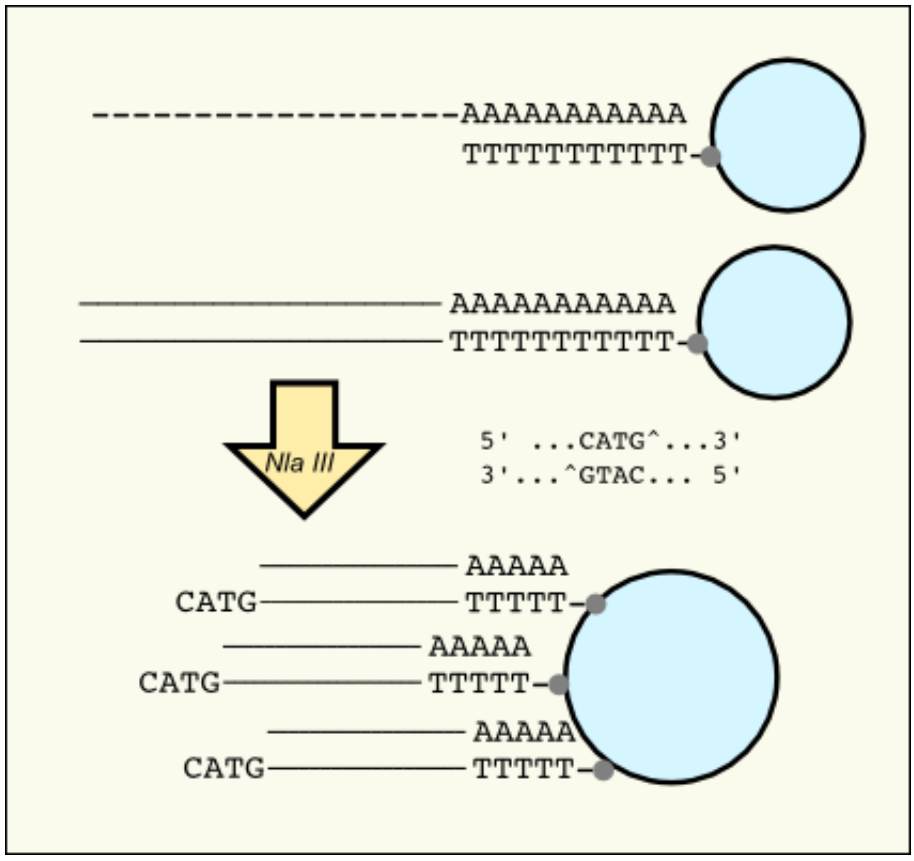

Figura 1.1: Isolamento do RNAm atravez de beads magnéticos e corte inicial

Formação das tags

Utilizando uma enzima de restrição, enzima que localiza uma seqüência especifica na fita e corta o ponto, neste caso a NlaIII, todos os sítios contendo a seqüência CATG são cortados. Utilizando o mesmo bead magnético o final do RNA, o ultimo trecho que contém uma seqüência CATG é isolado do resto.

Neste ponto as tags são separadas em duas alíquotas. Em cada uma um adaptador diferente é introduzido. Estes adaptadores se ligam ao sítio CATG na ponta da tag. Introduz-se então a enzima de restrição MmeI, que se une ao adaptador e secciona a seqüência 18 pares de base após o sítio CATG. Com isso temos seqüências com 18 pares de base cada, representando a última ocorrência de CATG no RNA, tendo assim uma identificação fixa para cada um deles.

