Métodos para construção de grupos de proteínas ortólogos e descrição de bases de dados públicas.

Luiz Thibério Rangel - número USP: 6996131
Disciplina: MAC5719
Professor: Alan Mitchel Durham

1 Introdução

A homologia esta dividida em: ortologia, paralogia e xenologia. Estas três relações são bastante importantes para os estudos de genômica comparativa, sobre tudo a ortologia, que além das informações evolutivas ligadas a ela, também permite que a função de um ortólogo seja transferida a outro. Duas proteínas são ortólogas quando foram originadas a partir de um processo de especiação, enquanto que são chamadas de parálogas caso tenham divergido em um processo de duplicação (Fitch, 1970; Fitch, 2000).

Existem diversas maneiras para identificarem-se proteínas ortólogas, as mais utilizadas na literatura serão discutidas neste trabalho, juntamente com os principais repositórios online de grupos de proteínas ortólogas.

2 Métodos para identificação de proteínas ortólogas

2.1 Melhor *hit* recíproco

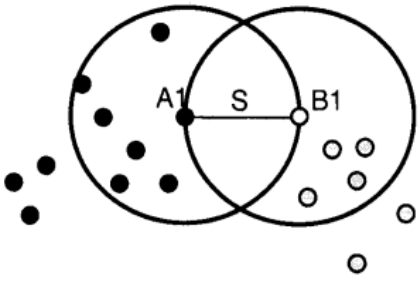
Um dos métodos mais utilizados para identificar proteínas ortólogas entre dois organismos baseia-se no fato de que proteínas ortólogas tendem a ser mais semelhantes entre si do que com qualquer outra proteína do outro organismo. Exemplos de ferramentas que utilizam essa metodologia, cada uma com a sua peculiaridade, são: INPARANOID (Remm, et al., 2001), OrthoMCL (Li, et al., 2003), HamSTR (Ebersberger, et al., 2009); alguns bancos de dados também aplicam essa metodologia durante a construção dos grupos ortólogos: COG/KOG (Tatusov, et al., 2003; Tatusov, et al., 1997); eggnog (Jensen, et al., 2008; Muller, et al., 2010).

2.1.1 INPARANOID

Possui um dos melhores resultados dentre as ferramentas para agrupamento de ortólogos (Altenhoff and Dessimoz, 2009; Chen, et al., 2007; Jun, et al., 2009), e realiza apenas a identificação de ortólogos entre dois organismos, para combinar os resultados pareados de diversos organismos utiliza-se o MULTIPARANOID (Alexeyenko, et al., 2006).

O usuário deve fornecer ao *software* os conjuntos de proteínas que se deseja comparar em dois arquivos FASTA, e também se pode utilizar um terceiro conjunto de proteínas como grupo externo. Para cada comparação entre dois organismos o INPARANOID executa quatro alinhamentos par-a-par utilizando o BLAST (Altschul, et al., 1997), cada organismos contra ele mesmo e contra o outro (por exemplo, A x A e B x B, seguido por A x B e B x A). As medias dos *hits* recíprocos são calculadas caso os alinhamentos possuam um escore maior que 50 e a região alinhada tenha pelo menos 50% do comprimento da maior proteína, as proteínas com maior média de escore recíproco são assinaladas como ortólogos, e são adicionados como parálogas aquelas proteínas que são mais similares aos ortólogos iniciais que qualquer proteína do outro organismo (Figura 1).

Figura 1 Adição dos ortólogos ao grupo inicial, apenas proteínas mais similares aos ortólogos iniciais que a outras proteínas do outro organismo.



No caso de querer-se analisar grupos ortólogos de mais de dois organismos aplica-se o MultiParanoid aos resultados do INPARANOID. No caso de compararmos três organismos (A, B e C) tomamos os grupos ortólogos do par A-B como base e procuramos os ortólogos-chave de A-B em A-C e B-C. Os grupos de A-B que possuírem ortólogos-chave em comum com os grupos de A-C e B-C serão enriquecidos com todas as proteínas dos grupos com ortólogos-chave em comum. Caso haja divergência em relação a alguma proteína (i.e. tenha sido assinalada a dois grupos ortólogos diferentes) ela será retirada do grupo no qual ela possua menor valor de confiança.

