Capítulo 5. Anotação de genomas

Frederico Schmitt Kremer [1], Luciano da Silva Pinto [1]

[1] *Laboratório de Bioinformática e Proteômica, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas*

Objetivos: Apresentar os principais conceitos referentes à anotação de genomas, com ênfase para genomas microbianos.

Anotação de Genes e Genomas

A anotação de um genoma consiste na identificação e caracterização das regiões funcionais, o que pode incluir genes, promotores, terminadores, regiões de DNA repetitivo, operons, dentre outras, sendo estas usualmente denominadas *features*.A identificação das *features* pode ser realizada através do uso de dados experimentais, como alinhamento de sequencias de transcritos (ex: RNA-Seq, *Expressed Sequence Tags*) ou proteínas, ou com base em ferramentas de predição *ab initio*. Cada uma destas possíveis fontes de informação é denominada evidência, e também é possível se combinar diferentes evidências para a geração de uma anotação consenso. No caso de genomas eucarióticos, por exemplo, devido à complexidade da estrutura gênica, e a existência de fenômenos como o splicing alternativo, torna necessária a combinação de diferentes dados experimentais para a geração de uma anotação confiável. Por outro lado, a estrutura relativamente simples dos genes de procariotos permite que estes sejam identificados com boa acurácia, unicamente através de ferramentas *ab initio* .

Identificação de regiões codificantes

Diferença na estrutura gênica de procariotos e eucariotos

O processo de identificação das regiões codificantes em um genoma está direramenta atrelado as características da estrutura gênica do organismo de interesse. Desta forma, os algoritmos desenvolvidos para predição de genes em eucariotos não são aplicáveis para procaritos, e vice-versa.

No caso de procariotos, a estrutura gênica é relativamente, sendo a porção codificante do gene contida em uma única região. Por outro, no caso de eucariotos, um mesmo gene pode conter diversas porções codificantes (exons) intercaladas por regiões não

codificantes (introns), sendo necessário um processo de *splicing* para que as regiões intrônicas sejam removidas. A identificação de introns dentro de uma sequência gênica é extremamente complexa, visto que os padrões de sequencia que indicam a presença de uma junção "exon|intron" e "intron|exon" são altamente variáveis, mesmo dentro de um mesmo genoma. Desta forma, a predição *ab initio* em eucariotos é geralmente acompanhada do uso de dados experimentais.