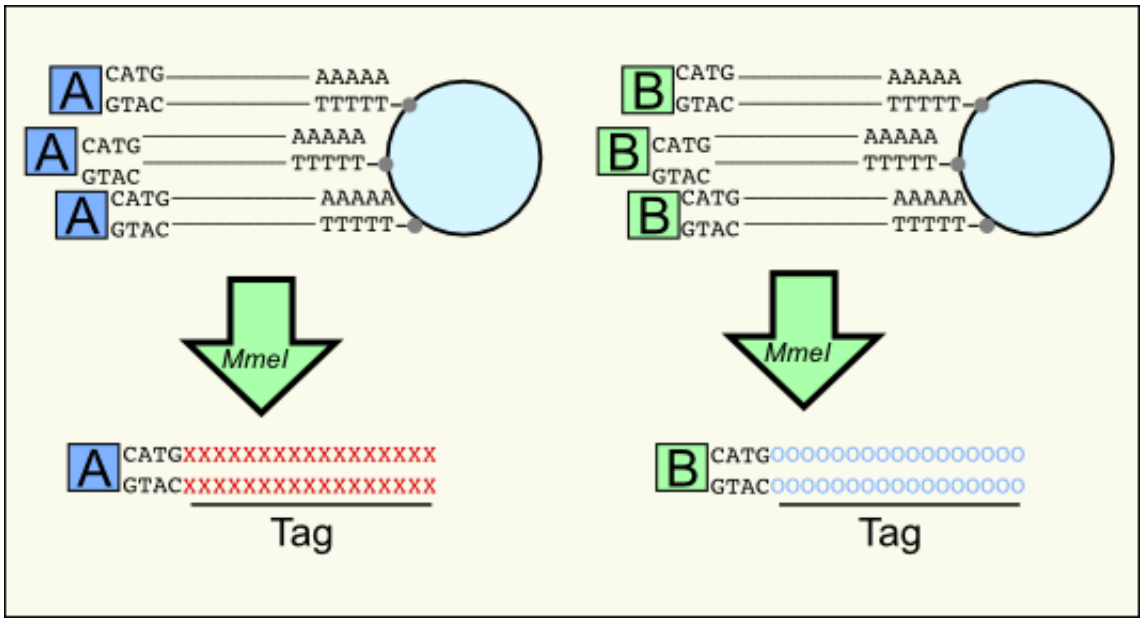

Figura 1.2: Inserção dos adaptadores nos dois conjuntos e corte final das tags

União em Ditags

As tags dos dois grupos são unidas duas a duas. Desta forma cria-se uma seqüência que contem um adaptador, a pontuação CATG, um ditag e depois, com código reverso, outra ditag, outro sítio de pontuação e o outro adaptador. Com os adaptadores garantimos que as tags não irão se ligar de forma espúria, mantendo a coesão entre o padrão pontuação seguida de tag. Utilizando os adaptadores das pontas da ditags, estas são amplificadas, clonadas, através do processo de PCR (Polymerase Chain Reaction). Desta forma é

possível trabalhar com as ditags, pois com um número pequeno de material gênico não é possível continuar a técnica.

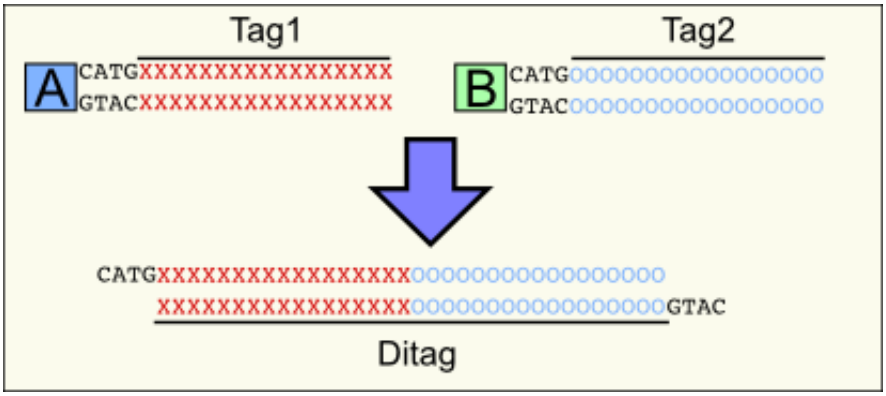

Figura 1.3: Formação das ditags a partir de tags de grupos diferentes. A tag do grupo B é invertida e depois de unidas os adaptadores são removidos

Finalização dos Concatâeros

Quando as ditags já estão formadas e flanqueadas pelos adaptadores, uma enzima é utilizada para que as os adaptadores sejam retirados e as ditags se unam em grandes seqüências Com isto temos seqüências grandes o suficiente para poderem ser seqüenciadas, um seqüenciador atualmente exige, em média, um mínimo de 100 pb (pares de base) para poder ser seqüenciado.

