OrthoFinder

数据准备

蛋白序列文件

- fasta格式
- 对于一个基因最好只保留一个最长的转录本
- 每个物种一个文件，所有物种的文件放在同一个文件夹下，文件后缀：.pep/.fasta/.fa/.faa/.fas
- 文件命名最好简洁，文件名会被用在结果中作为一个物种的id。例如，Homo_sapiens.GRCh38.pep.all.fa改为Homo_sapiens.fa或Human.fa

OrthoFinder：快速的基因家族分析工具
主要功能是找正交群(orthogroups)和直系同源(orthologs)，推断所有正交群的基因树，并能识别基因复制事件。
它推断出一个有根的物种树，并将基因复制事件从基因树映射到物种树的分支。
它还为比较基因组分析提供全面的统计数据。
OrthoFinder使用简单，运行它所需的只是一组 FASTA 格式的蛋白质序列文件。
参考:
https://github.com/davidemms/OrthoFinder
https://lxz9.com/2021/11/18/OrthoFinder/
https://xuzhougeng.top/archives/OrthoFinder2-fast-and-accurate-phylogenomic-orthology-analysis-from-gene-sequences

Homologs

- Orthogroups，正交群
  - 由物种的最后共同祖先进化而来的一组基因
  - 正交群的一个用途就是鉴定直系同源基因：因为常规鉴定同源基因的方法是使用基因树，而正交群中的基因正是用来建树的一组基因。
- Orthologues，直系同源
  - 由两个物种的最后一个共同祖先(LCA)中的一个基因进化而来的一对基因
- Paralogues，旁系同源
  - 在一个基因复制事件中从一个基因分离出来的一对基因

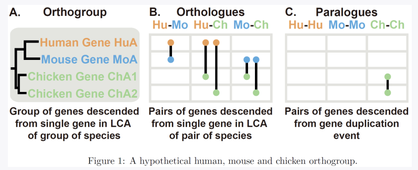
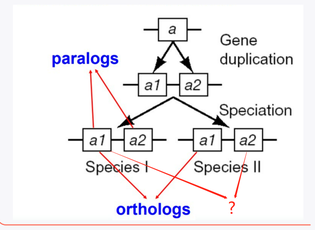
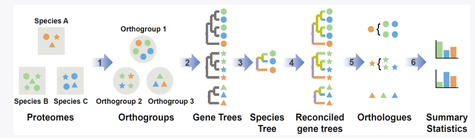


Figure 1: A hypothetical human, mouse and chicken orthogroup.





分析过程

运行：**orthofinder -f <fasta_files_directory>**

1. BLAST all-vs-all比对。使用BLASTP，以evalue=10e-3进行搜索，寻找潜在的同源基因（除了BLAST, 还可以选择DIAMOND和MMSeq2）
2. 基于基因长度和系统发育距离对BLAST bit得分进行标准化
3. 使用RBNHs确定同源组序列性相似度的阈值
4. 构建直系同源组图(orthogroup graph)，用作MCL的输入
5. 使用MCL对基因进行聚类，划分直系同源组
6. 系统发育的分析和统计
   - 为每个直系同源组构建基因系统发育树
   - 使用STAG算法从无根基因树上构建无根物种树
   - 使用STRIDE算法构建有根物种树
   - 有根物种树进一步辅助构建有根基因树
   - 基于Duplication-Loss-Coalescent 模型，有根基因树可以用来推断物种形成和基因复制事件，最后记录在统计信息中。

orthofinder 参数扩展

-M dendroblast	用于基因树推断的方法，默认是dendroblast。可选：msa
-S diamond	序列比对程序，可选: blast, diamond, diamond_ultra_sens, blast_gz, mmseqs, blast_nucl
-A mafft	多序列比对程序，可选: mafft, muscle。要求'-M msa'
-T fasttree	建树方法，可选: fasttree, raxml, raxml-ng, iqtree。要求'-M msa'

结果说明

Orthogroups

- Orthogroups.tsv：（主要的基因家族聚类文件）每一行是一个正交群
- Orthogroups.txt：同上，OrhtoMCL的输出格式
- Orthogroups.GeneCount.tsv：同上，基因数的统计
- Orthogroups_UnassignedGenes.tsv：物种特异性的基因家族
- Orthogroups_SingleCopyOrthologues.tsv：单拷贝基因家族
- N0.tsv

Phylogenetic Hierarchical Orthogroups

- N0.tsv
- N1.txt, N2.tsv, ...
- 具体说明https://github.com/davidemms/OrthoFinder#results-files-phylogenetic-hierarchical-orthogroups-directory

Orthologues

每个物种都有一个子目录，该子目录又包含每个成对物种比较的文件，列出该物种对之间的直系同源。

Gene Trees

为具有 4 个或更多序列的正交群推断出的有根系统发育树。

Resolved Gene Trees

为具有4个或更多序列的正交群推断出一个有根系统发育树，并使用OrthoFinder hybrid species-overlap/duplication-loss coalescent模型进行分析，以确定对基因树的更简洁的解释。

Species Tree

- SpeciesTree_rooted.txt：使用STAG算法推断出来的物种树，并使用STRIDE算法定根。
- SpeciesTree_rooted_node_labels.txt：物种树节点给出了标签(而不是支持值)，以允许其他结果文件交叉引用物种树中的分支/节点(例如基因复制事件的位置)。
- Potential_Rooted_Species_Trees：推断的可能的有根物种树

Comparative Genomics Statistics

- Duplications_per_Orthogroup.tsv：每个正交群识别出的重复次数
- Duplications_per_Species_Tree_Node.tsv：物种树每个节点识别出的重复次数
- Orthogroups_SpeciesOverlaps.tsv：两个物种之间共享的正交群，矩阵文件
- OrthologuesStats_*.tsv files：矩阵文件，给出每对物种之间一对一、一对多和多对多关系中的直系同源基因的数量
- Statistics_Overall.tsv：包含关于正交群大小和分配到正交群的基因比例的基本统计数据；总的来说，至少 80% 的基因被分配到正交群。少于此值意味着可能会遗漏某些剩余基因实际存在的直系同源关系，物种采样不佳是造成这种情况的最可能原因。 让我们也检查每个物种的百分比
- Statistics_PerSpecies.tsv：包含与 Statistics_Overall.csv 文件相同的信息，但针对每个单独的物种

Gene Duplication Events

- Duplications.tsv：列出了每个基因树的每个节点识别出的所有基因复制事件
- SpeciesTree_Gene_Duplications_0.5_Support.txt：每个节点或物种名称后面的数字是在导致节点/物种的分支上发生的具有至少 50% 支持度的基因复制事件的数量。

Orthogroup Sequences

每一个fasta文件包含一个Orthogroup里的所有蛋白序列

Single Copy Orthologue Sequences

只包含单拷贝基因的Orthogroup文件

WorkingDirectory

这包含了所有中间文件，用于断点检测、重复运行等，可以忽略。中间文件占空间较大，如果不需要可以删除。