Anotação de genomas e bases de dados de fungos

Prof. Dr. Angel Dominguez
Universidade de Salamanca

Desde o aparecimento das tecnologias de sequenciação em grande escala nos anos 90, têm sido sequenciados e anotados os genomas completos de centenas de organismos. O êxito destes programas de sequenciação tem sido impressionante. Hoje em dia têm sido catalogados genomas completos de organismos tão diversos como bactérias, fungos, plantas, peixes, mamíferos e o genoma humano. Por isto encontramo-nos no centro de uma revolução genómica. No entanto o trabalho de interpretação das sequências dos genomas realiza-se com dificuldade. Aproximadamente 40% dos genes predictos não têm assinada qualquer função. Em Ciência é raro ser capaz de delimitar claramente as fronteiras do conhecimento nos momentos em que se produzem sendo o que sucede hoje em dia em genomica. As sequências genómicas disponíveis neste momento representam recursos extraordinários e para compreender e explorar plenamente o seu potencial devem ser anotadas, adequada e correctamente, através de aproximações bioinformáticas e experimentais. Já que muitos genes encontram-se de alguma forma em mais de uma espécie, assinalar a função de cada gene individual pode reforçar o nosso conhecimento de organismos diferentes. Por isso tem-se realizado um esforço extraordinário na anotação funcional de genes individuais com grande impacto no conhecimento de espécies e sistemas. Entre as suas utilidades podemos mencionar a descrição de novas "dianas" para fármacos, de novas enzimas com aplicações biotecnológicas, descrever novos elementos reguladores e compreender a sua função, etc.

A anotação simples é só o primeiro passo – ainda que essencial- para entender se a complexidade do organismo completo. A anotação funcional deve ademais ser mais que um catálogo de funções proteicas. Deve incluir informação acerca de interacções entre produtos génicos que poderão sugerir hierarquias de funções a cerca das quais conhecemos muito pouco. As consequências destas interacções é um sistema muito mais amplo que a soma das partes e que resulta num organismo que se auto-replica. Por isso, determinar a função dos genes é proporcionar bases com os quais pode ser explorado e compreender como trabalha o organismo completo. Nisto se baseia, por exemplo, o conceito da Biologia de Sistemas.

A ANOTAÇÃO DE GENOMAS pode-se definir como: "O processo no qual se junta informação útil a sequência genomica de DNA crua obtendo uma imagem que podemos compreender e a que por sua vez podemos interrogar". deve-se fazer antes que a sequência de DNA em si mesma, pode determinar-se de forma objectiva e com as técnicas actuais, possui um elevado grau de exactidão qualquer anotação inclusive a mais simples, é uma interpretação da sequência e portanto sujeita a erros (interpretação equivocada). Enquanto que se anotam (anotamos) genomas tratamos de ser o mais objectivos e cuidadosos na sua interpretação, um bom cientista deve sempre olhar cuidadosamente a interpretação de dados realizada pelos anotadores e assegurar-se que entende os métodos pela qual foi realizada.

O processo de anotação de genomas pode-se dividir num sentido amplo em duas grandes secções:

A) Predição de genes, localização no genoma e determinação dos codões de iniciação e de "stop".
B) Descrição da função a cada produto génico

Ainda que ambos os processos estejam relacionados e sempre se sobrepõem no processo de anotação tratamos deles de forma separada.

Ainda que a predição da sequência de um gene pareça um processo simples e muitas vezes se apresente como um problema que resultou, embora este não seja o caso, especialmente em genomas de eucariotas. Uma grande proporção de genes em qualquer genoma pode-se predizer com um grau de consistência elevado utilizando os programas descritos na Tabela 1. No entanto

pelas razões que explicaremos, que não implicam crítica aos programas em si mesmos, sempre se encontram as chamadas áreas cinzentas donde a predição génica é difícil e nunca se alcança os 100 % de eficiência. Enquanto que para um estatístico, uma certeza maior que 97% pode ser perfeitamente aceitável, como biólogos sabemos que uma grande parte do interesse em qualquer sistema está não nas regiões centrais bem conhecidas mas ao seu redor, as áreas cinzentas.

