维普资讯 http://www.cqvip.com

文章编号:0253-3626(2001)03-0355-04

通过 INTERNET 获得人类基因组序列(草图)资源

谢丹玫,赵文龙

(重庆医科大学图书馆 重庆 400016)

【摘 要】目的:介绍通过 INTERNET 获得人类基因组资源的方法。方法:利用网络信息资料和 INTERNET 技术获得人类基因组序列信息。结果:可以获得正常的及与疾病有关人类基因序列及鉴别未知基因。结论:通过 INTERNET 免费获得人类基因组序列信息是廉价、快速而有效的方法。

【关键词】人类基因组;INTERNET;资源

【中国图书分类法分类号】G203 【文献标识码】B 【收稿日期】2000-08-15

Getting Human Genome Sequence(Draft)Resource via Internet

XIE Dan-mei et al

(Library of Chongqing Medical University)

【Abstract】 ***Objective***: The method getting human genome sequence resource via internet was introduced. ***Methods***: Using network information resource and internet techniques to get human genome sequence information. ***Results***: The information about normal and disease-related human genome sequences could be got, and unknown gene could be identified. ***Conclusion***: Getting human genome sequence information is a cheap, fast and efficient method.

【Key words】Human genome; INTERNET; Resource

人类基因组序列(Human genome sequence)是人类共同财富,本文简介人类基因组序列草图的意义及如何通过 INTERNET 获得人类基因组序列草图资源,以服务于科研和医疗[1]。

1 基因组序列草图的产生

现代的 DNA(脱氧核糖核酸)测序技术仅仅只能确定 DNA 短片(低于 1000 碱基对)的准确序列,因为人类基因(HG)的长度超过三十亿碱基对,所以基因组必须分成很多的小片排序,然后再拼装在一起。重新拼装的方法是比较每小节的末端排列以寻找能将它们连接在一起的相互嵌合部分。这项工程相当复杂,为了使得出的草图更为准确,DNA 短片必须取自全世界,它必须由组成人类基因组(HGP)国际集团的测序实验室通过国际合作来完成。

人类基因组已经完成了第一个人类基因草图,之所以说它是"草图",现在还认为它不是最精确的,但它却给科学家提供人类基因序列的概貌。本草图序列的产生很象是拼装一个复杂的、巨型的七巧板,必须将数百万 DNA 的短片拼装在一起以形成完整的基因组的总的序列。

由于持续跟踪成千上万的组成 HG 拼图的 DNA 片断是非常困难的,当测序必须进一步得到精确,其错误由世界范围内的测序实验室进行校正时,基因的序列是会改变的,要人工完成这项任务是极其困难的。由 EMBL-EBL 和 Sanger 中心联合开发的一个能自动跟踪 HG 的全部已排序片断,而且能够将其拼装成一个大的单一段,然后分析已拼装的 DNA 以寻找令生物医学研究人员感兴趣的基因和其它特征。通过 Ensembl(人类基因组序列数据库)这样的系统进行自动跟踪将是又快、又准确、又省时省力。

Ensembl 是怎样预测基因的呢? 它使用专门的叫做"Genscan"的基因寻找软件以预测基因序列的

作者介绍:谢丹玫,(1968-),女,馆员,四川邻县人。主要研究方向:情报咨询、计算机管理。

维普资讯 http://www.cqvip.com

位置,该软件研究 DNA 序列,识别 DNA 看似象基因的区域,然后将这些候选基因序列与公用数据库中的全部已知序列比较。如果发现匹配,则提示预测精确到某种程度的"匹配证据"。然后将已预测的基因存放于 Ensembl 数据库中以备免费检索。

为了使世界上各个国家的实验室里的科学家确信他们所使用的是同一基因,保持数据库中某些事物不改变是很重要的(如基因的名字),Ensembl 采取了重大的举措试图保持基因的基因和其他特征有不变的名字。这是非常困难的任务,因为 Ensembl 是在这样的环境中:DNA 序列和其他环境的不断变化,基本的 DNA 序列的改变可能产生新的基因,原来的基因被删除、替换或与另外的基因合并。只要有可能,Ensembl 就试图在 DNA 序列被订正后而保持名字的不变并且将保持一个版本号使其改变能够被完全跟踪。