Com isso temos seqüências grandes que contém tags duas a duas separadas por sítios de pontuação conhecidos, CATG, logo podemos recuperar os dados gerados através de uma análise computacional dos resultados. Uma vez que as tags foram devidamente extraídas pode-se identificá-las e verificar no genoma do organismo sua funcionalidade e utilidade. Desta forma podemos comparar o funcionamento de diversos organismos ou mesmo vários tecidos ou estágios de um mesmo.

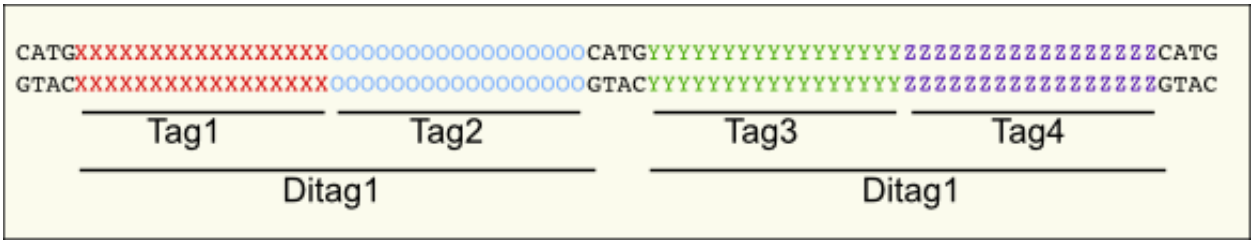

Figura 1.4: Exemplo abstrato de concatâmero contendo as tags e os sítios de pontuação

2 Conceitos e Resoluções

A análise de dados gerados pelo protocolo de SAGE poderia ser facilmente feito com os programas disponíveis para o mesmo. Entretanto, o grupo de pesquisa encontrou grande dificuldades na utilizadção dos programas disponíveis. Estes se mostraram insatisfatórios, apresentando interfaces problemáticas e pouco intuitivas. Mesmo depois de superar, ainda que parcialmente, as dificuldades encontradas e poder-se utilizar os programas para a análise, os resultados se mostraram inconsistentes e pouco confiáveis. Como os extratores de tags disponíveis possuíam código fechado, com licença acadêmica, ou no caso de licença aberta, extremamente mal documentada, não havia como saber como os mesmo processavam os dados, muito menos validar os métodos. Tendo estes fatos em vista o grupo de pesquisa resolveu criar seu próprio conjunto de programas para tratar da extração das tags.

Para desenvolver um conjunto de programas confiável foi necessário pesquisar mais profundamente o problema e verificar os dados obtidos. Estes resultados foram gerados pelo trabalho conjunto de todos os membros do grupo de pesquisa. Este é constituído pelos professores doutores Arthur Gruber e Alda Maria Backx Noronha Madeira e da doutoranda Jeniffer Novaes Gonçalves Dias, ambos do Instituto de Ciências Biomédicas da Universidade de São Paulo (ICB - USP). Os professores doutores Carlos Alberto de Bragança Pereira, do setor de estatística e Alan Michel Durham, do setor de computação, ambos do Instituto de Matemática e Estatística da Universidade de São Paulo (IME - USP), além do autor deste trabalho.

2.1 Inconsistência no Tamanho das Ditags

De acordo com o protocolo de LongSAGE as tags formada deveriam ter 18 pares de base de comprimento, logo seria esperado encontrar ditags de 36 pares de base (18 * 2). No entanto ao se realizar a extração das ditags dos concatâmeros, de forma bastante simples e pragmática, verificou-se que as ditags mais freqüentes eram as de 32, 33 e 34 pares de base, constituindo 35%, 48% e 15% do total de ditags extraídas, respectivamente. é de se

esperar a ocorrência de ditags com tamanhos diversos, devido à contaminação da amostra ao longo do processo de SAGE, mas o encontrado foi uma quantidade ínfima de ditags do tamanho esperado e uma quantidade muito expressiva de tamanhos inesperados, isso levou o grupo a procurar entender a causa de tal fato.