O primeiro e um dos pontos mais óbvios na predição génica é o estabelecimento de Marcos de Leitura Aberta (Open Reading Frames, ORFs). Os ORFs definem-se simplesmente como sequências de DNA maiores de um tamanho determinado começando com um códão de iniciação ATG (também GTG ou TTG em bactérias) e terminado com um códão de terminação /TAA, TAG ou TGA). Este pode ser calculado a partir da sequência de DNA por um grande número de programas (Tabela 1).

Tabela 1. Sítios de interesse no processo de anotação de genomas

---

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=genome
http://cbi.labri.fr/Genolevures/
http://psort.nibb.ac.jp/
http://www.broad.mit.edu/annotation/fungi/fgi/
http://www.ebi.ac.uk/
http://www.aspergillus.man.ac.uk/indexhome.htm?secure/WhatsNew/WhatsNewUpdates2.php~main
http://opal.biology.gatech.edu/GeneMark/
http://www.broad.mit.edu/annotation/fungi/magnaporthe/gene_finding.html
http://www.softberry.com/berry.phtml
http://www.sanger.ac.uk/
http://www.ch.embnet.org/software/TMPRED_form.html
http://phobius.cgb.ki.se/
http://www.fungalresearchtrust.org/
http://www.cbs.dtu.dk/services/TMHMM/
http://www.yeastgenome.org/
http://genolist.pasteur.fr/CandidaDB/
http://www.candidagenome.org/
http://www.galarfungail.org/
http://mips.gsf.de/
http://www.pfgd.org/pfgd/
http://genome.wustl.edu/projects/yeast/

A essência da predição génica é decidir qual dos possíveis ORFs é o que mais provavelmente codifique para una sequência de proteína. Ainda que isto possa parecer fácil é de realçar que nem todos ORFs são sequências codificantes (Coding Sequences, CDSs). Por exemplo, um elemento a considerar é o conteúdo em G+C do genoma. Se um genoma possui 30 % de G+C não vai produzir um numero menor de ORFs (de tamanho igual) que um que possua 70 % de G+C devido a que como existe um conteúdo muito menor de A + T e os codões de "stop" são mais ricos nestas bases sua presencia é reduzida. Ademais, nem todos os genes de um organismo são iguais e as suas propriedades podem variar por muitas razões. Entre elas estão os genes que codificam para proteínas hidrofóbicas grandes ou proteínas ribosómicas pequenas, etc.

Depois de se ter predicto um grupo de genes de um organismo o próximo passo é tratar de descrever uma função a cada um dos possíveis genes. Isto pode fazer-se directamente por comparação nas bases de dados ou indirectamente comparando as proteínas putativas contra modelos de domínios ou de motivos de proteínas conhecidas.

Os dois programas mais utilizados são FASTA (Pearson and Lipman 1988) e BLAST (Altschul, *et al*. 1997). Dado que se utilizam mecanismos diferentes de busca, as vezes chegam a resultados ligeiramente diferentes. FASTA trata de encontrar a melhor alienação global (i.e. começar no extremo da sequência) entre o gene a identificar e as sequências depositadas nas bases de dados enquanto que BLAST trata de encontrar os melhores alienamentos locais, isto é, os segmentos de maior identidade independentemente da sua longitude ou posição em cada uma das sequências. Cada um dos métodos têm as suas fortalezas mostrando identidade total no primeiro caso e proporcionando a possibilidade de encontrar domínios conservados no segundo. Por isto um investigador cuidadoso deve realizar e considerar buscas com ambos os programas em cada caso. Dada uma lista de dados de coincidência por cada programa é importante considerar o significado de identidade. Identidade significativa indica, neste caso, a identidade significativa pode ser a nível estatístico ou a nível biológico. É importante recordar que ainda que os valores geralmente coincidam há excepções. Ambos BLAST e FASTA descrevem a identidade significativa em cada base de dados como valor “E” ou “P”. A sua interpretação é similar, representando, em linguagem coloquial, o número de coincidências entre a sequência a identificar sobre a base de dados e as que poderiam esperar-se por casualidade. Obviamente quanto menor é o número, mais significativo é o resultado. As medidas estatísticas sugeren que valores de P> 0.05 são significativos e de P>0.01 muito significativos. Dada a incerteza ao redor das zonas cinzentas com frequência valores de $>10^{-10}$ indicam coincidência verdadeira. No entanto de novo é conveniente indicar que os resultados se devem analisar a luz do conhecimento biológico.

