网络出版时间：2015-06-26 16:39
网络出版地址：http://www.cnki.net/kcms/detail/11.2396.Q.20150626.1639.002.html

编者按：自2000年起，罗静初教授在北京大学生命科学学院和中国农业科学院研究生院开设“实用生物信息技术”课程，教学方法以上机实习为主，指导学生利用丰富的生物信息网络数据资源和软件工具，结合自己的研究课题，进行生物信息分析。本文以血红蛋白为例，介绍常用生物信息技术与方法的具体应用，对生物信息领域科研人员具有一定参考价值。

实用生物信息技术课程教学实例

罗静初
（北京大学生命科学学院 北京大学蛋白质与植物基因研究重点实验室 北京大学生物信息中心，北京 100871）

摘　要： 介绍“实用生物信息技术”研究生课程的5个教学实例。以血红蛋白序列和结构为例，介绍常用生物信息技术和分析方法，包括蛋白质和核酸序列相似性比对、蛋白质和核酸序列数据库检索、Blast数据库相似性搜索、分子系统发生树构建，以及蛋白质结构比较分析等。

关键词： 生物信息技术；血红蛋白；序列比对；数据库检索；Blast数据库搜索；系统发生树构建；蛋白质结构比较分析
DOI：10.13560/j.cnki.biotech.bull.1985.2015.07.001

Teaching Examples of Applied Bioinformatics Course

Luo Jingchu
(*College of Life Sciences*, *The State Key Laboratory of Protein and Plant Gene Research*, *Center for Bioinformatics*, *Peking University*, *Beijing* 100871)

Abstract: In this article, we introduce the basic bioinformatics analysis methods and tools taking the hemoglobin as an example. The methods include: 1) protein and DNA sequence alignment; 2) advanced search for UniProt and RefSeq database; 3) Blast database similarity search; 4) phylogenetic tree construction under MEGA; 5) protein structure comparison using Swiss-PdbViewer.

Key words: bioinformatics; hemoglobin; sequence alignment; database query; Blast database similarity search; phylogenetic tree construction; protein structure comparison

自2000年起，本人在北京大学和中国农业科学院研究生院开设“实用生物信息技术”课程[1]。本课程以从事分子生物学实验研究的硕士或博士研究生为教学对象，重点介绍最基本、最常用的生物信息技术和方法，主要包括：(1) 蛋白质和核酸序列相似性比对；(2) 蛋白质序列数据库UniProt和核酸序列数据库RefSeq高级检索；(3) NCBI数据库相似性搜索工具Blast的应用；(4) 利用MEGA软件构建分子系统发生树；(5) 利用Swiss-PdbViewer软件显示、比较和分析蛋白质三维空间结构。

本文以人、小鼠、大鼠、斑头雁、灰雁几个不同物种的血红蛋白序列和结构为例，介绍这些常用生物信息技术和方法的具体应用。学生通过这些实例，能够初步掌握这些方法的具体应用，并能举一反三，将这些方法用于自己的课题研究，学会如何利用丰富的网络生物信息资源和分析工具解决自己正在进行或即将开始的研究课题中的实际问题。

1　序列比对

1.1　研究背景

血红蛋白是人体血液中重要蛋白质分子，其主

收稿日期：2015-03-26
作者简介：罗静初，男，教授，研究方向：生物信息学；E-mail：luojc@pku.edu.cn

要生物学功能为运送氧气。血红蛋白分子为异源四聚体，可结合 4 个铁卟啉色素分子。成人血红蛋白分子由两个 α-亚基和两个 β-亚基组成。人类基因组中编码 α-亚基的血红蛋白基因有两个，位于 16 号染色体短臂的 α-珠蛋白基因簇中，其编码区核苷酸序列相同，所编码的蛋白质序列自然也相同，各含 142 个氨基酸残基。与人一样，小鼠和大鼠的血红蛋白也是四聚体，α-亚基也由 142 个氨基酸组成。小鼠和大鼠同属啮齿类动物，其共同祖先距今约 2 500 万年。而人属于灵长类动物，与啮齿类分歧时间约为 9 500 万年。对这 3 个物种 α-血红蛋白氨基酸序列及其编码基因的核苷酸序列进行比对，可探索血红蛋白分子及其编码基因演化的特点。

1.2 比对方法和结果

从国际蛋白质序列数据库 UniProt 中分别提取人和小鼠 α-血红蛋白的 FastA 格式序列，其序列条目名称分别为 HBA_HUMAN（人）、HBA_MOUSE（小鼠）。序列比对的软件很多，北京大学生物信息中心开发的综合序列分析平台 WebLab（http://weblab.cbi.pku.edu.cn/）包括 200 多个程序[2]。利用 WebLab 中基于 Needleman-Wunsch 全局序列比对算法的程序 Needle，采用默认蛋白质计分矩阵 BLOSUM62 和默认空位罚分值（起始空位罚分 10.0，延伸空位罚分 0.5），比对结果如图 1 所示。