Primeiramente foi descoberto que a enzima de LongSAGE, MmeI, ao seccionar o cDNA para formar as tags não realizava um corte igual em ambas as fitas do DNA, mas deixa uma diferença de duas bases na extremidade. Com isso ao unir as tags em ditags não teríamos mais ditags com o dobro do tamanho das tags, como ocorre no SAGE tradicional, mas sim o dobro exceto por duas bases, se são compartilhadas por ambas tags. Ainda assim era esperado encontrar apenas, ou uma quantidade majoritária, de ditags de 34 pares de base, não explicando a grande ocorrência de ditags de 32 e 33 pares de base.

Procurando em publicações de pesquisas envolvendo LongSAGE constatou-se que algumas dessas relatavam que as tags geradas possuíam um comprimento de 17 pares de base, enquanto outras afirmavam que este número era na verdade 18 pares. Diante deste impasse e com um pouco mais de pesquisa foi encontrado uma citação dizendo que a enzima MmeI seccionava a fita 20 ou 21 pares de base após a primeira base do sítio de reconhecimento CATG, ou seja, 17 ou 18 após o CATG completo. Desta forma seriam geradas tags com ambos os tamanhos. Não é incomum a existência de enzimas de restrição com certo grau de imprecisão, principalmente as que reconhecem locais muito extensos. Logo a conclusão mais lógica foi a de que a enzima MmeI na verdade gera tags de 17 e 18 pares de base, numa proporção de 60% de 17 e 40% de 18.

Com esta conclusão temos uma explicação coerente para os dados observados. Como 60% das tags possuem 17 pb e 40% possuem 18 pb, quando as ditags forem formadas teremos 36% com 32 pb [1] 16% com 34 pb [2] e finalmente 48% com 33 pb [3]. Com isto tivemos um fato conclusivo sobre o tamanho das ditags e das tags contidas nas mesmas. Este fato é de suma importância durante a extração das tags, se estes fatos não fossem considerados certamente os resultados da análise seriam falsos e inconsistentes.

[1] 17 pb + 17 pb - 2 pb = 32 pb, $60\% * 60\% = 36\%$

[2] 18 pb + 18 pb - 2 pb = 34 pb, $40\% * 40\% = 16\%$

[3] 17 pb + 18 pb - 2pb = 33 pb, $2 * (40\% * 60\%) = 48\%$

2.2 Tratamento de tags de tamanhos variados

Uma vez que foi descoberto que a técnica de LongSAGE gera tags de dois tamanhos criou-se uma nova dúvida, como tratar tags conflitantes. Um caso que deveria ser estudado e tratado é o fato de que duas tags de 18 pb poderiam diferir entre si por apenas a última base. Assim quando fossem geradas tags de 17 pares teríamos um conflito, pois uma mesma tag seria correspondente a duas tags de tamanho maior.

Para recuperar o maior número de dados possível, seria mais interessante manter as tags de tamanho maior, uma vez que estas representam mais especificamente o objeto de estudo. Então, para eliminar a redundância foi decidido que distribuir a contagem das tags menores proporcionalmente entre as contagens das tags equivalentes. Esta seria uma forma de aproximar o número de tags totais tendo todas com 18 pares de base.

Tendo decidido isso foi verificado nos dados coletados qual a proporção de colisões encontradas. Ao analisar os dados verificou-se que uma quantidade muito pequena de tags colidia com tags maiores e mesmo nas que acontecia a colisão o número de tags de 18 pb era muito pequeno, inexpressivo. Diante desta situação não havia muito que fazer quanto a manter as tags maiores. Logo foi decidido que todas as tags seriam diminuídas para 17 pares de base. Assim temos uma base de dados homogênea, e, apesar da pequena perda de dados, confiável.

2.3 Heurística da Remoção de Ditags Repetidas

Após a faze de clonagem das ditags, durante o seqüenciamento, pode-se seqüenciar uma ditags mais de uma vez, pois pode ocorrer de um clone de uma ditag ocorrer mais de uma vez. Muitos trabalhos realizados sobre SAGE e LongSAGE se baseiam numa heurística para a correção do problema. Num genoma como o humano, onde existem cerca de 25.000 genes e a maioria deles é expressa em qualquer momento numa célula comum, a chance de, ao acaso, duas tags se ligar numa ditag igual duas vezes numa biblioteca de cerca de 60.000 tags, um tamanho razoável, é mínima. Portanto parece razoável à primeira vista eliminar a ocorrência múltipla de ditags.