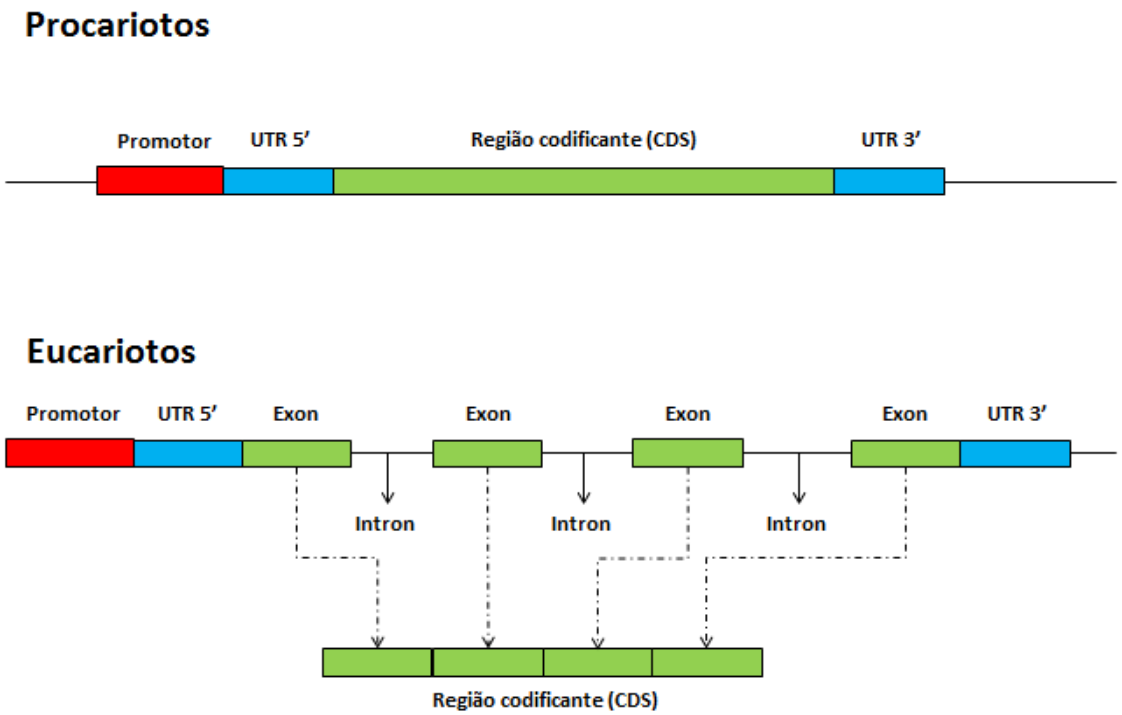

Figura 1. Estrutura gênica de procariotos e eucariotos

Identificação de genes codificantes de proteínas em procariotos

As ferramentas *ab initio* para predição de genes codificantes em procariotos, como GLIMMER, Genemark.hmm, FGenesB e Prodigal, utilizam modelos de estruturas de genes procarióticos para a realização das suas predições. No caso do GLIMMER e do Genemark.hmm, as estruturas são descritas na forma de modelos ocultos de Markov (HMM, *Hidden Markov Model*), sendo a identificação de novos genes dependente do uso de um modelo já construído ou gerado por auto-treinamendo do algoritmo. Já o Prodigal utiliza um algoritmo de computação dinâmica baseado na ocorrência de *motifs* de sítios de ligação à ribossomo (RBS, *ribosome binding sites*), conteúdo C+G e tamanho da região aberta para leitura (ORF, *Open Reading Frame*) para identificar as regiões com maior probabilidade de serem codificantes (CDS, *coding DNA sequence*).

Por apresentaram uma estrutura gênica relativamente simples, os programas de predição *ab initio* tendem a apresentar um elevado grau de acurácia, muitas vezes acima de 95% no que diz respeito a genes identificados.

Identificação de genes codificantes de proteínas em eucariotos

Como falado anteriormente, a identificação de genes em eucariotos feita através de abordagens *ab initio* é complicada pelos fatores intrinsecos a estrutura gênica destes organismos. Desta forma, é necessário muitas vezes se utilizados dados experimentais para suportar os modelos preditos. Ainda assim, ferramentas *ab intio* como Augustus e Glimmer.hmm ainda são muito utilizadas, sobretudo para genomas de organismos que ainda possuem poucas informações de proteínas ou RNA disponível.

Dos dados experimentais, até antes do surgimento das plataformas de NGS as principais informações experimentais utilizadas para a anotação de gene em eucariotos eram as sequencias de ESTs e de proteínas, sendo a obtenção de ambas um processo trabalhoso quando se objetiva o uso na identificação de genes. Entretanto, com o advento do NGS, a popularização das ténicas de RNA-Seq permitiu que este tipo de dados também fosse utilizado para a realização da identificação de genes em genomas eucariotos.

A limitação dos dados experimentais está diretamente associada com o fato do transcriptoma e o proteoma ser dinâmico, sendo necessário se extrair amostras de diferentes tecidos e condições para a identificação de uma parcela significativa de genes. Desta forma, é comum se integrar tanto dados experimentais quanto predições *ab initio*, de forma a se cobrir um maior número de genes, gerando-se assim uma anotação consenso.

