基因组分析

猜测严军老师会考察贝叶斯公式（比如17年的真题）。绝大部分课堂内容是算法层面，个人觉得不会考。

**钟杨：（严军）**

1.**. BLAST**

Blastn: 核酸序列🡪核酸序列

Blastp：蛋白质序列🡪蛋白质序列

Blastx：核酸序列（翻译）🡪蛋白质序列

Tblastn：蛋白质序列🡪核酸序列（翻译）

Tblastx：核酸序列（翻译）🡪核酸序列（翻译）

**Blast流程Blast---Basic Local Alignment Search Tool**

在浏览器中输入NCBI网址即可进入BLAST的主页，主页面按不同的搜索任务做了分类，选择所需的类别进行搜索。进入搜索页面后，把查询序列粘贴到“search”编辑框中。在database选择项中，有若干数据库可以选择，BLAST的缺省数据库是nr数据库，如果有特别需要可以选别的数据库。点击BLAST之后，程序就会把序列提交给BLAST服务器，服务器会返回给用户一个requestID，用于搜索BLAST 结果。输入ID，点击相应按钮，浏览器就会弹出新的窗口显示BLAST结果。

BLAST结果通常包括三个部分：

* 序列在数据库中搜索得到的Hits分布图，把用户序列在数据库中能找到的匹配序列从高到底用不同颜色的线条图形表示出来。
* 在数据库中检索到的匹配序列的从高到低的排序，BLAST显示所有E值小于设定的E值的匹配序列。这个部分包括三个数据：

------第一列：序列的名称

------第二列：Score（分值），用户序列和数据库序列中命中序列比对之后的相似性打分，打分越高则序列相似性越高。

------第三列：E值，是随机产生一个比所得分值高的对位排列的概率。是衡量分值可靠性的测度，E值越小, 所得相似序列的可靠性越高。

如果用户有大量的数据，则需要将BLAST程序下载到用户本地的计算机中，进行本地化分析。

**直系同源 / 并系同源**

直系同源基因(orthologous gene)是指在不同物种之间同源相似的基因（不同的物种中来源于共同的祖先），而并系同源基因(paralogous gene)是指一个物种内的同源基因（同一物种内由于基因复制而分离的同源基因）。

**双向BLAST:**用物种A的基因序列搜索(BLAST)物种B的基因序列，搜到的最高相似的序列反过来再BLAST物种A的基因组(阈值E小于1 e一20)。如果两条序列在这两个BLAST中均是最相似的，称为“双向BLAST最高相似蛋白”，则认为是直系同源关系。

**Mega-Blast：**可用于搜索近似完全的匹配, 可以处理一批核苷酸查询，比标准BLAST查询速度快，NCBI进行**基因组**BLAST查询时的默认程序。

**PSI-BLAST：位置特定的迭代 BLAST(Position Specific Iterated BLAST)，**搜索数据库以找出与查询序列同一蛋白质家族的成员，揭示亲缘关系较远的蛋白质间的关系。

**两个序列间的查询**

* 比较两个序列的相似性，不需要传统BLAST的数据库查询
* **BLAST2**
* 局部对位排列，获得结构域或序列内重复信息
* 建议不超过150 kb

**从GenBank等分子数据库中检索收录号为AF166093的DNA序列, 用BLAST等工具对该序列进行同源性搜索, 用所获得的序列构建进化树。**

**Flow：**

1.登录NCBI主页-点击BLAST-点击TBLASTX-在Search文本框中粘贴检测序列-点击BLAST!-点击Format-得到result of BLAST。

2．Mega建树   
构建进化树的主要步骤是比对，建立取代模型，建立进化树以及进化树评估。  
(1) 首先用Clustal X对上述检测到的序列进行比对.并将开始和末尾处长短不同的序列剪切整齐,得到目标同源序列。

(2)采用mega软件，建立系统树。

2.背景知识1：干扰素（Interferon，IFN）是一个具有多种生物学功能的蛋白质家族，具有抗病毒、细胞生长抑制和免疫调节的作用。目前，将IFN分为2个亚型：I和II，其中I型基因分为α、β、ω、κ、δ和τ。对所有哺乳动物研究发现，α干扰素家族成员均没有内含子，而且序列相似性高。人类IFN-α2的序列号是NM\_000605。