Entretanto, em organismos como a Eiméria, onde existem cerca de 4000 genes, sendo que apenas 30% a 40% é expresso por vez, a chance de uma ditag ser repetida e

razoavelmente maior. Neste caso a remoção de ditags repetidas não parece tão razoável. Portanto decidiu-se não eliminar ditags e aproveitar todo dado que se tem. De fato, algum tempo depois dessa decisão ser tomada um artigo foi publicado criticando a remoção de ditags repetidas, dando uma base maior para a decisão tomada.

3 Atividades realizadas:

Em paralelo ao estudo, análise e discussão do projeto estavam em desenvolvimento os softwares de análise. O grupo já havia desenvolvido um pequeno conjunto para a extração de tags, com este foi possível fazer todo o estudo em cima da técnica e problemas encontrados. Também foi necessária a criação de um conjunto de dados de teste para a validação do programa feito. Para gerar dados concisos e mais rapidamente foi criado um conjunto de programas simuladores de SAGE, com isso pode-se validar o extrator e as técnicas utilizadas.

Devido à facilidade do tratamento de strings da linguagem Perl, esta foi escolhida para o desenvolvimento de todo o sistema. Alem deste fato a característica multiplataforma desta linguagem também pesou muito na escolha, pois o objetivo final e ter um conjunto de programas completo para o uso em laboratórios em geral. Desta forma os programas podem ser facilmente portados de um sistema para outro.

3.1 SAGE Analysis

Para a extração de tags o grupo desenvolveu um conjunto de programas, mais tarde nomeado como SAGE Analysis. O conjunto consiste de dois programas principais, um para a análise dos concatâmeros, identificação dos sítios de pontuação e extração das ditags. A segunda parte do conjunto analisa as ditags e extrai as tags das mesmas, contanto as tags e gerando um arquivo de contagem[1].

Esta estrutura foi criada tendo em vista desmontar passo a passo o processo biológico, desfazendo cada estágio. Desta forma é possível avaliar cada ponto da extração e modificar o funcionamento de alguma etapa com mais facilidade. Assim se houver a necessidade de adicionar alguma nova funcionalidade, esta poderia ser introduzida facilmente aos componentes.

Este programa já havia sido criado pelo grupo de pesquisa e era usado para a extração dos dados e analise a priori. Depois dos estudos realizados o conjunto teve

[1] O arquivo de contagem é da forma tag=contagem

que ser adaptado para se adequar aos novos conceitos, na verdade aos conceitos corretos. Para realizar a manutenção do sistema foi necessário a analise dos programas já criados e entendimento do código já escrito. Alem da correção dos problemas encontrados foi feita uma refatoração do código, tornando-o mais claro e objetivo, melhorando sua estrutura e clareza.

A analise de códigos prontos e o entendimento dos mesmos é uma tarefa que exige tempo e domínio da linguagem que está sendo tratada. Realizar manutenção de um sistema já pronto e em funcionamento é muito diferente do trabalho de modificar algo feito por si mesmo. O contato com os conceitos usados e o problema tratado, bem como o desenvolvedor é muito importante para agilizar o trabalho e garantir que o trabalho será realizado corretamente.

3.2 GenSuite

Uma vez que todos os problemas foram resolvidos e o programa extrator de tags foi criado e estava em funcionamento surgiu uma questão. Como garantir que o programa funciona corretamente e modela todos os problemas encontrados de forma eficiente e confiável? No desenvolvimento de softwares convencionais, com um domínio completo e conhecido, é fácil criar um conjunto de testes pontuais e guiar o desenvolvimento do projeto por eles. Em bioinformática não se conhece o domínio de trabalho, este não é controlado e é dificilmente bem mapeado, pois o principal objetivo do trabalho é conhecê-lo.

Com isso tem-se que criar um meio de validar o programa com dados conhecidos, assim pode-se comparar o resultado obtido com o esperado e avaliar a eficiência do programa. Precisão total no resultado não é exigida, nem ao menos esperada, em bioinformática, espera-se obter a informação mais precisa possível, mas ainda assim existe um grau de precisão difícil de ser medido, uma vez que não se tem o resultado "ideal"para comparação.

