MAC5701 - Tópicos em Ciência da Computação

Relatório Final

Análise Comparativa de Seqüências de Genoma

Aluno: Jesús Pascual Mena Chalco

Orientador: Prof. Dr. Roberto Marcondes Cesar Junior

São Paulo, 25 de Junho de 2004

Resumo

Na atualidade, existe grande interesse na pesquisa genética dada pela importância global de seus descobrimentos e conclusões obtidas na procura de um maior entendimento dos organismos biológicos. Um importante avanço da Bioiformática deve-se ao seqüenciamento automático de grandes volumes de fragmentos de seqüências de DNA com o objetivo de identificar regiões onde estão localizados os genes, que são as regiões codificadoras, com sua função ou descrição . Entretanto, este crescimento exponencial de informação biológica superior ao da capacidade de análise propôs um desafio na utilização e desenvolvimento de novos métodos para a anotação completa de todos os organismos com as análises comparativas genômicas entre diferentes organismos para a identificação de seqüências conservadas com potencial funcional e regulador.

Neste relatório, é apresentado brevemente algumas estratégias de sequenciamento, a estrutura de um pipeline bioinformático, e métodos para a análise comparativa de seqüências genômicas.

1 Introdução

O genoma de um organismo codifica é o encarregado de codificar o conjunto completo das instruções genéticas para o desenvolvimento do próprio organismo. Esta estrutura é uma seqüência linear de nucleotídeos.

A obtenção da seqüência genômica de um organismo é um trabalho não-trivial e involucra muitos procedimentos biológicos e computacionais. Nos procedimentos biológicos se utilizam varias estratégias. Os procedimentos computacionais são agrupados em um *pipeline bioinformático*[1], e existe atualmente um desenvolvido, como parte dos objetivos do plano de estudos apresentado, para o processo e montagem de genomas de cloroplasto de eucalipto do Laboratório de Biotecnologia do Departamento de Ciências Biológicas da ESALQ-USP.

Apesar da existência de métodos seguros para prever a localização e a expressão de genes e da enormidade de seqüências montadas pelos diversos projetos genomas, vários já concluídos e muitos em andamento, ainda existe muitas seqüências que não se encontram nos bancos de dados, como também ainda não contam com sua anotação completa.

Assim, as análises comparativas de seqüências de genomas de organismos são utilizados como uma parte principal na anotação do genoma para encontrar as partes funcionais das seqüências dos genomas, permitindo-nos o estudo da sua evolução, das funções dos genes e da herança das doenças para aplicar o conhecimento dos organismos estudados aos não-modelados[2].

Este relatório está organizado da seguinte maneira. A seção 2 apresenta brevemente duas estratégias de seqüenciamento de genoma. Na seção 3 está descrita os procedimentos computacionais de um *pipeline bioinformático*. Os métodos da análise comparativa estão

[1] *Bioinformatic pipeline*, entenda-se como um conjunto de operações computacionais organizadas não precisamente lineares.

descritas na seção 4, o método da análise de correspondências na seção 5, os bancos de dados de propósito geral para genômica comparativa na seção 6, e os problemas que influenciam a genômica comparativa na seção 7. Finalmente, a conclusão está descrita na seção 8.

2 Seqüenciamento de DNA

O seqüenciamento é um processo para determinar a ordem ou a estrutura dos nucleotídeos (blocos que constituem a molécula de DNA ou RNA) em uma amostra.

Em projetos genoma é utilizado o método de fragmentar todo o genoma de um organismo em pequenos pedaços para seguidamente seqüenciá-los, com ajuda de ferramentas computacionais pode-se montá-los e assim reconstruir a informação genômica do organismo inicial, pois atualmente é impossível seqüenciar diretamente blocos contínuos de milhares de pares de base (geralmente são usados fragmentos cujas quantidades de bases superem em 14 vezes a quantidade do genoma total).

Existem várias estratégias disponíveis que em alguma medida tem erros, apresentando vantagens e desvantagens, os quais podem ser amplamente revisados em [5].

Em projetos genoma de procariotos, geralmente cópias múltiplas do DNA inteiro são cortados em fragmentos pequenos por enzimas de restrição, mediante o método *shotgun*, e depois são clonados em vetores plasmidiais que são seqüenciados em suas extremidades. Após uma primeira etapa da montagem desse genoma, fragmentos maiores são clonados em cosmídeos e seqüenciados. Essa segunda etapa é importante para a montagem do genoma completo do organismo porque a primeira normalmente produz uma seqüência incompleta apresentando alguns buracos de seqüência conhecidos como *gaps*. Os fragmentos sobrepostos criados são a parte fundamental para a montagem (Figura 1).