背景知识2：直系同源基因是指在物种形成过程中从祖先物种“继承”的基因，其在结构和功能等方面都具有保守性；并系同源基因是存在于同一个基因组中的同源基因。可以根据并系同源基因起源时间与物种形成时间的关系，进一步划分并系同源基因，其中“inparalog”基因特指那些物种形成之后形成的同源基因。

题目：

根据上述背景知识，用BLAST工具搜索α-干扰素基因相关序列，并据此构建哺乳动物α-干扰素基因的进化树。此外，根据并系同源基因的概念讨论α-干扰素基因的进化模式。

（提示：物种分类可以参考在线BLAST结果中的Taxonomy report）（2005）

**解答：**

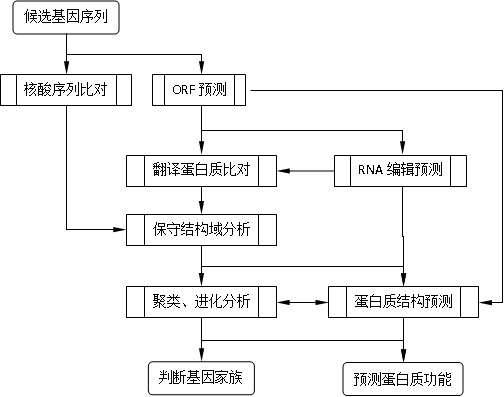
略

3.现在获得一个基因，试用生物信息学的知识对其功能，结构，来源等方面进行研究。（2007）

**解答：**

同4

4.从某个新近测序的基因组中发现了一个候选基因，需要进一步预测其功能并鉴定其是否为某个基因家族的成员。试用框图给出相关的生物信息学方法及步骤。（2008）



核酸序列比对：BLASTn，tBLASTx

ORF预测：ORF Finder，对找到的ORF进行评价（TATA盒，启动子）

RNA编辑预测：intro和exon预测，选择性剪切，RNA编辑（多框阅读）

翻译蛋白质比对：BLASTx，tBLASTn，tBLASTx

保守结构域分析：保守区和可变区预测，特征Motif，保守域序列预测

蛋白质结构预测：二级结构预测，三级结构预测（从头预测，同源预测）

聚类、进化分析：基因家族的判定

蛋白质功能预测：同源类比，虚拟筛选或模拟生物过程

6.请列举目前较为可靠的基因功能预测的方法，并分别说明其优势和局限性。（2011）

7.已知未知种类的鱼的全基因组序列，通过哪些生物信息学方法来研究生物的特性和功能？为什么？（2013）

贝叶斯公式



不及格计算

|  |  |  |  |
| --- | --- | --- | --- |
| 研究所 | 不及格率P(A|Br) | 学生比例  P(Br) | 不及格学生所在所的概率P(Br|A) |
| 生化所 | 0.05 | 0.33 | 0.45 |
| 神经所 | 0.04 | 0.17 | 0.18 |
| 营养所 | 0.025 | 0.13 | 0.09 |
| 健康所 | 0.05 | 0.13 | 0.18 |
| 药物所 | 0.02 | 0.17 | 0.09 |
| 计算所 | 0 | 0.07 | 0 |

对于生化所

看上面的贝叶斯公式

P(Br|A)=0.45267=0.05\*0.33/(0.05\*0.33+0.04\*0.17+0.025\*0.13+0.05\*0.13+0.02\*0.17+0\*0.07)

（2017 严军）CpG岛，在基因中存在的概率20%，正常基因CG比例为50%，给一段序列，用贝叶斯方程算出是CpG岛的概率

**解答：**

关键词：[CpG岛](https://www.baidu.com/s?wd=CpG%E5%B2%9B&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao" \t "/Users/brian/Documents\\x/_blank)（CpG islands）是指DNA上一个区域，此区域含有大量相联的胞嘧啶（C）、鸟嘌呤（G），以及使两者相连的磷酸酯键（p）。CpG岛的GC含量大于50%，长度500~1000bp。

P(CpG)=0.2, P(CG )=0.5 求 P(CpG | 序列)

由Bayes Equation : P(B|A) = P(B)P(A|B)/P(A)

假设已知 P(CG|CpG) 由公式直接得到

P(CpG | 某某CG含量的序列) = P(CpG)P(CG|CpG)/P(CG含量)

= P(CpG序列概率） P(CG在CpG序列的含量） / P(CG在所有基因中含量）

= 0.2P(CG | CpG)/0.5



参见多多的资料