Tendo isso em mente foi decidido criar um programa simulador de SAGE. Desta forma podem-se criar dados controlados que, após a analise, se conhece o resultado e possibilita uma comparação direta e avaliação da eficiência e eficácia do programa. Um conjunto de programas foi criado, cada programa simulando uma etapa de SAGE, permitindo avaliar a capacidade de recuperação do analisador em cada etapa, facilitando

a melhoria e testes do mesmo.

A estrutura do funcionamento do gerador, em componentes, foi feita para ter a mesma função que a estrutura do SAGE Analysis tem sobre o protocolo de SAGE, desfazer cada etapa para permitir a avaliação. O GenSuite funciona na ordem reversa do SAGE Analysis, ou seja, seguindo o processo de SAGE propriamente dito. Desta forma cada etapa é gerada individualmente permitindo a valiação pontual da análise.

3.2.1 tag_count_generator.pl

Primeiramente deve-se gerar um arquivo como uma biblioteca resultante de SAGE. Assim este componente gera uma listagem de tags. Cada base é sorteada aleatoriamente, com distribuição uniforme, uma a uma. Depois para cada tag gera uma contagem aleatória também é gerada. E necessário tomar cuidado na geração da seqüência pois esta não pode conter a combinação CATG, pois a técnica não gera tais tags.

Organismos diferentes podem possuir um perfil de expreção muito diferente. O oragnismo mais conhecido e estudado, o humano, possui uma distribuição bastante uniforme dos genes expressados. Por este maior conhecimento e estudo, o genoma humano é usado como base de comparação para muitos trabalhos, entretanto existem organismos com expreções muito diferentes à humana, assim devemos permitir a geração de dados compatíveis aos mesmo.

Para a contagem é possível escolher uma distribuição diferente. Usando aproximações estatísticas foram implementadas as distribuições, normal[6], gamma[4] e Poisson, esta através de uma distribuição exponencial. Alem da distribuição uniforme. Para o sorteio das contagens pode-se instituir um mínimo e um máximo, limitando a contagem. Estes limites devem estar de acordo com os parâmetros da distribuição, ou podem inviabilizar a geração do numero dentro do desejado.

3.2.2 trim_tags.pl

A partir de uma contagem de tags, como a gerada por tag_count_generator.pl, o usuário pode escolher a porcentagem de tags que terão a ultima base cortada, gerando assim uma nova contagem. Com isso podese simular a imprecisão da enzima de corte utilizada na técnica de LongSAGE. Após a eliminação da última base das tags extraídas, novas

tags são formadas e estas são recontadas. O formato da saída é igual a do gerador de contagens, mas possui tags com tamanhos desiguais.

Desta forma este componente permite a modificação dos dados de contagens ideais formando um perfil mais próximo do real. Assim é possivel avaliar o comportamento dos programas de análise diante esta variaçã, verificando se protocolos alternativos de SAGE, como o LongSAGE são devidamente tratados.

3.2.3 counts2ditags.pl

Uma vez que já se tem as contagens de tags é necessário juntá-las em ditags. Essa união pode ser feita de dois modos, com corte em extremidade cega, como no SAGE convencional, ou usando extremidades coesivas, levando em consideração o final das tags para a união, como no LongSAGE. Para cada tag da contagem é dado um número aleatório. A partir deste número as tags são ordenadas e então unidas duas a duas em seqüência. No caso de haver extremidade coesiva, as tags são colocadas em grupos e unidas apenas se houver uma outra que satisfaça a condição de pareamento da extremidade. Caso não haja uma tag compatível ela é inserida numa pilha, que é desempilhada quando ocorrer uma tag compatível.

3.2.4 ditags2concatamers.pl

Com este ultimo programa as ditags podem ser unidas em concatâmeros, assim já criando um conjunto de dados pronto para ser analisado por um analisador SAGE. Com a listagem das ditags são criados os concatâmeros através da união de duas ditags e o acréscimo do sitio de pontuação, CATG. Cada concatâmero é montado de acordo com o tamanho estipulado pelo usuário e o resultado é um arquivo no formato Multi-FASTA[10].