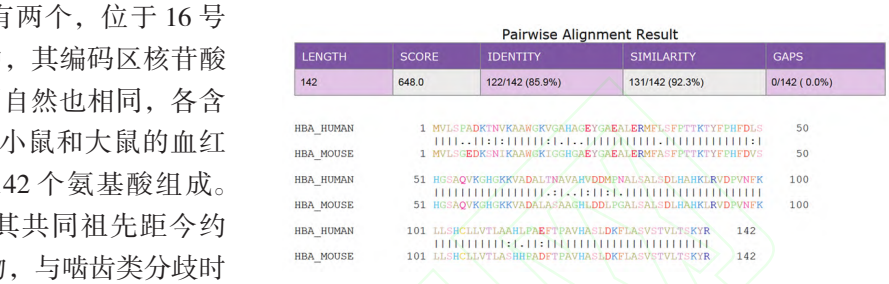

图 1 人和小鼠血红蛋白 α-亚基氨基酸序列比对输出结果

图 1 中上方为统计值，包括序列长度（LENGTH）、比对分值（SCORE）、相同位点（IDENTITY）、相似位点（SIMILARITY）和空位数（GAPS）。下方为两条序列的具体比对结果，“|”表示相同位点、“:”表示相似位点，“.”表示不同位点。所谓相似位点，是指该位点的两个氨基酸理化性质较接近，如苏氨酸“T”和丝氨酸“S”、缬氨酸“V”和异亮氨酸“I”等。

按上述方法，分别对人/小鼠、人/大鼠、小鼠/大鼠 3 个物种 α-血红蛋白进行序列比对，结果如表 1 所示。

表 1 人、小鼠、大鼠血红蛋白 α-亚基氨基酸序列比对结果

物种	序列名	登录号	得分	相同位点（比例）[a]	相同加相似位点（比例）[b]
人/小鼠	HBA_HUMAN / HBA_MOUSE	P69905 / P01942	648	122/142（85.9%）	131/142（92.3%）
人/大鼠	HBA_HUMAN / HBA_RAT	P69905 / P01946	587	111/142（78.2%）	120/142（84.5%）
小鼠/大鼠	HBA_MOUSE / HBA_RAT	P01942 / P01946	632	120/142（84.5%）	127/142（89.4%）

注：a：相同位点个数占全长序列的比例；b：相同加相似位点个数占全长序列的比例

从 NCBI 参考序列数据库中提取这 3 个物种 α-血红蛋白基因编码区序列，用 WebLab 中的 Needle 程序进行序列比对，注意选择核苷酸替换矩阵 EDNAFULL，将起始空位罚分改为 20.0，延伸空位罚分改为 2.0，比对结果如表 2 所示。

1.3 结果分析

表 1 为 3 个物种血红蛋白 α-亚基氨基酸序列比对结果。出乎意料的是，人和小鼠 α-血红蛋白共有 122 个相同位点，占全长 142 个位点的 85.9%；而小鼠与大鼠之间的相同位点数为 120 个，占全长 84.5%。换句话说，同为啮齿类的小鼠和大鼠，血红蛋白序列相似性低于啮齿类和灵长类。之所以出现这一结果，原因有许多，其中最主要的是密码子简并性，即同一氨基酸在不同物种或不同基因中可能由不同密码子编码，蛋白质序列相似性高低可能与其编码核苷酸的相似性高低并不一致。这 3 个物种血红蛋白编码基因的编码区核苷酸序列比对结果（表 2）显示，小鼠和大鼠之间的序列相似性为 89.3%，高于小鼠和人之间的序列相似性 81.6%。

2 数据库高级检索

2.1 研究背景

研究表明，有些基因在一个物种中只有一个拷

表 2　人、小鼠、大鼠 α-珠蛋白编码区核苷酸序列比对结果

物种	基因名	登录号	得分	相同位点（比例）[a]	空位数
人 / 小鼠	HsHBA1 / MmHba-a1	NM_000558 / NM_008218	1434	350/429（81.6%）	0
人 / 大鼠	HsHBA1 / RnHba1	NM_000558 / NM_013096	1335	339/429（79.0%）	0
小鼠 / 大鼠	MmHba-a1 / RnHba1	NM_008218 / NM_013096	1731	383/429（89.3%）	0

注：a：相同位点个数占全长序列的比例

贝，称单拷贝基因，而真核生物基因组中大部分基因按基因家族形式存在，有多个拷贝，它们或者分布在同一染色体上相邻区域，或者分散在整个基因组不同染色体上。基因家族的产生包括全基因组水平重复和染色体片段重复等多种机制，是生物演化重要途径。同一家族的基因往往具有相似生物学功能，通过复杂的调控机制，在不同组织、不同环境或不同发育阶段表达。例如，无脊椎动物的血红蛋白由一个基因编码，而脊椎动物的血红蛋白则由多个基因编码。以人血红蛋白基因家族为例，分为 α-珠蛋白（α-globin）和 β-珠蛋白（β-globin）两个基因簇，如图 2 所示。图 2 上方为 α-珠蛋白基因簇，位于 16 号染色体短臂正链 150–180 kb 区段，全长约 30 kb，按 5′–3′ 顺序依次为 ζ、μ、α_2、α_1 和 θ-珠蛋白基因。下方为 β-珠蛋白基因簇，位于 11 号染色体短臂互补链 5.22–5.27 Mb 区段，全长约 50 kb，依次为 ε、γ_2、γ_1、δ 和 β-珠蛋白基因。此外，α-珠蛋白基因簇上有两个假基因 Ψ_ζ 和 Ψ_α；β-珠蛋白基因簇上有 1 个假基因 Ψ_β。这 10 个珠蛋白基因在不同发育阶段表达，θ、μ、β 和 δ 在成人血红细胞中表达，γ_1 和 γ_2 在胎儿血红细胞中表达，ζ 和 ε 在胚胎血红细胞中表达，而 α_1 和 α_2 在成人和胎儿血红细胞中均表达。