2.1.2 OrthoMCL

O OrthoMCL possui uma metodologia muito próxima a do INPARANOID, mas diferencia-se principalmente por utilizar o algoritmo de clusterização de Markov (MCL) para determinar a granularidade dos grupos. Ao invés de usar escore e/ou alinhamento como limiar o OrthoMCL utiliza um p-valor de 1e-5 como limiar, escolhido empiricamente. As prováveis relações de homologia determinadas pelo BLAST recíproco são convertidas em um grafo, onde os nós são as seqüências protéicas e as arestas são as médias dos $-\log_{10}$(p-valor). A clusterização de Markov faz uso de simulações e considera todas as relações simultaneamente para separar os outparalogos, ortólogos distantes e outros falsos positivos em geral.

Um parâmetro importante do OrthoMCL é o valor de dilatação, que controla a granularidade dos grupos. Quanto maior for o valor de dilatação mais coeso é o grupo formado. E os clusters com proteínas de ao menos duas espécies são consideradas ao final da análise.

2.1.3 HaMStR (Hidden Markov Model based Search for Orthologs using Reciprocity)

Esta ferramenta realiza comparações bidirecionais para identificar proteínas ortólogas. Entretanto, por seu objetivo principal ser o agrupamento de ESTs em grupos definidos inicialmente por proteomas as comparações são diferentes das implementadas nas ferramentas anteriores.

Além do conjunto de ESTs que se deseja classificar de acordo com suas relações de ortologia, o usuário deve fornecer um conjunto de grupos ortólogos previamente gerados por outra ferramenta (i.e. INPARANOID ou OrthoMCL) e as seqüências protéicas utilizadas na geração dos mesmos. As proteínas de cada grupos ortólogo são alinhadas e utilizadas para gerar um HMM de perfil, o qual será alinhado contra o conjunto de ESTs alvo. Os ESTs que obtiverem um alinhamento acima do limiar em seguida são alinhados, via BLASTp, contra o conjunto de proteínas inicial mais próximo. Desta maneira comparam-se os conjuntos de seqüência em duas direções, sendo uma delas mais restrita e outra permissiva em relação aos alinhamentos. O próximo passo é avaliar se a proteína com a qual o EST obteve o melhor alinhamento foi utilizada para gerar o HMM de perfil com o qual ele, o EST, obteve o melhor alinhamento. Caso a avaliação seja positiva, o EST consultado será inserido ao grupo ortólogo contra o qual apresentou mais similaridade.

2.2 Identificações baseadas em filogenia

Existem diversas ferramentas que utilizam esta metodologia para identificar proteínas ortólogas em um conjunto de organismos, como: RIO (Zmasek and Eddy, 2002) e Orthostrapper (Storm and Sonnhammer, 2002), estas metodologias tendem a possuir uma baixa taxa de falsos positivos e uma alta taxa de falsos negativos (Chen, et al., 2007). Além da grande da falta de equilíbrio entre sensibilidade e especificidade encontrada nestas metodologias outra desvantagem é a necessidade de uma filogenia bem determinada e completa entre as espécies.

2.3 Considerações

Dentre as ferramentas descritas neste trabalho as duas ferramentas mais utilizadas são INPARANOID/MultiParanoid e o OrthoMCL, estas duas ferramentas são tidas como referências para identificação de proteínas ortólogas, seja entre dois organismos ou mais (Altenhoff and Dessimoz, 2009; Chen, et al., 2007; Ebersberger, et al., 2009; Jun, et al., 2009). A ferramenta HaMStR por ser recente, não foi levada em conta pelos trabalhos comparativos, e até mesmo por seu objetivo (agrupar ESTs), provavelmente não será tão utilizada quanto o INPARANOID ou OrthoMCL, já que a própria geração de ESTs não ocorre mais com a mesma freqüência, pelo próprio barateamento das tecnologias de seqüenciamento.