Neste estagio é possível adicionar erro ao sinal de pontuação, com uma taxa definida. Para gerar um sinal de pontuação se adiciona um erro simples, apenas trocando uma base do sinal por outra diferente. Um relatório indicando os erros e as localizações é gerado. Assim pode-se avaliar a capacidade de recuperação de erros do programa extrator de tags.

Este acrécimo de erro ao sinal de pontuação serve para avaliar a capacidade de recuperção dos analisadores, pois no processo de seqüenciamento dos concatâmeros podem

haver erros. Caso este erro coincida no sinal de pontuação este será perdido, levando consigo as ditags vizinha, causando a perda de quatro tags. Desta forma é importante que os programas extratores levem este erro em consideração e modelem corretamente a recuperação, com o intúito de não introduzir erro à amostra.

3.3 Interfaces

Todos os programas desenvolvidos rodavam em modo texto, através de linha de comando. Mesmo possuindo instruções e uma lista de comando facilmente acessível, ainda era um pouco trabalhoso mudar os parâmetros e redirecionar os arquivos de saída e entrada. Como o intuito do conjunto é disponibilizar algo simples e fácil para a comunidade de biológica que trata de SAGE, era de se esperar que este possuísse uma interface mais agradável e intuitiva. Tendo isto em vista foi criada uma interface gráfica, com o pacote Tk[11] para os programas desenvolvidos.

Foi terminado o desenvolvimento da interface gráfica para o pacote GenSuite. A interface unifica os quatro módulos e forma de abas, facilitando o acesso. Cada um dos parâmetros está disponível para o usuário para ser modificado como quiser. A geração dos dados é executada com apenas um clique, podendo ser repetida facilmente com facilidade para modificações. O resultado é apresentado imediatamente na tela permitindo ao usuário avaliar a até mesmo alterar algo a sua escolha.

Está ainda em desenvolvimento a interface gráfica para o analisador, SAGE Analysis. A interface será de fora similar ao do GenSuite. Um módulo de analise estatística, estagio posterior a extração das tags, será incorporado. A analise estatística foge ao escopo deste trabalho e foi desenvolvida separadamente por outra seção do grupo de pesquisa e desenvolvimento.

Todos os pacotes, e as interfaces gráficas, dependem da instalação dos devidos interpretadores e pacotes no sistema do usuário. Como nem sempre é fácil adicionar e instalar pacotes pelo usuário está em avaliação a possibilidade do desenvolvimento de uma interface web. Através dessa o usuário poderíam submeter seus dados e receber a analise completa, sem a necessidade de qualquer instalação local. Neste serviço também sería incluso a requerimento para dados virtuais. A viabilidade deste sistema ainda está em avaliação, pois exige a disponibilização de memoria e processamento de um servidor para

abranger a demanda.

O grupo de pesquisa tem como objetivo no desenvolvimento deste pacote criar um conjunto de programas confiáveis e eficiêntes para o tratamento de SAGE. Disponibilizando assim, para a comunidade ciêntifica, ferramentas de boa qualidade equipada para o tratamento *in-silico* dos resultados de SAGE.

4 Resultados

Com o conjunto de geração de dados pronto tem-se em mãos a ferramenta para a validação do analisador. Com isso pôde-se verificar a eficiência e confiabilidade do mesmo. Além disso, é possível fazer uma comparação objetiva com outros analisadores e verificar qual o grau de confiabilidade e robustez destes.

Para esta tarefa foram criados vários casos de teste, um para cada combinação de características a avaliar. Primeiramente três testes principais foram executados. O primeiro com uma biblioteca padrão de SAGE, sem extremidade coesiva, corte seco, ou erro de qualquer forma. O segundo apenas com extremidade coesiva e sem erros nenhum e um ultimo com extremidade coesiva e tamanho de tags variados, com dezessete e dezoito pares. O analisador SAGE Analysis foi submetido a três testes de cada categoria[1], com tamanho progressivamente maior, para facilitar a identificação dos erros e fazer um tese mais robusto. O aproveitamento em todos os testes foi de 100%, todas as tags foram recuperadas com a contagem correta.

Após a bateria de testes foram realizados três testes exaustivos com uma biblioteca com 1.000.000 de tags com contagem variando entre 1 e 200 tags, gerando no total aproximadamente um cem milhões de ditags. Após passar pelo teste exaustivo, cada qual demorou cerca de vinte minutos para processar, o analisador teve um aproveitamento perfeito mais uma vez.