Em projetos genoma de eucariotos, que possuem freqüentemente uma enorme quanti-

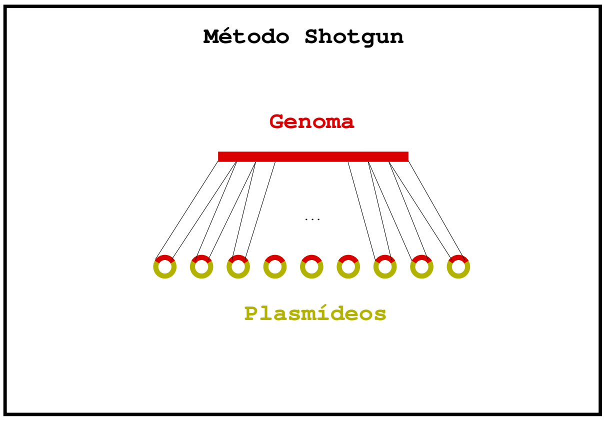

Figura 1: Método *shotgun* para organismos procariotos.

dade de DNA, atualmente é usado o método *shotgun hierárquico*. Aqui o DNA inteiro do organismo é primeiramente inserido em grandes vetores de clonagem, como cromossomos artificiais de bactérias (*Bacterial artificial chromosome* - BACs) ou leveduras (*Yeast artificial chromosome* - YACs).

Depois é realizado o *shotgun* desses grandes fragmentos dos vetores, gerando fragmentos menores que são clonados em vetores plasmidiais para o seqüenciamento. Assim, tais projetos consistem de duas etapas, a montagem de cada um dos fragmentos clonados nos BACs e YACs, e a montagem final que reunirá as seqüências completas dos BACs e YACs montados para a reconstituição da informação genômica inicial[8] (Figura 2).

Um fragmento pode ser único, parte de um gene ou uma das muitas seqüências repetidas. Diferentes partes do genoma são bastante similares na seqüência, e certamente essas regiões são as mais difíceis de montar.

O genoma deve ser seqüenciado nessa grande escala para garantissar estatisticamente a existência de redundância entre os fragmentos e reduzir as chances que se perca regiões no genoma.

Os resultados do seqüenciamento são conhecidos como fragmentos ou leituras que

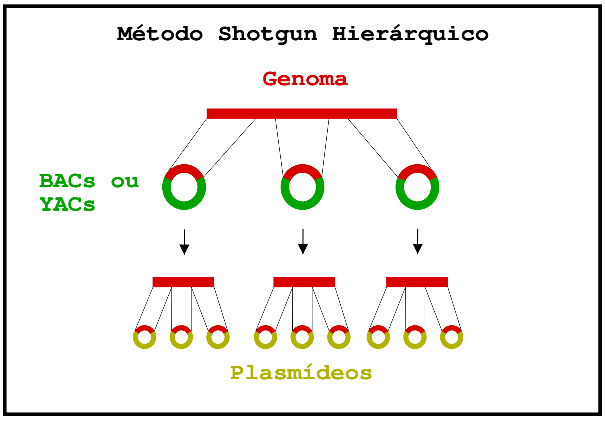

Figura 2: Método *shotgun hierárquico* para organismos eucariotos.

suscintamente é a estrutura a nível atômico do organismo seqüenciado. Conseqüentemente, um pipeline bioinformático deverá ser usado para determinar a longa seqüência original que será o genoma.

3 Pipeline Bioinformático

Para a montagem e anotação genômica de um organismo geralmente é adotado o método de seqüenciamento fragmentado em milhares de pequenos pedaços (e.g. 3.1 milhões de fragmentos para o genoma da Drosophila), como foi descrito na seção 2, para depois serem lidos por um equipamento de seqüenciamento automático.

Um *pipeline bioinformático* de dados de genomas de organismos pode ser organizado como um conjunto de etapas, computacionalmente custosos, mostrados na figura 3:

3.1 Etapa 1: Identificação das bases (Base-calling)

As entradas para o *pipeline* são as leituras (*reads*) dos fragmentos produzidos pelo seqüenciador automático de DNA com informações analógicas que representam as bases lidas