Identificação do produto de cada ORF

Após a identificação, as ORFs podem ser comparadas com bancos de dados de genes, proteínas e domínios para a identificação de seus respectivos produtos. Ferramentas como BLAST, BLAT , USEARCH e HMMER podem ser usadas para se alinhar as regiões de interesse contra bancos de dados com Genbank, Uniprot e Pfam para se verificar se a região identificada possui similaridade com alguma proteína previamente caracterizada e identificar sua possivel função. Além disso, é também possivel se utilizar estas ferramentas para se remover possíveis erros no processo de predição de genes. Neste caso, ORFs falso positivas, denominadas, spurious ORFs, podem ser identificadas através da comparação com bancos de dados específicos, como o Antifam.

Identificação de RNAs não codificantes

Da mesma forma que genes que codificam para proteínas, genes de diferentes classes de RNAs não-codificantes (ncRNAs) (Ex: rRNAs, tRNAs, tmRNAs, miRNAs) podem ser identificadas através ferramentas especificas de predição. Abordagens mais simplificadas de identificação, como através de ferramentas de alinhamento local como o BLAST, permitem uma identificação rápida, mas desconsidera muitos aspectos estruturais, o que pode levar à um grande número de falto positivos. Desta forma, modelos probabilísticos, como HMMs, e preferencialmente baseados em estrutura secundária, como modelos de covariância (CM, *Covariance Models*), são preferidos para a identificação destes RNAs.

RNAmmer

A ferramenta RNAmmer utiliza como base o HMMER, baseado em modelos ocultos de Markov, para a identificação de unidades de RNA ribossômico (rRNAs) de bactérias, archeas e eucariotos. Neste caso, não são considerados aspectos estruturais devido à alta conservação deste grupo de RNAs em diferentes espécies.

tRNAscan-SE e Aragorn

Os programas preditores de RNAs transportadores (tRNAs) tRNAscan-SE e Aragorn utilizam como base diferentes algoritmos. O tRNAscan-SE usam um modelo probabilístico denominado Covariance Model (CM), similar ao modelo oculto de Markov (HMM), mas que também considera a variação simultâneas entre diferentes posições para a estimativa de estruturas secundárias. Já o Aragorn usa como base uma simulação de estrutura secundária das regiões de loop do tRNAs para determinar se identificar na sequência sub-regiões que sejam compatíveis estruturalmente com esta classe de ncRNAs. Além de tRNAs, o Aragorn também realiza a busca de RNAs transportadores-mensageiros (tmRNAs), que consistem em RNA transportadores que possuem uma região de leitura aberta responsável pela síntese de um pequeno peptídeo de sinalização.

INFERNAL e Rfam

O INFERNAL é um pacote de ferramentas de alinhamento baseada em modelos de covariância. Um de seus programas, o cmsearch, permite que um conjunto de sequência

seja comparado à um banco de dados de CMs para a identificação de regiões similares aos modelos, de forma análoga ao BLAST (Altschul et al., 1990) e ao HMMER. Para a anotação de ncRNAs em um genoma, é possível se utilizar esta ferramenta em conjunto com o Rfam, um banco de dados para estruturas de RNAs. Entretanto, a busca através de CMs do INFERNAL é consideravelmente lenta se comparada ao BLAST, o que levou ao desenvolvimento de abordagens híbridas para reduzir o tempo necessário para a anotação, como a implementada na ferramenta rfam_scan.pl (ftp://ftp.sanger.ac.uk/pub/databases/Rfam/).

Anotação estrutural

A identificação de regiões funcionais, também denominados motifs ou domínios, em um determinado genes ou proteína é denominada anotação estrutural. No caso de proteínas, ferramentas como SignalP (peptídeos sinais), TMHMM (hélices transmembrânicas), Interproscan (domínios), e bancos de dados de regiões conservadas como Pfam, PRODOM e SMART, podem ser usados para uma melhor caracterização estrutural.

Figura 2. Resultado de uma análise estrutural gerada pelo Interproscan, indicando a presença de peptídeo sinal e de um domínio previamento descrito em outras proteínas.