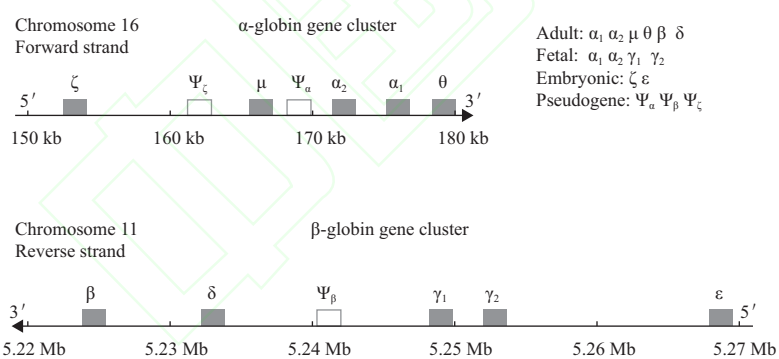

图 2　人 α-珠蛋白和 β-珠蛋白基因家族染色体定位

2.2　*检索方法*

上述 α 和 β-珠蛋白基因编码的血红蛋白氨基酸序列，均存放在国际蛋白质序列数据库 UniProt 中。利用该数据库提供的高级检索功能，可以快速有效地检索到这些蛋白质序列条目。具体检索步骤如下：

（1）点击 UniProt 数据库主页上方检索框右侧 Advanced 下拉式菜单，打开弹出式高级检索子窗口（图 3-A）。

（2）点击高级检索窗口最上方下拉式选择菜单中的 All，选择 Protein Name［DE］，在其右侧的文本输入框中输入血红蛋白的英文 Hemoglobin。

（3）点击第 2 个下拉式选择菜单中的 All，选择基因名 Gene Name［GN］，在其右侧的文本输入框中输入血红蛋白的基因名缩写 hb（不分大小写），并在后加通配符星号，即 hb*。

（4）点击该选择菜单输入框右侧增加选择项符号“+”，弹出第 3 个选择菜单（图 3-B）。

（5）点击第 3 个选择菜单中的 All，选择物种名 Organism［OS］，在其右侧输入 Human，系统列出该数据库中与输入文本 Human 相关的所有物种，选择 Human［9606］。9606 为人在 NCBI 分类学数据库中的登录号。

（6）点击检索窗口右下侧检索按钮（图标为放大镜），提交检索策略，页面显示 UniProt 数据库中收录的所有人血红蛋白序列条目。

（7）点击页面左侧 Reviewed 图标，页面显示检索结果（图 3-C）。

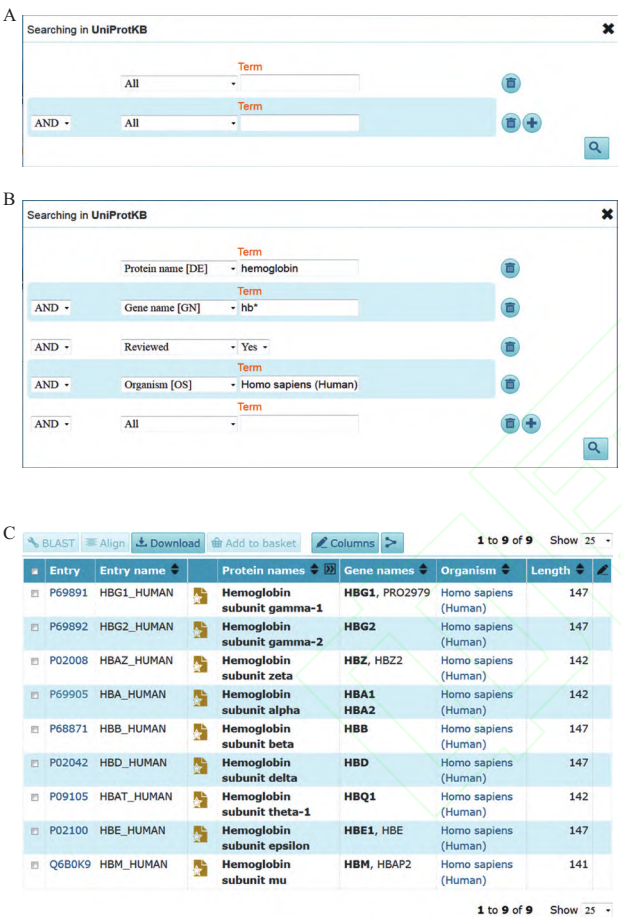

图 3 利用 UniProt 高级检索界面检索人血红蛋白 9 个序列条目

2.3 检索结果

检索结果中列出已经通过人工审阅的 9 个人血红蛋白序列条目。UniProt 数据库包括 Swiss-Prot 和 TrEMBL 两个子库，其中 Swiss-Prot 中的序列条目均已经通过人工审阅，而 TrEMBL 中的序列条目则是利用计算机对核酸序列数据库 EMBL 中的蛋白质编码序列翻译得到的，未经人工审阅。截止 2015 年 3 月，Swiss-Prot 子库中的数据条目总数为 547 599 条，而 TrEMBL 子库中的数据条目总数为 90 860 905 条。显然，这两个子库的数据量差别极大。点击 UniProt 网站主页面下方 UniProt data 栏目下的 Statistics 图标，可以找到这两个子库的统计资料文档 UniProt/Swiss-Prot statistics 和 UniProt/TrEMBL statistics，文档中有许多图表，详细叙述这两个子库的基本情况。