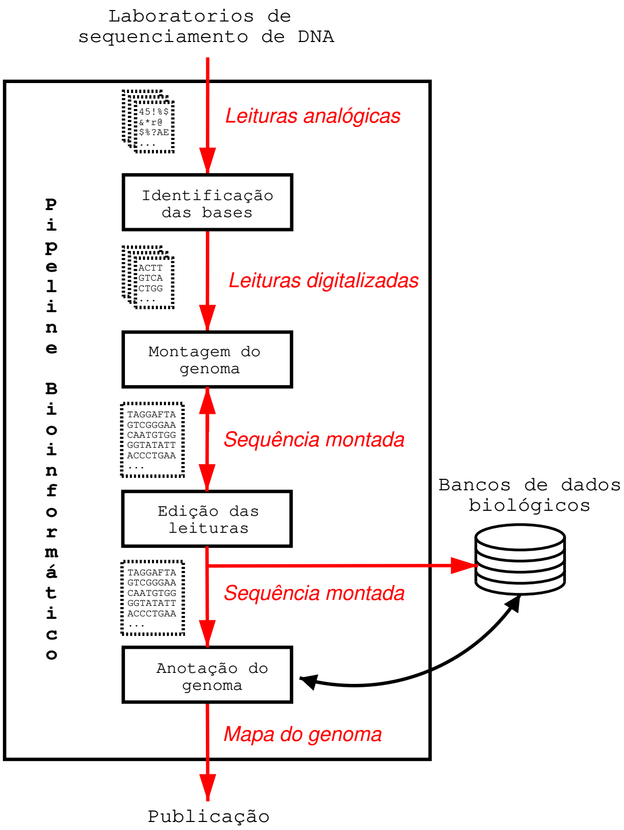

Figura 3: Esquema de um *pipeline bioinformático* para o procceso, montagem, edição e anotação genômica.

deste equipamento (*raw data*), chamados de arquivos de cromatograma, de um dos fragmentos próprios do método de seqüenciamento.

Para converter esses dados analógicos em fragmentos de bases são submetidas diretamente a um programa de identificação das bases, denominado *base-caller*, o qual os identifica como `A`, `C`, `G` ou `T`, correspondentes às bases Adenina, Citocina, Guanina e Timina, atribuindo um valor numérico de qualidade para cada um identificado, no caso contrario atribui o valor `N` como é mostrado na figura 4.

Aqui, a orientação de um fragmento é desconhecido e é freqüente que as máquinas de

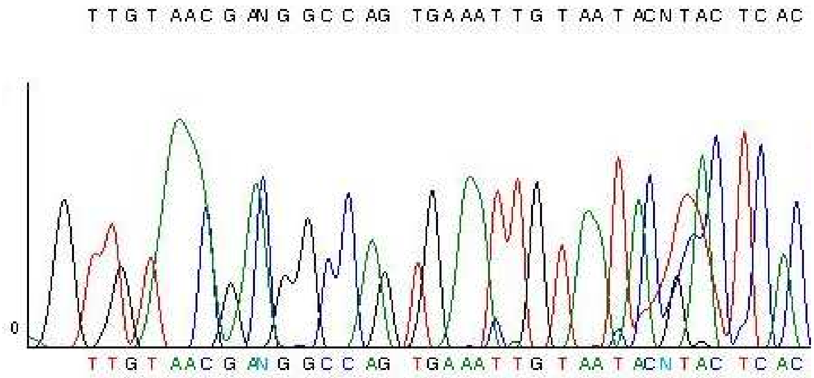

Figura 4: Identificação das bases desde a leitura de um fragmento arbitrário no *Base-caller*.

seqüenciamento automático tenham erros nas leituras dos fragmentos de DNA. Uma boa parte dos fragmentos contém erros numa taxa de 1-5%. Muitos mais erros ocorrem no final das leituras como é apresentado na figura 5.

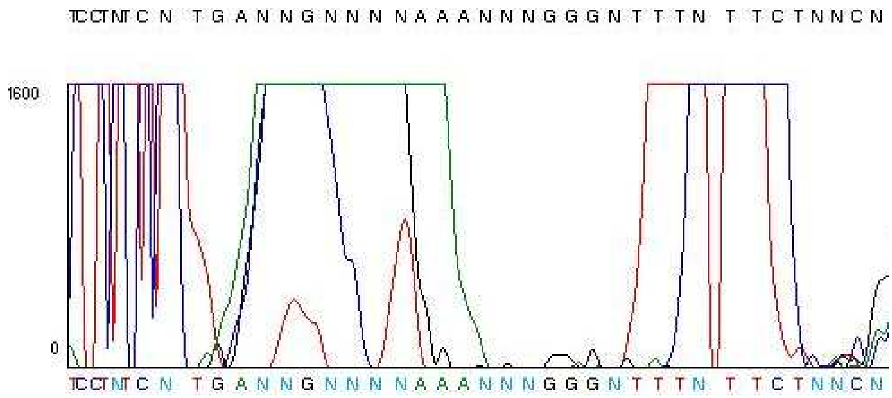

Figura 5: Erros no final das leituras depois da identifição de bases.

Nos trabalhos de Ewing[9, 12] as atribuições dos valores confiáveis aos nucleotídeos são dadas por um algoritmo que tem como base os métodos da análise de Fourier. O valor de qualidade de uma base em um fragmento é $q = -10 \log_{10} p$, onde p é a probabilidade de erro estimado para a base. Assim os valores de qualidade estão definidos para inteiros positivos pequenos para cobrir um amplio ranjo de probabilidades de erro. Para um valor de qualidade 10 corresponde a uma taxa de erro de 1 em 10, e o valor de qualidade 40 corresponde a uma taxa de erro de 1 entre 10000. Esse valor é importante para determinar si um fragmento precisa ser reseqüenciado.