第一个人类基因组工作草图已经提前完成。Wellcome 托拉斯总裁称这一成就"不仅是我们这一时代的杰出成就,而且也是人类历史上的杰出成就"。

这一草图绘出 97% 的人类基因组,其 85% 基因组序列的准确度达 99.9%。草图包括 3.8 万个已肯定的基因及 11.5 万个可能的基因。前不久,两组竞争对手联手在美国首都华盛顿和英国伦敦宣布这一成就[2]。

2 基因组序列草图的意义

Nature 网络版编辑 MacFarlane[3] 等认为,基因组序列对下列研究工作有益:

* 在特殊的模型动物研究中,基因组序列使人们能有把握地进行实验设计,确定生物学途径,无须测序便可确定变异体。

* 对于物种进化研究,《比较基因组学》将为物种关系的研究提供迄今为止最全面的资料。

* 在细胞生物学方面,人们可以彻底研究构成细胞的各成分之间是如何相互作用的以及环境状态是如何影响这些成分的。

* 在医学领域,基因组序列提供了研究疾病易感性、发病机制的基础,同时,也为开发新的防治方法奠定了基础。

* 基因组以后的研究规划将包括结构基因组学,开发蛋白芯片以研究蛋白 - 蛋白之间的相互作用。

总之,几乎生物学和生物医学的所有领域,以及它们之间的相互作用都会受到基因组学的影响。

目前,癌基因项目组正在系统地将草图中的基因序列与癌细胞的基因进行比较,基因方面的差别有可能成为探索癌症发生的生物学机制的线索,可能有助于寻找新的诊断手段和新的治疗方法。

3 基因组草图是人类的共同财富

人类基因组项目是在 8 个公立大学科研中心进行的国际合作项目。该项目自 1990 年开始绘制基因图。1998 年,由 Venter 投资成立的美国生物技术公司 Celera 宣称自己具备更快速的制图技术,并称可于 2000 年 6 月将全部基因组测序完毕。

代表众多科学家态度的英国剑桥大学 Sanger 中心主任 Sulston 博士(该中心负责提供此项目 1/3 的基因测序数据)主张基因信息不能落到某个单位或某个企业的控制之下。

人类基因组项目已经把其基因草图公布在互联网上,美国前总统克林顿和英国首相布莱尔主张将基因测序数据作为全球资源保留在公共范围内。在基因草图能否使人们迅速找到许多疾病的治疗办法方面尚需作很多的工作。

4 如何在 INTERNET 上免费查找基因组草图

我们可以通过任何一台同 INTERNET 联网的计算机上免费获得基因组草图。该网站地址是 www.ensembl.org[1],该网站提供如下服务:

①它提供了快速而容易浏览的 Ensembl 数据库内容或寻找其感兴趣的特殊项目的方法;②有许多连接到 Ensembl 数据库的入口点;③DNA 相似性搜索("BLAST"搜索),如果你已经有一个 DNA 或蛋白质,并且你想知道在 Ensembl 数据库中是否存在类似物;④Ensembl 标识符搜索,如果你已经知道 ID 号,你能够直接检索得到它;⑤可以从染色体水平浏览到 DNA 序列水平。

4.1 Ensembl 主页界面(见图 1)

4.2 从这个 Ensembl 网站中,可以获得如下信息:

4.2.1 染色体视图表示出一对染色体图形和一系列已被映射到该染色体的"标记"。

4.2.2 单击不同的染色体就能进入不同的描图。

维普资讯 http://www.cqvip.com

单击右面一系列蓝色标记就能获得那一点的 DNA 序列。

4.2.3 解释标记信息 查看标记序列,查看无图的 DNA 序列,查看该区域序列片段描图,存储标记的数据库,标记两边的序列。

4.2.4 序列显示 单序列重复,序列长度,正向双螺旋特征,与其他已知序列的同源序列,基因。

4.2.5 序列描图显示 除序列显示外,DNA 片段描图显示基因的位置,单击图上的基因位置,当前序列显示为图中高亮度的线框,在图块中点击就能改变序列显示。在标记显示图的克隆或描图链上点击就可以获得该序列的显示。