3 数据库序列相似性搜索

3.1 研究背景

利用上述蛋白质序列数据库高级检索方法，可以快速高效地找到人血红蛋白基因家族 9 个成员所编码的蛋白质序列。近年来发现，除了运送氧气的血红蛋白和储存氧气的肌红蛋白外，人体中还有另外两种珠蛋白分子，一种为细胞红蛋白，或简称胞红蛋白（Cytoglobin），普遍存在于各种组织，可能具有氧储存、氧感受、一氧化氮运输、抗自由基等多种功能。另一种为神经红蛋白（Neuroglobin），多见于脑组织，因此也称脑红蛋白。胞红蛋白基因位于 17 号染色体长臂 25 区（17q25），编码 190 个氨基酸残基；脑红蛋白基因位于 14 号染色体长臂 24 区（14q24），编码 151 个氨基酸残基。X 衍射晶体结构研究证明，这两种蛋白质分子的三维空间结构与血红蛋白、肌红蛋白具有相同折叠模式，同属珠蛋白家族（Globin family）。序列比对发现，两者与血红蛋白序列相似性均很低。胞红蛋白与血红蛋白 α-亚基的相同位点共 42 个，约占 22%；脑胞红蛋白与血红蛋白 α-亚基的相同位点仅 31 个，不到 20%。

3.2 搜索方法

利用 BLAST 数据库相似性搜索，可以通过局部序列比对方法，从数据库中找到相似性较高的序列或序列片段。例如，以人血红蛋白 α-亚基 HBA_HUMAN 为检测序列，可以从 Swiss-Prot 数据库中搜索到与其相似性较高的其它物种血红蛋白 α-亚基序列。而对于脑红蛋白这样相似性很低的序列，则需要通过选择搜索程序、确定搜索数据库、限制搜索物种、设置适当的搜索参数，才能搜索到。具体步骤如下：

（1）打开 NCBI BLAST 服务器主页面，在常用 BLAST 选择区（Basic BLAST）中选择蛋白质 BLAST（protein blast），将人血红蛋白 α-亚基 HBA_HUMAN

序列粘贴到检测序列输入框。

（2）在数据库选择框（Database）中选择 Swissprot protein sequence（swissprot），在物种选择框（Organism）中输入 Human。

（3）在程序选择区选择位点特异迭代型 BLAST（Position-specific Iterated BLAST），即 PSI-BLAST。

（4）打开参数选择（Algorithm parameters）窗口，将错误率（Expected threshold）由缺省值 10 调为 0.001。

（5）点击运行 BLAST 按钮递交作业，搜索结果得到 11 个珠蛋白分子。

（6）点击“运行第 2 次 PSI-Blast”（Run PSI-Blast iteration 2 with max 50）按钮（Go），新一轮搜索结果中包括脑红蛋白（Neuroglobin，Siwss-Prot 数据库登录号 Q9NPG2.1）。

3.3　搜索结果

搜索结果（图 4）显示，人 12 个珠蛋白均在搜索结果中，而与珠蛋白无关的其它序列则没有列在搜索结果中。也就是说，搜索结果既无假阳性（False positive）结果，也无假阴性（False negative）结果。

上述搜索过程说明，BLAST 是一个功能强大的序列相似性数据库搜索系统。但要用好 BLAST，必须对其基本算法有所了解，例如位置特异性迭代 BLAST 的原理、计分矩阵、错误率 E 值的选取等。

Select: All None Selected:0　　Yellow: sequences scoring below threshold on previous iteration

Alignments　Download　GenPept　Graphics　Distance tree of results　Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
RecName: Full=Hemoglobin subunit zeta; AltName: Full=HBAZ; AltName: Full=Hemoglobin zeta ch	195	195	100%	4e-64	60%	P02008.2	☑	✓
RecName: Full=Hemoglobin subunit delta; AltName: Full=Delta-globin; AltName: Full=Hemoglobin	189	189	97%	2e-61	42%	P02042.2	☑	✓
RecName: Full=Hemoglobin subunit alpha; AltName: Full=Alpha-globin; AltName: Full=Hemoglobir	188	188	100%	4e-61	100%	P69905.2	☑	✓
RecName: Full=Hemoglobin subunit gamma-2; AltName: Full=Gamma-2-globin; AltName: Full=Hb	187	187	97%	9e-61	39%	P69892.2	☑	✓
RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin b	187	187	97%	1e-60	42%	P68871.2	☑	✓
RecName: Full=Hemoglobin subunit gamma-1; AltName: Full=Gamma-1-globin; AltName: Full=Hb	185	185	97%	6e-60	39%	P69891.2	☑	✓
RecName: Full=Hemoglobin subunit theta-1; AltName: Full=Hemoglobin theta-1 chain; AltName: Fu	183	183	100%	3e-59	62%	P09105.2	☑	✓
RecName: Full=Hemoglobin subunit epsilon; AltName: Full=Epsilon-globin; AltName: Full=Hemoglo	181	181	97%	3e-58	37%	P02100.2	☑	✓
RecName: Full=Hemoglobin subunit mu; AltName: Full=Hemoglobin mu chain; AltName: Full=Mu-g	178	178	99%	3e-57	45%	Q6B0K9.1	☑	✓
RecName: Full=Myoglobin [Homo sapiens]	169	169	100%	1e-53	26%	P02144.2	☑	✓
RecName: Full=Cytoglobin; AltName: Full=Histoglobin; Short=HGb; AltName: Full=Stellate cell acti	164	164	96%	4e-51	28%	Q8WWM9.1	☑	✓
RecName: Full=Neuroglobin [Homo sapiens]	58.9	58.9	92%	1e-11	23%	Q9NPG2.1	☑	