O programa mais utilizado nesta etapa é o *Phred* que pode gerar arquivos de saída em diferentes formatos, sendo FASTA o formato mais utilizado com informação da identificação e os valores de qualidade de cada base.

3.2 Etapa 2: Montagem do genoma

Esta etapa consiste na geração da seqüência genômica desde os fragmentos já digitalizados. Os programas mais utilizados são o *Phrap*, *CAP3* e *Celera* baseados em algoritmos heurísticos e de programação dinâmica de alinhamento de seqüências que essencialmente procuram encontrar redundâncias entre fragmentos, agrupando os que tenham uma parte comum formando fragmentos maiores chamados consensos (*contigs*).

As operações gerais para a montagem, como é usado no *Celera*, são:

- **Mascaramento (*Screener*)**. São mascarados (marcados ou ocultados) com a letra `X` as bases das regiões contaminantes nos fragmentos produzidos. Regiões contaminantes são partes do fragmento que no representam o DNA que se deseja analisar. Estas regiões devem-se aos vetores de clonagem onde as seqüências de interesse foram inseridas ou pedaços de DNA adaptadores utilizados durante a construção das bibliotecas. A seqüência a mascarar deve com antecedência ser apresentada para que altere as bases de todas as similaridades com os fragmentos digitalizados e não influa nos processos posteriores.

- **Detecção de sobreposições (*Overlapper*)**. Cada fragmento é comparado com todos os fragmentos restantes para encontrar regiões sobrepostas entre eles (fragmentos onde o prefixo de um é o sufixo do outro).

A existência de sobreposições, devido ao método de seqüenciamento, é o fundamento da montagem, e é freqüente nos projetos de montagem genômico a definição previa do tamanho da sobreposição (e.g. 40 bp para o Projeto Genoma Humano).

Algumas dessas sobreposições são "verdadeiros" e outras induzidas por repetição (Figura 6). Nas sobreposições "verdadeiras" a seqüência compartilhada envolve fragmentos que são seções sobrepostas do genoma, os quais devem ir juntos. E nas induzidas por repetição a seqüência compartilhada envolve parte de uma repetição que ocorre em diferentes partes dispersas do genoma.

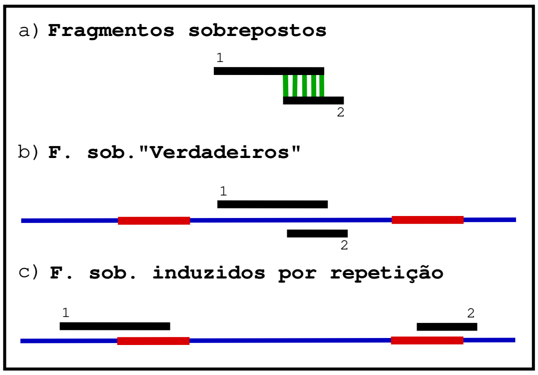

Figura 6: Suposições dos fragmentos sobrepostos.

Se fosse evidente conhecer qual sobreposição for "verdadeira", a montagem seria uma etapa trivial[7]. Para auxiliar esta detecção de sobreposições foram construídas varias ferramentas, e o *Basic Local Alignment Search Tool* (BLAST)[11] é a amplamente utilizada.

- **Agrupamento dos fragmentos (*Unitigger*).** Os fragmentos com regiões sobrepostas que conjuntamente formam uma única seqüência comum não-ambígua são agrupados e chamados de *unitigs* (Figura 7-a). (e.g. para a montagem da Drosophila cada *unitig* estava conformada por um promedio de 30 fragmentos).

Os fragmentos não-sobrepostos com regiões que disputam a seqüência comum são agrupados com seu valor de disputa (Figura 7-b).

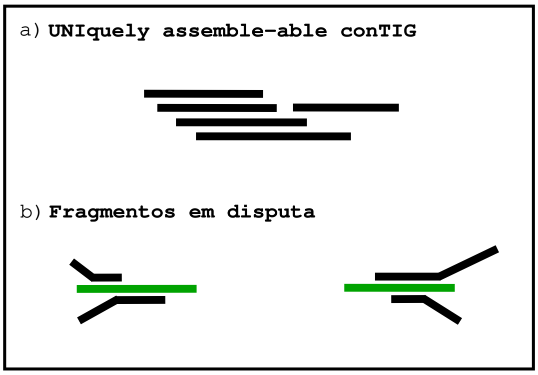

Figura 7: Agrupamento dos fragmentos.