Anotação funcional

É possível se estender as informações referentes à um determinado gene através da identificação das funções e processos biológicos associados a ele. Bancos de dados como *Clusters of Orthologous Groups* (COG) e o *Gene Ontology* (GO), organizam em estruturas hierárquicas estas funções, e usam um conjunto limitado e curado de termos

para a identificação de cada função e processo biológico. Além disso, algumas ferramentas como o BLAST2GO, permitem que a função de uma determinada proteína seja predita através da comparação dos seus resultados de BLAST contra um banco de dados de referência, com os termos GO.

Além da determinação da função de uma proteína, é possível também se reconstruir rotas metabólicas através das funções preditas para cada proteína do genoma. Ferramentas como KAAS, MinPaths e PathPred utilizam bancos de dados como o KEGG Pathways e o SEED como base para a identificação de proteínas ortologas relacionadas à rotas já elucidadas.

Anotação Automática

Como a anotação de um genoma pode envolver um grande conjunto de ferramentas e a integração de dados de diferentes fontes, sua realização de forma "manual" (não automatizada) tornaria o processo exaustivo e sujeito a falhas. Sendo assim, diversas ferramentas foram desenvolvidas para automatizar, ao menos parcialmente, a execução e integração dos resultados de cada ferramenta.

Para anotação de genomas bacterianos as ferramentas podem ser classificadas em webserver, como o RAST, xBASE, BASys e o NCBI Prokaryotic Genome Annotation Pipeline (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/), e as ferramentas de uso local, como o Prokka, Eugene-PP, Maker e o BG7. Os webserver de anotação possuem como principal vantagem a fácil utilização, mesmo por usuários com pouca experiência com anotação de genomas e bioinformática, mas costumam oferecer pouca flexibilidade quanto à customização de seus parâmetros. Já as ferramentas de uso local, exigem um maior conhecimento técnico por serem executadas, em sua maioria, através de linhas de comandos e serem restritas a sistemas Linux / UNIX-Like / POSIX.

Visualização de anotações

A anotação de um genoma pode ser visualizada com auxílio de ferramentas denominadas genome browsers, como CGView, Artemis, JBrowse, GenomeView, que representam graficamente a organização dos cromossomos (no caso de genomas finalizados) e contigs/scaffolds (no caso de genomas rascunho), e indicados as informações referentes à cada feature, possibilitando uma análise mais acurada. Outras ferramentas, como Circos, DNA Plotter, BRING também possibilitam uma melhor

customização na forma com a qual a anotação é representada, permitindo a geração de gráficos de alta qualidade para publicações científicas. Além disso, ferramentas com o Artemis Comparison Tool (ACT), integram as funções de genome browsers com análise de sintenia, sendo útil para análises de genômica comparativa.