As restrições existentes para aplicação de metodologias baseadas em filogenia são os fatores limitantes para o uso dessa metodologia, já que uma filogenia coesa e com uma larga quantidade de organismos não esta disponível para a maioria das espécies.

3 Bancos de dados

Existem diversos bancos de dados de grupos de proteínas ortólogas, construídos de maneiras diversas e que permitem, ou não, a classificação de outras proteínas em seus grupos. Os bancos de dados discutidos aqui são todos formados buscando-se homologia por alinhamentos.

3.1 COG/KOG

Mesmo tendo sido descontinuada desde 2003 continua sendo uma das bases mais citadas na literatura. O COG contém 66 genomas de bactéria e archaea, e a formação dos grupos ortólogos se da utilizando a metodologia de melhores *hits* recíprocos. Todos os organismos são alinhados uns contra os outros, e os melhores *hits* recíprocos entre cada espécie são identificados, e os conjuntos de proteínas que contenham pelo menos três proteínas de organismos diferentes formam os grupos ortólogos. Como não apenas o melhor hit é utilizado para formar os grupos ortólogos, os "triângulos" com proteínas em comum são mesclados. Cada grupo ortólogo é curado manualmente visando corrigir agrupamentos errôneos e reagrupar possíveis proteínas com domínios múltiplos. Já o KOG contém proteínas oriundas de sete eucariotos, e foi formado inicialmente baseado nos grupos criados anteriormente para bactérias e archaea, enquanto que apenas as proteínas que não foram agrupadas nos antigos grupos foram submetidas ao procedimento de comparações bilaterais.

Um dos fatores que tornou este banco de dados popular foram às ferramentas *online* Cognitor/Kognitor, que classificam proteínas em seus grupos ortólogos, informando seus prováveis ortólogos e funções.

3.2 eggNOG

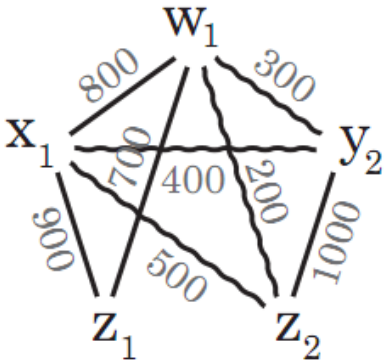
Este banco de dados de proteínas ortólogas pode ser compreendido como uma extensão do COG/KOG. Possui seqüências protéicas de 630 genomas, e como passo inicial essas proteínas foram adicionadas aos grupos previamente geradas pelo COG/KOG analisando os melhores hits encontrados por alinhamentos Smith-Waterman. As proteínas que não foram assinaladas a nenhum grupo ortólogo pré-existente foram alinhadas entre si, utilizando Smith-Waterman, e os conjuntos de pelo menos duas proteínas que forem os melhores *hits* recíprocos formam um novo grupo ortólogos. Todo o processo empregado pelo eggNOG é não supervisionado, da criação dos grupos ortólogos as avaliações de suas consistências.

3.3 OMA (Schneider, et al., 2007)

Este banco de dados anunciou este ano que possui implementado em seu banco de dados proteínas provenientes de 1.000 genomas (Altenhoff, et al., 2010), e a construção de seus grupos esta dividido em quatro passos (Roth, et al., 2008): (1) todas as proteínas são

alinhadas utilizando-se Smith-Waterman e matriz PAM-224, as proteínas cujos alinhamentos possuírem características acima do limiar são assinaladas como ortólogos-candidatos; (2) os ortólogos-candidatos são avaliados se a distância evolutiva entre eles é menor que a distancia para qualquer outra proteína do outro organismo, e caso sim são assinalados como pares-estáveis; (3) compara-se a distância entre os pares-estáveis com suas distâncias para possíveis ortólogos em um terceiro organismo com a finalidade de manter apenas os pares-estáveis mais recentes, que passam a ser assinalados como pares-verificados; (4) o agrupamento das proteínas é realizado de maneira não obrigatoriamente transitiva e, ao convertermos as relações dos pares verificados em um grafo (onde os nós são as seqüência e as arestas os escores dos alinhamentos), selecionamos os sub-grafos com a maior soma das arestas como grupos ortólogos (Figura 2).