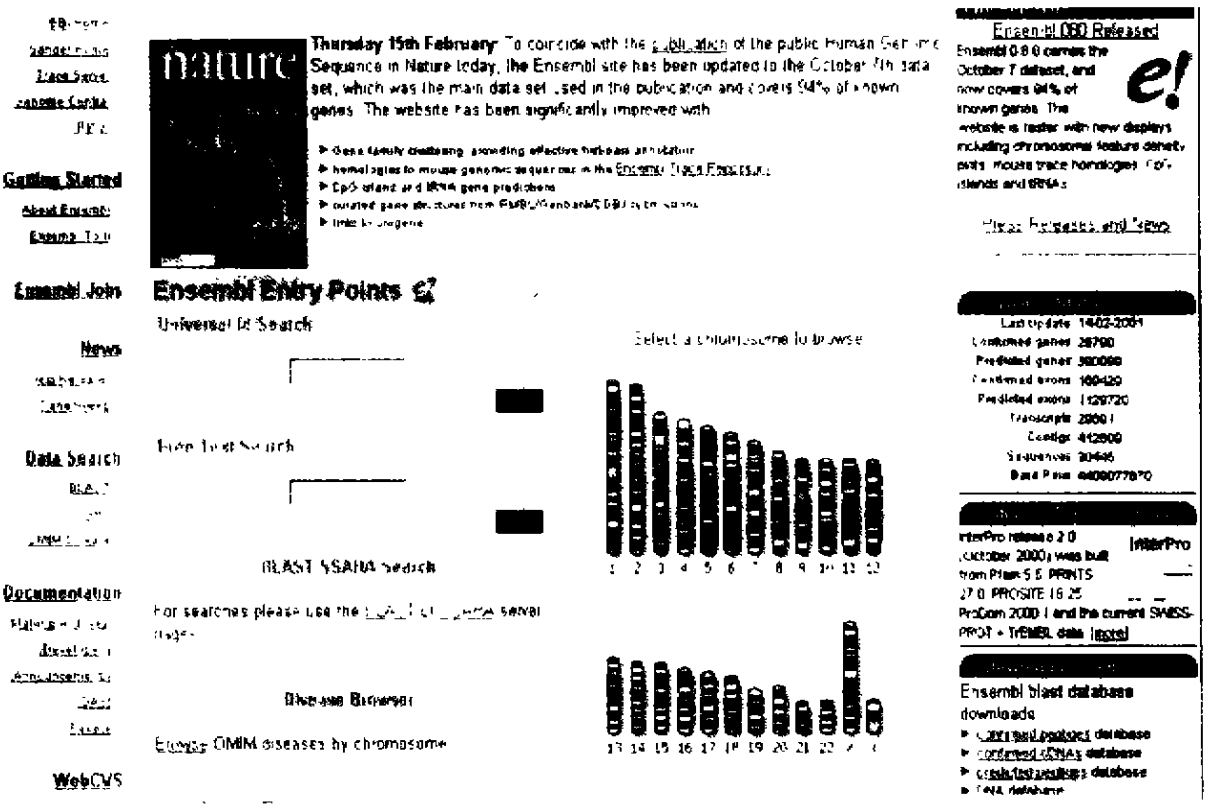


图 1 Ensembl 主页

4.2.6 基因视图 已预测的性质,蛋白质产品序列,基因结构,导致预测该基因的"支持证据"。

4.2.7 基因的支持证据 以灰色显示低可靠性证据,支持按其可靠性排序的基因的预测证据。可靠数据以绿色显示,支持数据源,支持数据 ID,该证据支持的基因预测部分的图解表示,数据的摘要。

4.3 Ensembl 提供的搜索方式

Ensembl 主页提供几种搜索方式 ①浏览染色体(Select a chromosome to browse);②通用标识符搜索(Universal ID Search);③自由词搜索(Free Text Search);④BLAST/SSAHA 搜索(BLAST/SSAHA Search);⑤按疾病方式浏览(Disease Browser)。

4.3.1 浏览染色体(Select a chromosome browse) 点击任意染色体就能直接进入染色体视图,左边显示带有特征密度图(FDP)的单个染色体的表意符号(表示基因、复制、SNP 分布)以及序列装配(Golfrn Path 表示),并以 FISH(荧光原位杂交)和 RH(辐射杂合子)标记左对齐,灰色区域表示着丝点或终聚体上的高度重复的异染色质。

4.3.2 通用标识符(Univernal ID Search)

通用标识符搜索有三种途径:①在主页中的"Univernal Id Search"对话框中输入要查找的标识符;②点击[Date Search]输入要查找的标识符;③在每页的顶部的搜索框中输入要查找的标识符。标识符类型包括 chromosome, clone, contig, disease, domainentry, gene, marker, snp, transcript 等。