Genes Ortólogos, Parálogos e Sintenia

Para assinalar a função todavía podemos aplicar outras considerações. Entre os conceitos mais úteis são os conceitos de Ortología e Paralogía que podem utilizar-se como guias para analisar a informação de um genoma completo frente a outro. A definição mais simples de genes Ortólogos, é que são genes equivalentes em dois organismos diferentes que descendem directamente do mesmo gene do antecessor comum dos dois organismos (Fig. 1-I). Seguindo esta definição Parálogos são genes que estão relacionados através de um acontecimento de duplicação génica no organismo parental (Fig. 1-II) ou em um dos seus descendentes (Fig. 1-III). Assume-se que os genes ortólogos realizam a mesma função molecular ou celular enquanto que os genes parálogos podem realizar a mesma função (conduzindo a redundância funcional) ou podem ter divergido depois da sua duplicação para realizar uma função diferente ou para actuar sobre substratos diferentes. Os genes ortólogos podem-se identificar em genomas completos verificando as melhores identidades entre todos os genes dos organismos a analisar. Deve-se ter em conta que a identificação de ortólogos mediante análises computacional não produz certeza a 100 % já que em alguns casos se podem encontrar genes parálogos. A ortología pode-se verificar em genomas completos verificando-se a Sintenia. A Sintenia pode-se definir como: “A conservação da ordem e a orientação dos genes em genomas de organismos relacionados”. Os genes ortólogos podem-se encontrar no mesmo contexto em genomas de organismos similares. Assim, por exemplo, o grau de sintenia é elevado entre *Saccharomyces cerevisiae* e *Saccharomyces bayanus* e escasso entre *Magnaporthe grisea* e *Neurospora crassa*.

Outras características dos genomas completos devem identificar-se por propriedades específicas e por exemplo os telómeros identificam-se pelo motivo (TTAGGG) n (repetido n vezes).

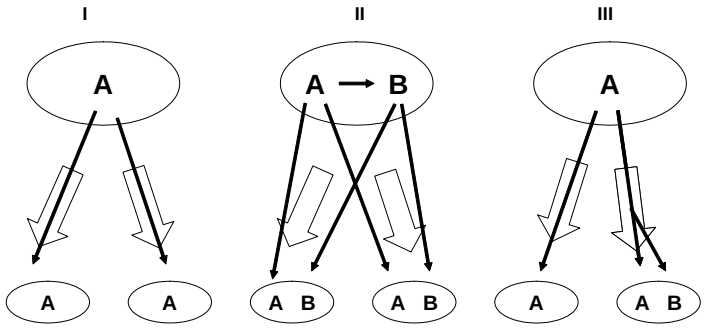

Figura 1. - Ilustração de Ortología e Paralogía. Os ovóides superiores representam as espécies antecessoras e os inferiores os descendentes. A e B representam genes. Em cada par de descendentes A e A são Ortólogos enquanto que A e B são Parálogos. I representa um descendente simples; II, representa uma duplicação génica de A a B na espécie antecessora e III representa uma duplicação génica de A a B somente numa linhagem.

Referencias

Altschul SF, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; 25: 3389-3402.

Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988; 85: 2444-2448.

Dean, R A, *et al*. The genome sequence of the rice blast fungus *Magnaporthe grisea Nature* 434; 980-986: 2005.