Na montagem da Drosophila, o 99% dos *contigs* foram agrupados corretamente, mas uma pequena porcentagem devio-se a existência de DNA repetitivo no genoma do organismo.

O montador identifica *unitigs* incorretamente agrupados que conformam repetições da seqüência de DNA, com seu nível de "profundidade" do número total de fragmentos nos *unitigs*. Um modelo estatístico chamado de Discriminador é utilizado para encontrar pilhas de fragmentos com grande profundidade (Figura 8-b).

Os *unitigs* agrupados corretamente que conformam DNA único são chamados de *u-unitigs* (Figura 8-a). Os *unitigs* restantes são separados e utilizados para fechar as aberturas na seqüência. Assim, os *u-unitigs* são seqüências de DNA que estão prontas para ser ordenados no genoma.

- **Determinação de orientação e ordenação de *contigs* (*Scaffolder*).** No seqüenciamento muitos dos fragmentos foram agarrados do genoma em pares chamados

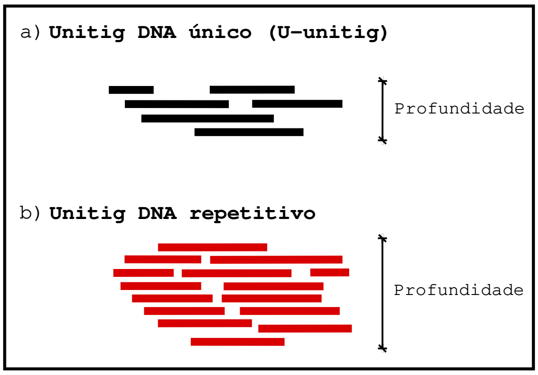

Figura 8: Profundidade dos agrupamentos de *unitigs*.

mate-pairs [2] estando sempre separados pela mesma quantidade de bases (e.g. 9-10Kb para o genoma da Drosophila), e como as repetições são menores a estas separações a utilização desses *mate-pairs* é uma maneira de levar as repetições.

Entretanto, nem sempre todos os *mate-pairs* são emparelhados numa dada distância devido a erros inevitáveis na manipulação computacional dos fragmentos.

Uma seqüência contígua de *unitigs* é um *contig*, e durante esta operação o montador orienta *contigs* usando os *mates-pairs* (Figura 9-a).

Se um par de um *mate-pair* está num diferente *contig*, os *contigs* provavelmente serão vizinhos o 99% das vezes. Se dois ou mais *mate-pairs* indicam a mesma direção, então os *contigs* envolvidos serão quase certas para ser vizinhos.

Com a comparação de todos os *mate-pairs*, a posição do *contig* torna-se aparente. Assim, conjuntos de *contigs* são ordenados e orientados usando esses *mate-pairs*, com respeito ao genoma, em uma estrutura chamada *scaffold* que é contínua exceto pelos buracos ou aberturas (*gaps*) (Figura 9-b).

Alguns desses *gaps* são devido às inevitáveis seqüências faltantes, como foram des-

[2] Um par das seqüências derivadas das duas extremidades de um único clone.

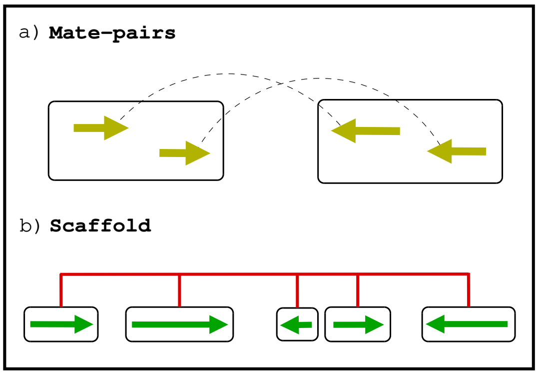

Figura 9: Scaffold.

critas na seção 2, ou seqüências repetitivas que agora podem ser fechadas usando os *contigs* que foram separados pelo Discriminador no agrupamento dos fragmentos.

- **Resolução das repetições (*Repeat rez*).** Onde se resolvem as repetições, que são seqüências que aparecem duas ou mais vezes em uma molécula de DNA, e certamente esta operação a tarefa mais difícil de toda a montagem do genoma.

  A classificação é feita de fragmentos repetidos por tamanho e confiabilidade, chamando os maiores e mais confiáveis de *rocks*, logo as *stones*, e finalmente as peças menores e menos confiáveis *pebbles*. *Rocks* são as primeiras para fechar os *gaps* e são enlaçados aos *contigs* sobre qualquer lado por dois ou mais *mate-pairs* (Figura 10-a). *Stones* são enlaçados aos *contigs* somente por um *mate-pair*, e sua posição em um *gap* é confirmada pelas sobreposições (Figura 10-b). *Pebbles* são posicionadas em um *gap* sobre a qualidade das sobreposições entre cada um e os *contigs* adjuntos (Figura 10-c).

