The Biostar Handbook: A Beginner's Guide to Bioinformatics[1]

#course

1. **Bioinformatics**

   Bioinformatics is a data science that investigates how information is stored within and processed by living organisms.

   Bioinformatics requires multiple skill sets, extensive practice, and familiarity with multiple analytical frameworks.

   Tools change. Concepts don't. Over time tools implement the same concepts better and better.

   No project was identical, and we were surprised at how common one-off requests were. There were a few routine procedures that many people wanted, such as finding genes expressed in a disease. But 79% of techniques applied to fewer than 20% of the projects. In other words, most researchers came to the bioinformatics core **seeking customized analysis, not a standardized package**.

   Data management and analysis are meaningless without accurate and insightful interpretation. Bioinformaticians discover or support biological hypotheses via the results of their primary analyses, and so they must be able to **interpret their findings in the context of ongoing scientific discourse**.

   "Understanding data formats, what information is encoded in each, and when it is appropriate to use one format over another is an essential skill of a bioinformatician."

   information+format

   随记 "生物信息学"

2. **Organizing project**

```
mkdir -p bin src doc results data genome
```

Organizing projects

3. **Starting an analysis**

   - download just a subset of the data
   - build random samples
   - use a segment of the target genome
   - generate data with known properties

4. **Data reproducibility**

   your analysis is reproducible if and only if you too are able to reproduce it many times over with ease (and others can reproduce it as well)

   The so-called "scientific narrative" is simplified and misleading - it depicts the bioinformatics analysis as a linear chain of decisions: "first we did this, then we did that." In this way it paints a picture that does not correspond to reality. **The reality is : analysis is not a linear , predicable or step-by-step work**.

   within the scientific paper the long and branching chains of decisions and actions are described as straightforward choices, and **the failed attempts are never mentioned.**

   Once you know what type of insights are in your data there is almost always a simpler and more efficient method to find them. (**there is always existing alternatives**)

5. **GO/SO**

   The MISO Sequence Ontology Browser
   GitHub - The-Sequence-Ontology/SO-Ontologies: Collect of SO Ontologies
   QuickGO
   GeneCards - Human Genes | Gene Database | Gene Search

   "The **Sequence Ontology** (SO): a vocabulary for information related to sequence features. The **Gene Ontology** (GO), a vocabulary for information related to gene functions."

   Ontology本体论， 除了SO，GO，还有其它本体论

“consistency”

一致性比正确性更重要。如果双方对同一个概念各抒己见，那么讨论只会浪费时间。

“ **HUGO Gene Nomenclature Committee (HGNC)** available via http://www.genenames.org/ is the only worldwide authority that assigns standardized nomenclature to human genes”

nomenclature命名系统，HGNC唯一有权利

“What a piece of DNA is: annotations or classifications.
What a piece of DNA does: functional analyses.”

这个dna片段是什么-> annotation, classification
这个dna片段做什么->function
“de-novo genome assembly” **构建基因组，然后为其注释**
“RNA-Seq” **通过差异表达的转录本解释表型**

“Does all sequencing data obey the rules of SO?”

No, 不是所有数据都遵守SO，如CDS

“For that and other reasons the majority of life scientists do not make use of the SO relationships and use the SO terms only.”

一般不用 SO relationship 只用 SO terms

“The ontology is intended to categorize gene products rather than the genes themselves. ”

GO用语分类基因产物，而非基因本身。因为一个基因可以有不同的产物，行使不同的功能

“The sub-ontologies are as follows:”

GO的组织结构（namespace）：
CC: cellular component （在哪里发挥）
MF: molecular function （如何发挥功能）
BP: biological process （为什么需要该产物/功能）

"The GO vocabulary is designed to be species-agnostic, and includes terms applicable to prokaryotes and eukaryotes, and to single and multicellular organisms."