Figura 2 Grafo gerado a partir das relações dos pares-verificados, a soma das arestas que ligam as proteínas W_1- X_1-Z_1 é a maior o possível, e dentre o restante nenhuma combinação é maior que Y_2-Z_2. Desta forma, os grupos ortólogos formados são {W_1, Z_1 e X_1} e {Y_2 e Z_2}



Recentemente foi adicionado ao escopo deste banco de dados uma busca de similaridades com as proteínas classificadas e caracterização funcional dos seus grupos ortólogos

3.4 OrthoDB (Kriventseva, et al., 2008)

É um banco de dados de grupos de proteínas ortólogas em eucariotos, possui 115 genomas de vertebrados, artrópodes e fungos (Waterhouse, et al., 2010). Seus grupos ortólogos são formados baseados na metodologia melhores *hits* recíprocos utilizando Smith-Waterman com o PARALIGN (Saebo, et al., 2005). Metodologia próxima a aplicada no COG/KOG, onde para o delineamento dos grupos ortólogos deve haver ao menos uma

triangulação das proteínas com alinhamentos maiores que 30 aminoácidos e e-value menor que 1e-3, caso não haja a triangulação o limiar do *e-value* passa a ser 1e-6.

A versão do banco de dados lançada em setembro de 2010 foi incrementada com: caracterização funcional e evolutiva dos grupos ortólogos. A caracterização funcional dos grupos ortólogos foi realizada submetendo as proteínas agrupadas à análises de GO e domínios do InterPro, e a partir destas informações os grupos foram classificados. Desta maneira, 95% das proteínas agrupadas estão em algum grupos com alguma caracterização funcional de GO e/ou InterPro. A caracterização evolutiva dos grupos esta representada por: (1) uma análise da taxa de divergência de cada grupo, que é calculada como sendo a média das identidades entre cada proteína do grupo normalizadas pela média de todos os grupos; (2) análise da distribuição filética de cada grupo ortólogos; e (3) grau de relação entre os grupos, que é determinado pelos alinhamentos par-a-par das proteínas de um determinado grupos contra todos de um outro grupo.

3.5 Considerações

As construções das bases de dados de grupos ortólogos descritas neste trabalho não divergem muito em relação a sua metodologia básica para formação dos grupos. A base COG/KOG é mais antiga, e ainda a mais utilizada, e por muito tempo foi a única a caracterizar funcionalmente seus grupos ortólogos, outras bases mais antigas, que não foram descritas, neste trabalho como o InparanoidDB (O'Brien, et al., 2005) e OrthoMCL-DB (Chen, et al., 2006) não caracterizam diretamente seus grupos ortólogos em relação a função. A presença mais freqüente deste tipo de informação nos bancos de dados é um reflexo da utilização cada vez maior da ortologia na anotação de proteínas, deixando de ser usada apenas em análises evolutivas.

A base OMA é a que possui a maior quantidade de organismos representados, e avaliando-se os algoritmos expostos em seus respectivos artigos é uma das mais bem planejadas e com melhores resultados de acertos de relações de ortologia identificadas (Altenhoff and Dessimoz, 2009).