Ensembl 数据库检索系统支持通配符检索,用

维普资讯 http://www.cqvip.com

符号“*”。如以“disease”为检索途径,在检索提供框中输入“leuk*”,系统将会检索出数据库中包括“Adrenoleukodystrophy”,“Interleukin - 2 receptor deficiency”,“Leukemia”等含有“leuk”字符的内容。查询中可含一个以上的通配符,但这样的查询要花很长的时间,可能返回一些无用的结果,查询应尽可能具体一些。

4.3.3 自由词搜索(Free Text Search)

Ensembl 自由文本搜索能够检索到包括 OMIM, Swissport 和 Interpro 数据库中的文摘信息。支持通用的检索语法“AND”、“OR”、“AND NOT”、“+”、“-”、“NEAR”和“WITHIN”等。

4.3.4 BLAST/SSAHA 搜索(BLAST/SSAHA Search):

4.3.4.1 BLAST 搜索

Ensembl 支持直接浏览和以 E-mail 方式进行碱基序列相似性检索。用户将需要查对的碱基序列通过系统提供检索界面提交给系统,系统完成检索后会自动或以 E-mail 方式返回结果。步骤如下:

(1)点击[BLAST]进入检索页面。

(2)将数据粘贴到 DNA 序列输入框,或直接调入存放在本地序列的文件(在序列输入框的下方有〔浏览〕,点击〔浏览〕按钮,系统会弹出文件选择对话框,选择存放碱基序列的文件,打开即可)。

(3)确定结果返回方式。直接浏览(browser)或 E-mail 返回,如果是以邮件方式返回,则在 Email to 后面的空白框中填写 E-mail 地址。

(4)选择查对的数据库(database)。包括 Ensembl confirmed cDNAs, Ensembl confirmed peptides, Ensembl predicted peptides, Ensembl predicted cDNAs 等。

(5)确定检索程序(excutable)。这个下拉菜单提供适当的搜索程序:①blastn: DNA 数据库的 DNA 查询序列;②blastp:蛋白质数据库的蛋白质查询序列;③blastx:蛋白质数据库的 DNA 查询序列(全部 6 种结构的虚拟转录);④tblastn: DNA 数据库的蛋白质查询序列(全部 6 种结构的虚拟转录);⑤tblastx: DNA 数据库的 DNA 查询序列(全部 6 种结构的两种虚拟转录)。

(6)确定返回结果条数(report)。Alignments 表示检索结果中高积分片断配对(High-scoring Segment Pairs, HSPs)的数据库序列的条数,默认值为 100。

(7)其他参数设置。BLAST 搜索设有多种参数供选择,一般情况下选择默认值。

Matrix(矩阵):为 BLASTP, BLASTX, TBLASTN 和 TBLASTX 程序指定的一个评分矩阵,默认值为 blosum62。

Description(描述):该参数说明程序将打印出的评分数(描述行数)。默认值是 100 行。

Expect(期望值):该参数说明选中数据库的统计意义阈值。默认值为 10。

HSP score:用以估计多重 HSP 的统计意义。

Filter type:过滤器类型。

(8)完成上述选择后,点击左上方的〔search〕按钮。

4.3.4.2 SSAHA 序列搜索

SSAHA 是 DNA 序列快速匹配和校验的软件工具,由 Hashing 算法完成序列搜索和校验。SSAHA 算法最适用于两序列间需精确或几乎精确的匹配的应用领域,该算法不检查短于 Hash 字长(默认值为 14 对碱基)的相继匹配的碱基的伸展度。

检索步骤如下:

(1)点击[SSAHA]进入检索页面。

(2)在编辑框中输入单个或多个 FATSA 序列,或直接调入存放在本地文件(在序列输入框的下方有〔浏览〕,点击〔浏览〕按钮,系统会弹出文件选择对话框,选择存放本地的文件名,打开即可)。

(3)参数设置

使用最小长度匹配截断值以仅显示多于给定长度的匹配数,选择要搜索的数据库包括最新的人类“golden path”数据库和 Ensembl Trace Repository。

(4)完成上述选择后,点击左上方的〔search〕按钮

4.3.5 OMIM 疾病(OMIMDisease):

在主页中点击[OMIM Disease]能浏览 OMIM 病种及其相关的标识符及定位信息,也可由染色体浏览病种或浏览搜索结果。

参 考 文 献

[1] http://www.ensembl.org.

[2] Yarney, G. Scientists unveil first draft of human genome [J]. BMJ, 2000; 321: 7

[3] http://www.nature.com.