Os genomas de diferentes organismos da mesma espécie não são idênticos devido ao polimorfismo entre eles (e.g. é estimado uma deferência em uma taxa de 1 em 1000 bp para o genoma de duas pessoas diferentes). Assim, as regiões no genoma com alta taxa

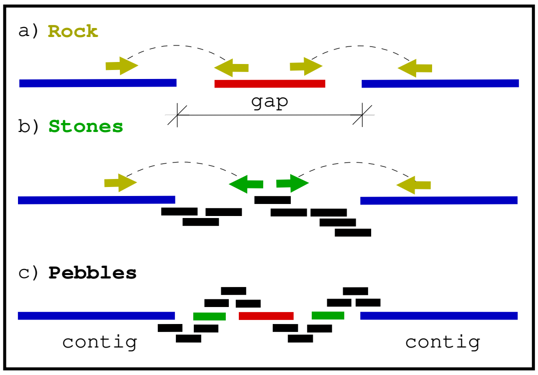

Figura 10: Classificação de fragmentos repetidos.

de polimorfismo são um problema neste etapa da montagem.

O seqüenciamento termina quando o aumento no número de fragmentos não contribui para o alongamento dos *contigs* resultando ainda regiões onde não se conhece a seqüência devido a contaminantes nas flutuações de clonagem, presença de seqüências repetidas, etc. Devem ser usados métodos biológicos alternativos para adicionar mais fragmentos, podendo também ser retirados do conjunto de entrada.

A montagem termina quando produz tantos *contigs* quanto *replicons*[3] do genoma e todas as bases tenham um valor de qualidade aceitável dependendo do valor estabelecido para o projeto.

Existem diferentes maneiras para certificar que uma seqüência genômica seja a correta. A seqüência pode ser comparada com pequenas partes de um genoma que haja sido seqüenciado e anotado previamente ou com regiões de mapas genômicos. Se uma montagem é consistente com regiões dispersas de informação conhecida, então é bastante provável que a seqüência toda seja a correta.

Entretanto, por muito valiosa que essa montagem seja, ela não é imediatamente útil si não estiver devidamente caracterizada em termos funcionais.

[3]Uma molécula de DNA que é capaz de se replicar e se manter.

3.3 Etapa 3: Edição das leituras

Essa etapa permite a correção de possíveis erros cometidos na identificação das bases e nos erros da própria montagem, com as operações de substituição, remoção ou inserção de bases. Em geral isso pode ser feito devido à existência de um ou mais fragmentos que tenham bases adequadas com qualidade alta para inferir a correção da base incorreta[10].

O programa mais utilizado para a edição das leituras e genomas é o *Consed*.

3.4 Etapa 4: Anotação do genoma

Essa etapa é o mais importante porque consiste em identificar regiões onde estão localizados, quais são as funções dos genes, e procurar um significado, descrição ou categorização biológica para a construção do seu mapa do genoma do organismo seqüenciado, sendo as operações comuns de:

- **Identificação dos genes**. Que pode ser automatizado com um processo de reconhecimento de genes em uma seqüência, ou seja, de seqüências codificadoras de proteínas *Open Ready Frames* (ORFs).

  Os programas mais utilizados aqui são o Glimmer, GenMark e GenScan.

- **Descrição de genes**. Essa é feita geralmente com ajuda de outras montagens de outros organismos cujas funções já foram conhecidas devido que seqüências de organismos estruturalmente similares devem ter também funções similares[3, 8] podendo ser esta tarefa semi-automática para a ajuda na classificação das novas ORFs.

- **Geração do mapa de genes**. Mostrando a disposição, orientação e categoria geral dos genes do organismo montado.

Os genes são regiões codificadores que expressam ou controlam uma proteína, ou algumas vezes, controlam o que é expresso por outros genes.

O principal banco de seqüências conhecido é o GenBank[6] construído pelo *National Center for Biotechnology Information* (NCBI).

4 Análise comparativa de seqüências de genoma

Na biología um resultado importante é que ao menos uma terceira parte dos genes de cada genoma dos organismos seqüenciados sua funções são conhecidas ou confiáveis, sendo unicamente possível uma predição funcional geral para os genes restantes.

Características comuns de dois organismos codificam dentro do DNA que é conservado entre essas espécies. As funções de varios genes pode ser predita mediante comparações de diferentes genomas, transferindo as anotações funcionais das proteínas dos organismos melhor estudados aos seus organismos ortólogos menos estudados. Assim, a análise comparativa de seqüências de genoma (*genômica comparativa*) ajuda na descoberta de relações entre genes e entre espécies, na compreensão dos mecanismos de evolução molecular e procura indícios de funções e interações dos genes, permitindo conseguir um melhor entendimento dos genomas e, conseqüentemente da biologia dos organismos respectivos.