4 Identificação de ortólogos utilizando sintenia

Uma forma de identificar ortólogos utilizada por (Jun, et al., 2009) consiste em comparar-se apenas as regiões vizinhas de cada gene. Neste trabalho a região sintênica considerada abrange os seis genes mais próximos ao gene-alvo (três de cada lado). Ao comparar dois genes-alvo, considera-se que a sintenia foi conservada caso o alinhamento dos genes vizinhos possua um *e-value* menor que 1e-5, e para que os genes-alvo sejam considerados ortólogos uma taxa mínima de conservação sintênica deve ser encontrada entre os genes-alvo, definida pelo usuário.

Segundo o trabalho apresentado, esta metodologia apresentou resultados muito similares aos do INPARANOID e OrthoMCL, entretanto há um fato relacionado a duplicação não explicado durante a discussão do artigo. Caso haja a duplicação de uma proteína e a copia seja alocada em uma região distante da original, é mais provável que a copia sofra uma menor pressão para manutenção da função que a original, já que sua localização é diferente. Entretanto, não há garantias de que este fato ocorra em uma freqüência relevante e o trabalho não cita nenhum trabalho que avalie esta situação.

5 Referências

Alexeyenko, A., Tamas, I., Liu, G. and Sonnhammer, E.L. (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes, *Bioinformatics*, **22**, e9-15.

Altenhoff, A.M. and Dessimoz, C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods, *PLoS Comput Biol*, **5**, e1000262.

Altenhoff, A.M., Schneider, A., Gonnet, G.H. and Dessimoz, C. (2010) OMA 2011: orthology inference among 1000 complete genomes, *Nucleic Acids Res*.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Chen, F., Mackey, A.J., Stoeckert, C.J., Jr. and Roos, D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups, *Nucleic Acids Res*, **34**, D363-368.

Chen, F., Mackey, A.J., Vermunt, J.K. and Roos, D.S. (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes, *PLoS One*, **2**, e383.

Ebersberger, I., Strauss, S. and von Haeseler, A. (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs, *BMC Evol Biol*, **9**, 157.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins, *Syst Zool*, **19**, 99-113.

Fitch, W.M. (2000) Homology a personal view on some of the problems, *Trends Genet*, **16**, 227-231.

Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. and Bork, P. (2008) eggNOG: automated construction and annotation of orthologous groups of genes, *Nucleic Acids Res*, **36**, D250-254.

Jun, J., Mandoiu, II and Nelson, C.E. (2009) Identification of mammalian orthologs using local synteny, *BMC Genomics*, **10**, 630.

Kriventseva, E.V., Rahman, N., Espinosa, O. and Zdobnov, E.M. (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs, *Nucleic Acids Res*, **36**, D271-275.

Li, L., Stoeckert, C.J., Jr. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res*, **13**, 2178-2189.

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powell, S., von Mering, C., Doerks, T., Jensen, L.J. and Bork, P. (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations, *Nucleic Acids Res*, **38**, D190-195.

O'Brien, K.P., Remm, M. and Sonnhammer, E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs, *Nucleic Acids Res*, **33**, D476-480.

Remm, M., Storm, C.E. and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, *J Mol Biol*, **314**, 1041-1052.

Roth, A.C., Gonnet, G.H. and Dessimoz, C. (2008) Algorithm of OMA for large-scale orthology inference, *BMC Bioinformatics*, **9**, 518.

Saebo, P.E., Andersen, S.M., Myrseth, J., Laerdahl, J.K. and Rognes, T. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology, *Nucleic Acids Res*, **33**, W535-539.

Schneider, A., Dessimoz, C. and Gonnet, G.H. (2007) OMA Browser--exploring orthologous relations across 352 complete genomes, *Bioinformatics*, **23**, 2180-2182.

Storm, C.E. and Sonnhammer, E.L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability, *Bioinformatics*, **18**, 92-99.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. (2003) The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families, *Science*, **278**, 631-637.

Waterhouse, R.M., Zdobnov, E.M., Tegenfeldt, F., Li, J. and Kriventseva, E.V. (2010) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011, *Nucleic Acids Res*.

Zmasek, C.M. and Eddy, S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs, *BMC Bioinformatics*, **3**, 14.