图 4　利用 BLAST 从 Swiss-Prot 数据库中搜索 12 个珠蛋白

4　系统发生树构建

4.1　研究背景

研究表明，人、小鼠和大鼠 3 种哺乳动物中，均有血红蛋白、肌红蛋白、胞红蛋白和脑红蛋白 4 类珠蛋白基因家族成员，其中肌红蛋白、胞红蛋白和脑红蛋白在这 3 个物种基因组中均为单拷贝基因，而血红蛋白 α 和 β-两个亚家族均包含多个拷贝，在 3 个物种基因组中的数目、分布也不相同。美国宾夕法尼亚州立大学从事血红蛋白研究多年的哈迪森教授 2012 年发表的“血红蛋白及其基因的演化”综述中，对人和其它脊椎动物的血红蛋白起源、演化、表达和功能做了详细介绍[3]。图 5 是根据该论文中的插图改编的人、小鼠、大鼠 3 个物种基因组中 α-和β-珠蛋白基因家族成员名称和在染色体上的排列次序。

上述 3 个物种中，人类基因组的血红蛋白基因家族研究得比较清楚，而小鼠和大鼠血红蛋白的基因家族的大部分成员是根据基因组、转录组序列预测所得，尚无实验证据。表 3 列出这 3 个物种中已经确定的 37 个成员。

需要说明的是，小鼠脑红蛋白基因有两个剪接变体，RefSeq 参考序列数据库中 mRNA 序列登录号为 NM_022414 和 NM_001294308。NM_022414 编码区长度 453 bp，编码 151 个氨基酸；NM_001294308 编码区长度 465 bp，编码 155 个氨基酸。表中只

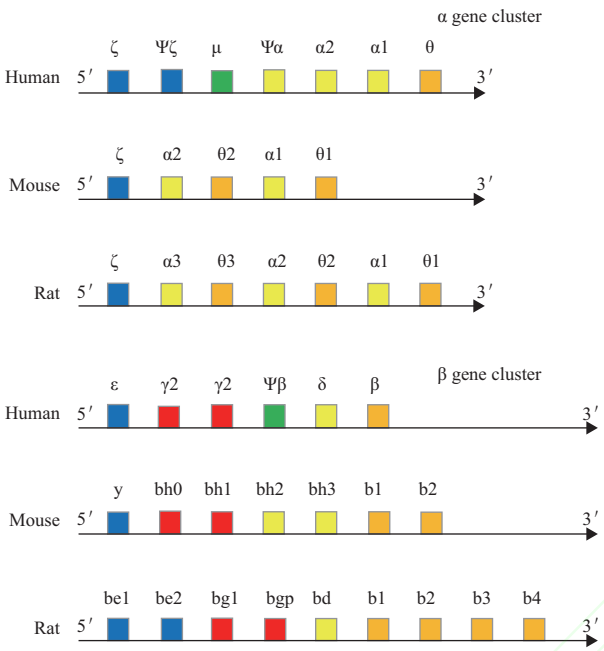

图 5 人、小鼠、大鼠 α-珠蛋白和 β-珠蛋白基因家族

列出其中一个 NM_022414。小鼠 β-珠蛋白家族成员 MmHbb-b1 和 MmHbb-b2 为单倍体型 C57BL/- 株系基因组中测得的序列，RefSeq 参考序列数据库中 mRNA 序列登录号为 NM_001278161 和 NM_016956。小鼠基因组计划测序样本所用的为融合体 BALB/c 和 129Sv 株系。小鼠基因组信息系统（MGI）中所列小鼠 β-珠蛋白家族成员基因名为 MmHbb-bs 和 MmHbb-bt，RefSeq mRAN 登录号为 NM_001201391 和 NM_008220，表中未予列出。

此外，大鼠基因组中 α-珠蛋白家族共有 7 个成员[3]，位于 10 号染色体 15.468–15.508 Mb 区段，长度约为 40 kb；表 8 中只收录已有转录数据的 3 个，即 RnHbz（NM_013096）、RnHba1（NM_013096）和 RnHba2（NM_001007722）。另 4 个尚无确切证据，也无确定的基因名，未在表中列出。这 4 个基因中，一个为 α-珠蛋白，RefSeq 数据库中 mRNA 序列登录号为 NM_001013853，大鼠基因组数据库 RGD 中暂定基因名为 LOC287167；其它 3 个为 θ-珠蛋白，尚无实验证据。大鼠基因组中，β-珠蛋白共有 9 个成员，位于 1 号染色体 175.095–175.170 Mb 区段，约 75 kb（图 6），其中 1 个为假基因，4 个为串联重复排列的 α-珠蛋白，推测由近期发生的基因倍增机制产生。