Métodos estatísticos e combinatoriais são usados na genômica comparativa para modelar a evolução molecular mediante estruturas combinatoriais e processos estocásticos freqüentemente ajudado de simulações [4].

Os seguintes são métodos estatísticos e combinatoriais utilizados em genômica comparativa:

4.1 Medidas de distância evolucionária entre espécies

A distância evolucionária é utilizada para medir o tempo desde onde duas espécies divergiram (i. e. a idade do seu ultimo antepassado comum) unicamente mediante a comparação dos genomas das espécies contemporâneas, baseados na suposição da existência do “relógio molecular” dos organismos cuja taxa de mudanças evolucionárias é sempre constante.

Assim, a quantidade das mudanças relacionadas aos genes no nível das seqüências deverá ser proporcional ao tempo da sua divergência, e também a quantidade de mudanças do genoma total no nível do ordem dos genes.

Para estimar essa distâncias evolucionária do ordem dos genes pode ser medido a similaridade superficial contando o número dos *breakpoints* (onde o gene comum é seguido por dois genes diferentes nos dois genomas), ou calculado o número de eventos evolucionários necessário para explicar as diferenças no ordem dos genes.

4.2 Reconstrução da árvore da vida

A construção das árvores filogenéticos a partir dos dados genômicos é uma tarefa muito comum para o trabalho dos biólogos em genômica comparativa. Aqui é feito uma estimação das árvores baseados nos dados genômicos das diferentes espécies.

4.3 Detecção da quantidade de genes transferidos horizontalmente

A transferência de genes entre genomas é conhecido como *transferência horizontal* ou *transferência lateral*. Existem duas aproximações para a detecção dessa transferência:

- **Métodos filogenéticos**, envolvem comparações das árvores filogenéticas de diferentes genes do genoma. Para cada gene selecionado no genoma são coletados seus genes correspondentes (genes ortólogos) de outras espécies, e é reconstroido a filogenia baseada em similaridade das seqüências. Um gene transferido horizontalmente poderia manter significativamente uma árvore diferente ao dos outro genes.

- **Métodos paramétricos**, os quais não realizam comparações de genes entre organismos. É unicamente utilizado o contexto genômico atual e é feita a busca de genes que pareçam atípicos que são provavelmente transferidos horizontalmente.

4.4 Detecção das duplicações genômica ocorridas

É sabido que o maior recurso do novo material genético é devido à duplicação de partes do genoma, seguidos pela evolução divergente das copias dos organismos.

As duplicações quanto as eliminações, são ocorridas em todas as escalas (simples nucleotídeos, genes inteiros, regiões cromossomáticas, e até o genoma inteiro). Aqui é tentado inferir a historia dos organismos baseados nessas duplicações e eliminações genômicas ocorridas durante a evolução.

5 Análise de correspondências

Na genômica comparativa é fortemente utilizada a análise de correspondências como um método descritivo para análise global das funções dos genes. O análise de correspondências é um método poderoso para exploração multivariada de dados em grande escala, sendo aplicado em várias áreas de pesquisa, incluindo análise genômico.

Para a extração de informações relevantes dos dados, o análise de correspondência baseia-se na projeção de informação de alta dimensão sobre espaços de baixa dimensão. Tais projeções, no plano, permitem inspeção visual de tendências significantes, as quais são freqüentemente difíceis de enxergar em espaços de alta dimensão.

A dimensão dos espaços considerados são relevantes ao número de variáveis e observações envolvidas nas comparações. Neste método multivariado pode-se construir um sistema ortogonal chamado eixos fatoriais, correspondendo a projeções de baixo ordem sobre planos chamados planos fatoriais. Uma vantagem importante dessa construção é que as propriedades características das observações e as variáveis são apresentadas simultaneamente sobre planos fatoriais.

O método é chamado de "correspondência" entre a análise das observações e as variáveis. Nesta análise, cada eixo fatorial representa uma fração da informação total contida na matriz analisada. A significância estatística dessa fração determina a confiança relativa unida às observações e/o variáveis mostradas sobre o eixo correspondente. A ortogonalidade dos eixos fatoriais permitem a soma das suas correspondentes frações de informação.

Análises de correspondência permite também considerações de variáveis e observações simuladas (chamadas de variáveis e observações ilustrativas respectivamente), como variáveis adicionais que não contribuam à construção do espaço fatorial, mas pode ser mostrado como espaço fatorial. Assim, com uma representação como a descrita, é possível determinar a proximidade entre observações e variáveis, e variáveis ilustrativas e observações.