Desta forma garantimos que o analisador funciona sem falhas para o caso mais básico. Ainda não foram implementadas formas de recuperar sítios de pontuação de forma aceitável, por isto não foram gerados dados de teste com erros. Com o contínuo desenvolvimento do analisador novas funções de recuperação de dados serão utilizadas e com isto novos casos de teste devem ser criados.

Com estes conjuntos de dados pude-se avaliar também os programas de extração de tags que foram utilizados no começo do projeto. Até o momento apenas um conjunto de testes foi utilizado, o primário sem erros ou tags de tamanhos diferentes e ex-

[1] Os testes possuíam aproximadamente 2.000, 10.000 e 100.000 tags cada com contagens variando entre $[1 : 50]$.

tremidades coesivas. O analisador recuperou com precisão quase todas as tags mas com contagem errada na maioria delas. Obteve-se cerca de 85% de precisão, um resultado satisfatório para dados reais mas no caso simples é bastante decepcionante. Entretanto o programa não adicionou nenhuma tag errada ou extrapolou a contagem de nenhuma delas. Testes mais profundo serão realizados a fim de avaliar melhor o seu funcionamento, que no momento é qualificado apenas como aceitável.

5 Conclusão:

Depois de todo o estudo dos dados obtidos e de resultados anteriores na utilização do protocolo de SAGE o grupo possúi um entendimento pleno sobre a técnica e suas nuâncias. Somente com a descoberta e resolução dos problemas encontrados pode-se criar resultados corretos sem introduzir erros por falhas conceituais.

O conjunto todal de programas, incluíndo o analisador, SAGE Analysis, gerador, GenSuite e um módulo de análise estaística dos dados esta em faze final de desenvolvimento. Uma vez pronto e publicado o conjunto SAGE Suite estará disponível para uso na comunidade científica. Até a publicação o conjunto se encontra em desenvolvimento fechado. Quando publicado o conteúdo do pacote, assim como todas as ferramentas auxiliares utilizadas ao longo do processo estarão disponíveis no sítio do laboratório[1]

[1] http://www.coccidia.icb.usp.br/

Referências

[1] Alberts, B. ; Bray, D.; Lewis, J.; Raff, M.; Roberts, K. & Watson, J. D. Molecular Biology of the Cell. New York, Garland (1994)

[2] Dinel, S.; Bolduc, C.; Belleau, P.; Boivin, A.; Yoshioka, M.; Calvo, E.; Piedboeuf, B.; Snyder, E.E.; Labrie F.; St-Amand, J. Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. Nucleic Acids Res., **16**, 26, (2005)

[3] Emmersen, J.; Heidenblut, A. M.; Laursen A.; Hahn, S. A.; Welinder, K. G. & Nielsen, K. L. Discarding duplicate ditags in LongSAGE analysis may introduce significant error, BMC Bioinformatics, **8**, 92 (2007)

[4] Marglasia, G. Expressing a random variable in terms of uniform random variables. Ann. Math. Stat. 82, **894** (1961).

[5] Mount, D. W. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, New York (2001).

[6] Neumann, J. von, "Various techniques used in connection with random digits. Monte Carlo methods", Nat. Bureau Standards, **12**, pp. 36-38 (1951).

[7] Ojopi, E. P. D.; Oliveira, P. S. L.; Nunes, D. N.; Paquola, A.; DeMarco, R.; Gregório, S. P.; Aires K. A.; Menck, C. F. M.; Leite, L. C. C.; Almeida, S. V. & Dias-Neto, E. A quantitative view of the transcriptome of Schistosoma mansoni adult-worms using SAGE (2007).

[8] Velculesco, V.E.; Zhang, L.; Volgestein, B.; Kinzler, K.W. Serial analysis of gene expression, Science. **270**, 484 (1995).

[9] Zhang, Y. & Gilles, P. l-SAGE Long Kit for digital genome-wide expression profiles from smaller samples sizes. Focus, **93 2**, 11 (2003)

[10] `http://www.ncbi.nlm.nih.gov/blast/fasta.shtm`

[11] `http://www.perltk.org/`