4.2 建树方法

利用上述 3 个物种基因组中的血红蛋白及同一家族的肌红蛋白、胞红蛋白和脑红蛋白序列信息，可以构建分子系统发生树。系统发生树是以树状图表示不同物种之间系统发生关系的常用方法。达尔文“物种起源”一书中唯一的一幅插图，就是用树的形式表示物种多样性及其起源和演化。因此，系统发生树，有时也称“进化树”或“演化树”。其实，系统发生树不仅可以用来表示不同物种之间的亲缘关系和演化途径，也可以用来表示同一物种内部某个基因家族的不同成员之间的关系及演化。

利用 MEGA 软件包[4]，可以构建人的珠蛋白基因家族 12 个成员系统发生树，所用序列为蛋白质序列，用全局比对程序 ClustalW 进行多序列比对，用 GONNET 蛋白质计分矩阵，空位罚分和其它参数均采用默认值。用邻接法（Neighbor-Joining）建树，采用差异位点比例（p-distance）为距离模型，选择自举法（Bootstrap）100 次作为稳定性检验。

利用 MEGA 软件包中的邻接法（Neighbor-Joining）方法构建人、小鼠、大鼠 3 个物种珠蛋白基因家族 37 个成员系统发生树（图 8），所用序列为编码区核苷酸质序列。序列比对采用 ClustalW Codon，即基于密码子的序列比对，比对过程中密码子 3 个核苷酸不打断，双序列和多序列比对的起始空位罚分均调为 20，延伸空位罚分均调为 2.0，以减少不必要的空位插入。建树过程中采用差异位点比例（p-distance）为序列差异模型，用转换加颠换（transition + transversion）为核苷酸替换模型，选择自举法（Bootstrap）100 次作为稳定性检验。

4.3 结果分析

图 7 所示的系统发生树为基因树。结果表明，人的 12 个珠蛋白基因可以分为 5 个分支，其中 α-珠蛋白亚家族包括 4 个成员，β-珠蛋白亚家族包括 5 个成员，而肌红蛋白、胞红蛋白和脑红蛋白各有 1 个成员。α-珠蛋白和 β-珠蛋白有共同祖先，而肌红蛋白和胞红蛋白有共同祖先。α-珠蛋白亚家族 4 个成员中，α-珠蛋白和 θ-珠蛋白之间的距离较近，而 β-珠蛋白亚家族 5 个亚家族中，γ_1-珠蛋白和 γ_2-珠蛋白

表 3　人、小鼠、大鼠 3 个物种珠蛋白家族基因信息

编号	基因名	RefSeq 登录号	基因长度 /bp	编码区	基因登录号	蛋白质登录号
Hs01	HsHBZ	NM_005332	589	56–484	3050	NP_005323
Hs02	HsHBM	NM_001003938	524	25–450	3042	NP_001003938
Hs03	HsHBA2	NM_000517	622	67–495	3040	NP_000508
Hs04	HsHBA1	NM_000558	627	67–495	3039	NP__000549
Hs05	HsHBQ1	NM_005331	653	154–582	3049	NP_005322
Hs06	HsHBE1	NM_005330	816	254–697	3046	NP_005321
Hs07	HsHBG2	NM_000184	583	54–497	3048	NP_000175
Hs08	HsHBG1	NM_000559	584	54–497	3047	NP_000550
Hs09	HsHBD	NM_000519	774	196–639	3045	NP_000510
Hs10	HsHBB	NM_000518	626	51–494	3043	NP_000509
Hs11	HsMYB	NM_005368	1078	81–454	4151	NP_005359
Hs12	HsCYGB	NM_134268	2166	364–936	114757	NP_599030
Hs13	HsNGB	NM_021257	1885	376–831	58157	NP_067080
Mm01	MmHba-x	NM_010405	630	65–493	15126	NP_034535
Mm02	MmHba-a1	NM_008218	569	36–464	15122	NP_032244
Mm03	MmHbq1a	NM_175000	604	50–487	216635	NP_778165
Mm04	MmHba-a2	NM_001083955	587	33–461	110257	NP_001077424
Mm05	MmHbq1b	NM_001033981	748	50–478	544763	NP_001029153
Mm06	MmHbb-y	NM_008221	619	56–499	15135	NP_032247
Mm07	MmHbb-bh1	NM_008219	610	53–496	15132	NP_032245
Mm08	MmHbb-bh2	NM_001127686	559	95–538	436003	NP_001121158
Mm09	MmHbb-b1	NM_001278161	630	55–498	15129	NP_001265090
Mm10	MmHbb-b2	NM_016956	630	55–498	15130	NP_058652
Mm11	MmMb	NM_001164047	1121	205–699	17189	NP_001157519
Mm12	MmCygb	NM_030206	2331	395–967	114886	NP_084482
Mm13	MmNgb	NM_022414	1630	280–735	64242	NP_071859
Rn01	RnHbz	NM_001172845	589	51–479	287168	NP_001166316
Rn02	RnHba1	NM_013096	556	34–462	25632	NP_037228
Rn03	RnHba2	NM_001007722	545	34–462	360504	NP_001007723
Rn04	RnHbe1	NM_001008890	444	1–444	293267	NP_001008890
Rn05	RnHbe2	NM_001024805	444	1–444	502359	NP_001019976
Rn06	RnHbg1	NM_172093	444	1–444	94164	NP_742090
Rn07	RnHbb	NM_033234	620	48–491	24440	NP_150237
Rn08	RnHbb-b1	NM_198776	659	57–550	361619	NP_942071
Rn09	RnMb	NM_021588	1015	28–492	59108	NP_067599
Rn10	RnCygb	NM_130744	2138	171–743	170520	NP_570100
Rn11	RnNGB	NM_033359	1773	410–865	85382	NP_203523