6 Bancos de dados de propósito geral para genômica comparativa

Atualmente existem bancos de dados disponíveis que oferecem aceso às mesmas seqüências dos genomas, mas devido à especialização dos grupos de pesquisa na análise de genoma são mantidos bancos de dados com informação como a organização dos operons, predição funcional, estrutura tridimensional, e reconstruções metabólicas.

Os principais bancos de dados com informação genômica para as análises comparativas são:

- **PEDANT** (Protein Extraction, Description, and ANalysis Tool), contém a coleção mais detalhada e exata dos genomas sistematicamente anotados e classificados disponíveis.
- **COGs** (Clusters of Orthologous Groups of proteins), que foram delineados mediante comparação das seqüências de proteínas nos genomas completos, representando os principais linhagens filogenéticos.
- **KEGG** (Kyoto Encyclopedia of Genes and Genomes), contém uma coleção detalhada das vias metabólicas de genomas completamente seqüenciados.
- **MBGD** (Microbial Genome Database), contém uma classificação de genes ortólogos utilizando as relações de similaridade entre genes de múltiplas genomas.
- **ERGO** , contém as ligações a informações sobre o rol funcional das enzimas.

Banco de dados	URL
PEDANT	http://pedant.gsf.de/
COGs	http://www.ncbi.nlm.nih.gov/COG/
KEGG	http://www.genome.ad.jp/kegg/
MBGD	http://mbgd.genome.ad.jp/
ERGO	http://www.ergo-light.com/

7 Problemas que influenciam a genômica comparativa

Inerentemente a anotação genômica está propenso a erros quando esse processo é automatizado. Esses erros podem ser minimizados mediante sistemas de anotação semi-automatizada e o treinamento apropriado dos anotadores.

Os problemas sistemáticos que influenciam a análise comparativa, como parte da anotação genômica, são [1]:

- **Propagação de erros e informação incompleta nos bancos de dados**, devido que a anotação errônea de uma proteína pode causar muitos erros quando usado para a anotação dos novos genomas, permitindo assim a amplificação dos erros.

- **Falsos-positivos e falsos-negativos nas buscas de bancos de dados**, devido aos critérios (geralmente estatísticos) de apresentação das respostas nas buscas por similaridade nos bancos de dados.

- **Genoma, proteína e o contexto do organismo como fonte de erros**, devido às más interpretações das informações funcionais dos organismos.

8 Conclusão

Na anotação genômica, um dos problemas de interesse biológico é descobrir padrões que possam revelar alguma similaridade ou ponto de mutação genética e polimorfismo (detecção de variações) para auxiliar os estudos de seus mecanismos de evolução molecular e relacionamentos filogenéticos.

Para tal, o conhecimento acerca do conteúdo e da organização de outras espécies é um importante ponto de partida na análise comparativa de seqüências de genoma, que com o aumento do número de seqüências genômicas completas disponíveis, essa será uma estratégia influente para a anotação genômica total.

Referências

[1] Andreas D Baxevanis and B F Francis Ouellette. *Bioinformatics: A practical guide to the analysis of genes and proteins*. Wiley interscience, 2001.

[2] Danila M Melloto-Passarin Tercilio Calsa-Junior and Helaine Carrer. Complete chloroplast dna sequence of sugarcane reveals similarity and diversity among saccharum species, sorghum, maize, rice and wheat. *Plant and Animal Genomes XII Conference*, 2004.

[3] Dirce Maria Carraro and João Paulo Kitajima. Sequenciamento e bioinformática de genomas bacterianos. *Biotecnologia Ciência e Desenvolvimento*, 28 - setembro/outubro, 2002.

[4] Kimmo Eriksson. Statistical and combinatorial aspects of comparative genomics. *Invited Lecture at the 19th Nordic Conference on Mathematical Statistic*, 2002.

[5] Bruce Alberts et al. *Molecular Biology of the Cell*. Garland Science, Englewood Cliffs, 2002.

[6] Dennis A Benson et al. Genbank. *Nucleic Acids Research*, 28(1):15-18, 2000.

[7] Eugene Myers et al. A whole-genome assembly of drosophila. *Science*, 287: 2196-2204, 2000.

[8] Francisco Prosdocimi et al. Bioinformática: Manual do usuário. *Biotecnologia Ciência e Desenvolvimento*, 29 - novembro/decembro, 2002.

[9] Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Research*, 8: 186-194, 1998.

[10] Vagner Katsumi Okura. Bioinformática de projetos genoma de bactérias. *IC, Universidade Estadual de Campinas*, Master thesis, 2003.

[11] Altschul SF Madden TL Schaffer AA Zhang J Zhang Z Miller W and Lipman DJ. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389-3402, 1997.

[12] Brent Ewing LaDeana Hillier Michael C. Wendl and Phil Green. Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Research*, 8: 175-185, 1998.