GO stores the deposited knowledge on different organisms.
GO本体被构造为有向的非循环图，其中每个术语定义了与同一域中的一个或多个其他术语的关系，并且有时与其他域有关。

"There are so called reduced (slim) vocabularies that simplify the data to broader terms at the cost of losing some of its lower level granularity."

more readingWhat can I do with a GO slim? | EMBL-EBI Train online

"The underlying data for the gene ontology consists of two files: 1. Gene ontology definition file. 2. A gene association file."

1.GO definition file: 词条定义
2.GO association file: 不同命名体系与GO词条的映射关系

"In the gene association file, each gene product may be associated with one or more GO terms in each category."

a gene product ID is connected to one or more GO functions

"There it goes to show how not everything is all and well in the GO world…"

GO的注释数据库会不断地更新

"Keep that in mind the next time something completely surreal seems to happen."

then keep searching

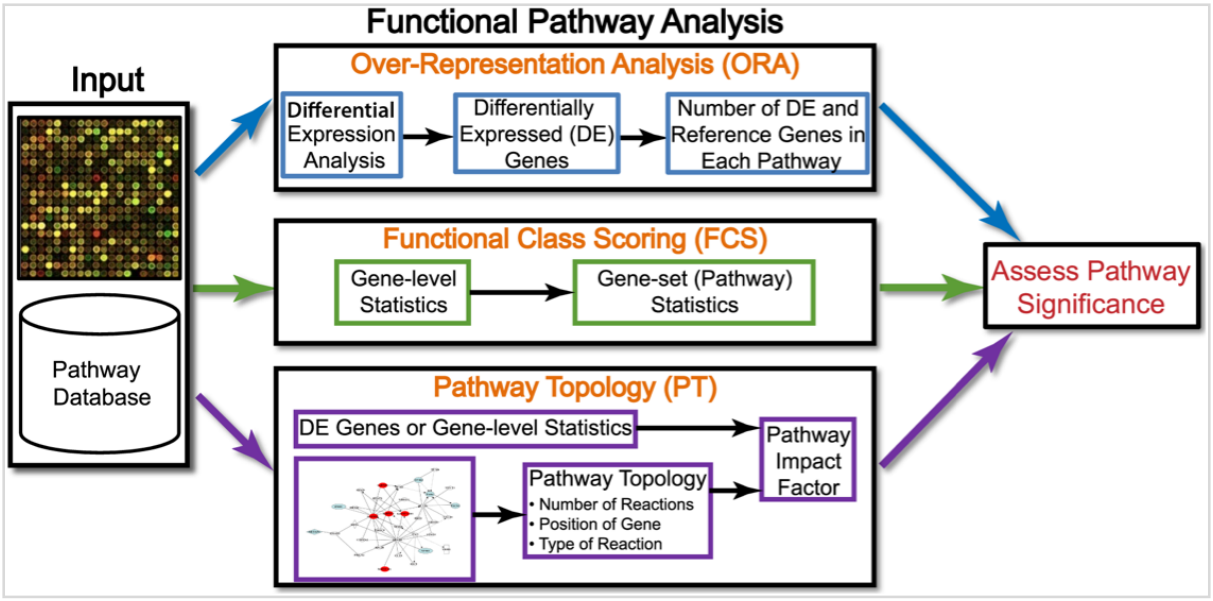
6. Data format(1)

obo : "The obo format is plain text and is NOT column oriented. "
GAF： GO Annotation File

7. Functional analysis （algorithms）

Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges

Functional analysis / Pathway analysis



"An **Over-Representation Analysis** (ORA) attempts to find representative functions of a list of genes."

ORA 过表征分析

"ORA analysis computes four numbers"

ORA分析需要你提供4类输入：
1.一共有多少个基因，也就是背景
2.属于某分类的基因有多少个
3.样本一共有多少个基因
4.样本属于某分类的有多少个基因

然后通过超几何分布或者2X2独立表进行检验

"ORA analyses should be used as a hypothesis generation and testing framework and not as a method that provides the final answer to a problem."