的距离最近，其次为 α-珠蛋白和 δ-珠蛋白。

图 8 所示的系统发生树包括 3 个物种，每个物种均有多个基因，共 37 个基因。结果表明，37 个基因总体可以分为 5 个分支，即 α-珠蛋白、β-珠蛋白、肌红蛋白、胞红蛋白和脑红蛋白。3 个物种的肌红蛋白、胞红蛋白和脑红蛋白各聚为一支；3 个物种所有 α-珠蛋白聚在一起，所有 β-珠蛋白聚在一起。这一结果说明，这 5 类基因在 3 个物种形成以前就已经出现，即“先有基因、后有物种”。α-珠蛋白分为 3 支，第一支为 ζ-珠蛋白，3 个物种各有一个成员，即人的 HsHBZ、小鼠的 MmHba-x 和大鼠的 RnHbz；第二支又分两支，一支为 α-珠蛋白，另一支为 θ-珠蛋白。3 个物种的 α-珠蛋白各有两个成员，如人的 HsHBA1 和 HsHBA2，θ-珠蛋白各有 1

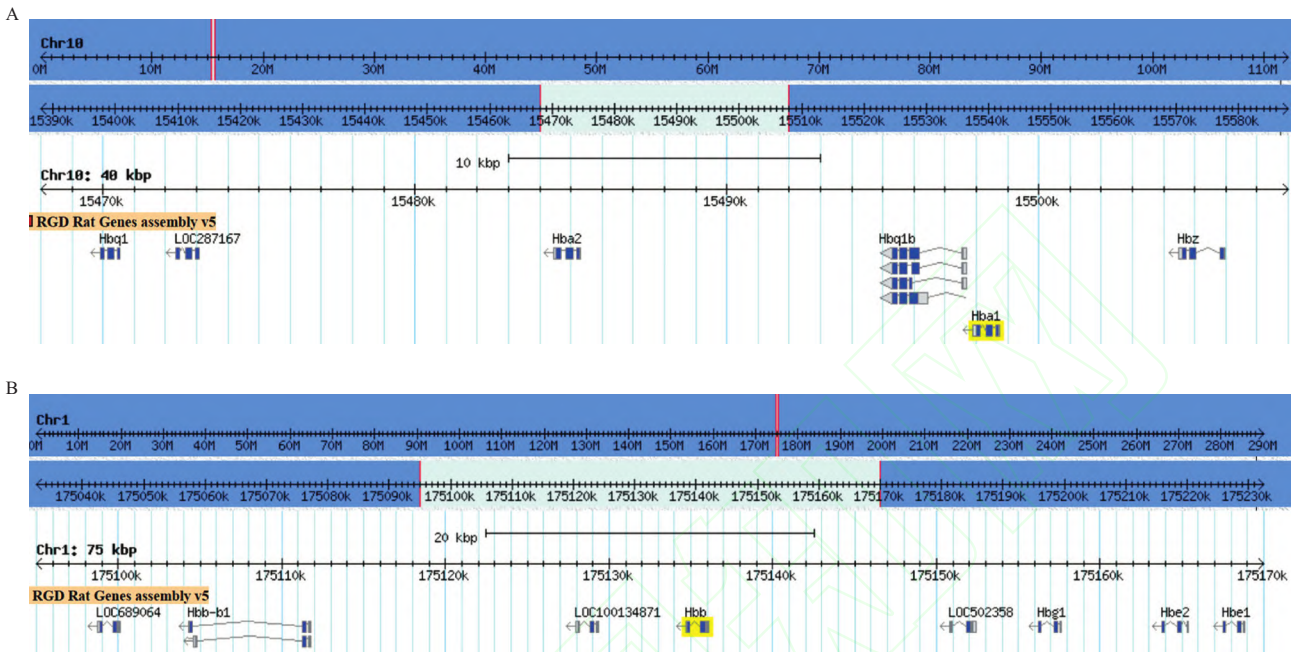

图 6 大鼠基因组数据库 RGD 中 α-珠蛋白（A）和 β-珠蛋白（B）基因家族信息

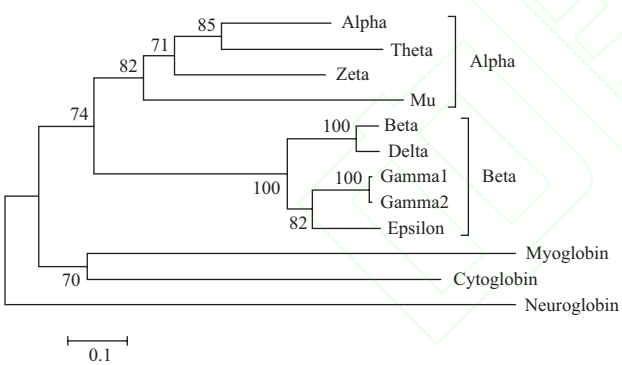

图 7 人珠蛋白家族 12 个蛋白质序列系统发生树

个成员。可以推断，α-珠蛋白的两个成员是在灵长类和啮齿类分化以后通过基因倍增机制产生的，即“先有物种、后有基因”。β-珠蛋白基因簇在这 3 个物种的起源和演化留给读者自行分析。