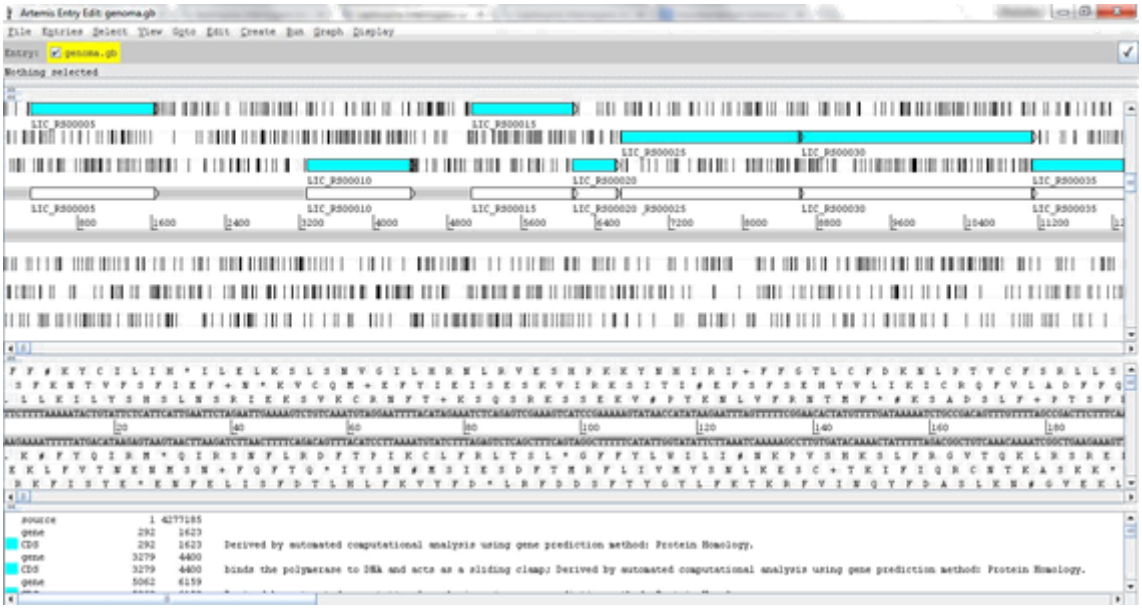

Figura 3. Artemis, um exemplo de ferramenta para visualização de anotações de genoma.

Submissão para bancos de dados públicos

Em muitos casos, como requisitos para publicação, artigos que relatam o sequenciamento de novos genes ou genomas deve incluir os respectivos códigos de acesso em bancos de dados públicos, como Genbank e EMBL. A submissão de sequências para estes bancos de dados segue um conjunto de recomendações e padrões internacionais, o que inclui o nome que será usado para as features e seus respectivos qualificadores (qualifiers), como são tratadas as regiões de gaps, quais informações foram usadas para a ligação das scaffolds, dentre outras.

No caso de submissões de genomas para o Genbank, o processo envolve o cadastro da amostra no banco de dados BioSample, do projeto de sequenciamento, no banco de dados BioProject, a geração de um arquivo genbank submission template (.sbt) com dados dos autores do projeto (https://submit.ncbi.nlm.nih.gov/genbank/template/submission/), a formatação do arquivo contendo a anotação e as sequencias através do programa TBL2ASN (http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/), e a submissão do resultado deste, um arquivo em formato .sqn, para o servidor do Genbank através do seu portal de submissão (http://www.ncbi.nlm.nih.gov/).

Exemplos de abordagens de anotação de genomas procarióticos

Predição de regiões codificantes (CDS) em genomas de bactérias com o Prodigal

O prodigal é uma ferramenta para predição de regiões codificantes em genomas de organismos procarióticos, e seu código fonte pode ser obtido a partir do endereço: http://prodigal.ornl.gov/downloads.php.

Para instalar e compilar o programa, descompacte o arquivo baixado e utilizando o comando "`tar -xvf <arquivo do código fonte>`", entre na pasta descompactada e use os seguintes comandos:

```
$ make
$ sudo make install
```

Agora que o prodigal está instalado, basta utilizar a seguinte linha de comando para realizar a anotação de um genoma. As `-i`, `-o`, `-f` e `-a` indicam, respecitivamente, o arquivo de entrada (em formato FASTA), o arquivo de saída, o formato do arquivo de saída (este caso, GBK, indica o formato Genbank) e o arquivo que conterá as sequencias da proteínas codificada por cada gene identificado.

```
$ ./prodigal -i genoma.fasta -o resultado_prodigal.gb -f gbk -a
proteinas_prodigal.fasta
```

Predição de genes de tRNA com o tRNAscan-SE

O programa tRNAscan-SE, para predição de genes de tRNAs, estando seu código fonte disponível para download a partir do endereço http://lowelab.ucsc.edu/tRNAscan-SE/. Apos descompactar o arquivo baixado, entre no diretório e utilize o comando "`make`" para compilar o programa. A utilização do tRNAscan-SE deve ser feita de dentro do seu diretório.

Para realizar a predição do tRNAs presentes em um genoma, basta se utilizar o seguinte comando:

```
$ ./tRNAscan-SE -o resultado.tab genoma.fasta
```