1.ORA没有考虑到基因的表达水平，仅仅关注基因是否属于分类
2.ORA仅仅使用部分数据，存在主观臆断
3.基因和功能被认为是相互独立。这只是一种统计学假设而已，实际情况下并非如此。

"FCS"

functional class scoring

第一步： 通过实验计算出单个基因的基因水平(gene-level)的统计值，比如说基因差异表达衡量会用到的ANOVA，Q-statistic, 信噪比， t-test, Z-score等。
第二步： 同一条通路上所有基因的基因水平(gene-level)统计值聚合成单个通路水平（pathway-level)的统计值。可选方法有，Kolmogorov-Smirnov statistic [21,29],基因水平统计值的和, 均值或中位数, Wilcoxon rank sum，maxmean statistic。
第三步：评估通路水平统计显著性。这一步所需要的统计学思想是重抽样（bootstrap）。也就是对于一个特定通路而言，随机排序和按照一定规则排序是否有差异。

"FCS methods account for dependence between genes in a pathway, which ORA does not"

它的基本假设是：虽然单个基因的巨大改变会对通路有显著性影响，但是那些功能相关的类似微效基因累加后也能有显著效果。

缺点：
第一，它是单独分析每个通路，而不是多通路组合分析。第二，FCS也只将基因表达的差异用做给定通路的排序而已。比如说A和B的表达量分别改变了2倍和20倍，但是对于不同的通路而言，A和B的排名就有可能相同。

"Pathway Topology (PT)-Based Approaches"

基于通路拓扑学的方法

8. Functional analysis （tools）

"Gene set enrichment analysis" refers to the process of discovering the common characteristics potentially present in a list of genes."

Functional enrichment

"One of the most commonly used is over-representation analysis (ORA). An ORA examines the genes in a list, summarizes the GO annotations for each, then determines whether there any annotations are statistically over-represented (compared to an expectation) in that list."

"Overall GO enrichment is surprisingly subjective and different methods will produce

different results.”

->clusterProfiler,它支持ORA和FCS两类算法。函数为:
enrichGO, gseGO: GO富集分析
enrichKEGG, gseKEGG: KEGG富集分析
enrichDAVID: DAVID富集分析
clusterProfiler

9. Database

International Nucleotide Sequence Database Collaboration | INSDC

NCBINational Center for Biotechnology Information
EMBLEMBL - European Molecular Biology Laboratory - The European Molecular Biology Laboratory
DDBJBioinformation and DDBJ Center
USCSUCSC Genome Browser Home

GenBank：存放所有注释和已被发现的DNA序列信息
SRA： 存放高通量测序产生的短读数据
PDB： 蛋白3D结构数据库
uniprot： 最权威的蛋白序列数据库
UCSC Genome Browser： 脊椎动物相关数据库
FlyBase： 果蝇相关数据库
WornBase： 蠕虫相关数据库
SGD： 酵母相关数据库
RNA-Central： RNA相关资源汇总数据库
TAIR：拟南芥相关数据库
EcoCyc： 大肠杆菌数据库

“using bioinformatics data resources is frustrating, requires patience, and often requires the suspension of disbelief.”

suspension of disbelieve

“biological meaning”

use a biological meaning name

10. Data submission

Sequence Read Archive (SRA): raw sequence reads from NGS
Gene Expression Omnibus (GEO): functional genomic data like RNASeq, ChIP-seq etc
Database of Short Genetic Variations (dbSNP): variation specific data
Database of Genomic Structural Variations (dbVar)
Database of Expressed Sequence Tags (dbEST)
Transcriptome Shotgun Assembly Sequence Database (TSA): transcriptome assemblies
Whole Genome Shotgun Submissions (WGS)
Metagenomes
GenBank
Genomes

11. Download reference genomic data

NCBI web: https://www.ncbi.nlm.nih.gov/
NCBI FTP: ftp://ftp.ncbi.nlm.nih.gov/
Ensembl web: http://useast.ensembl.org/index.html
Ensembl FTP: ftp://ftp.ensembl.org/pub/release-86/
Biomart: http://www.ensembl.org/biomart/martview/
UCSC Downloads: http://hgdownload.cse.ucsc.edu/downloads.html
UCSC FTP: ftp://hgdownload.cse.ucsc.edu/goldenPath/