5 蛋白质结构比较和分析

5.1 研究背景

基于蛋白质和核酸序列，我们已对人、小鼠和大鼠 3 个物种的血红蛋白进行了比较分析。下面，我们以斑头雁和灰雁为例，利用生物信息方法和结构分析软件，对血红蛋白的序列、结构和功能关系进行分析。

斑头雁在分类学上为鸟纲（Aves）、雁形目（Anseriformes）、鸭科（Anatidae）、雁属（*Anser*），拉丁文学名分别为 *Anser indicus*，英文名为 Bar-headed goose。斑头雁为典型的候鸟，夏季生活在我国西部青海湖，每年 9 月初往南迁徙，经过近两个月的长途跋涉，飞跃喜马拉雅山，大约 10 月中下旬飞抵印度平原过冬。每年春季开始又往北迁徙，飞回青海湖，周而复始，年年如此。灰雁（英文名为 Grayleg goose，美国英语多用 Greyleg goose）的拉丁文学名分别为 *Anser anser*，与斑头雁同为鸭科、雁属，主要生活在印度平原[5]。我们知道，地球表面氧分压随海拔增高而降低，斑头雁飞跃的喜马拉雅山巅，氧分压不到平原地区的一半。斑头雁这种高空长度迁徙的能力，是否与其血红蛋白分子的特征有关，是一个值得研究的有趣问题。

1983 年，英国剑桥分子医学研究实验室已故著名血红蛋白研究专家佩鲁茨（Max Perutz）在分子生物学和演化杂志（Molecular Biology and Evolution）创刊号上发表的题为“从蛋白质分子看物种的适应性”综述中指出，斑头雁和灰雁的血红蛋白氨基酸序列仅有 4 个位点差异，其中 α-亚基的 119 位比较特殊[6]。斑头雁 α-亚基该位点位序氨酸（A119Ala），

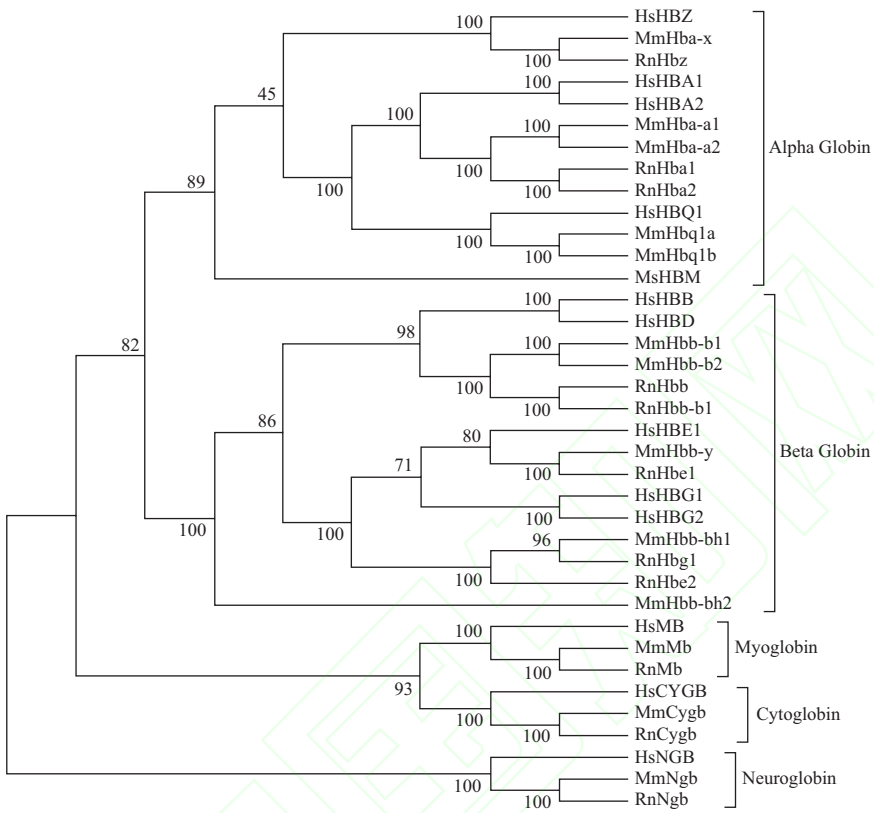

图 8　人、小鼠、大鼠 3 个物种珠蛋白家族系统发生树

而灰雁该位点为脯氨酸（A119Pro）。蛋白质三维空间结构分析表明，该位点与 β-亚基第 55 位的亮氨酸（B55Leu）空间距离较近。我们知道，成熟的血红蛋白为四聚体，由两个 α-亚基和两个 β-亚基组成，各含一个血色素卟啉环，环中央的五价铁离子用于结合氧气。结合氧气和释放氧气过程中，血红蛋白四个亚基构象发生变化，并通过协同作用，提高结合氧气的效率。佩鲁茨指出，斑头雁 α-亚基 119 位的丙氨酸侧链仅有一个甲基，与 β-亚基 55 位亮氨酸侧链距离较远，有利于构象变化；而灰雁该位点侧链脯氨酸有 3 个甲基，与 β-亚基 55 位亮氨酸侧链距离较近，不利于构象变化。这两种鸟类血红蛋白序列结构的差异，可能与其结合氧气的能力有关。20 世纪 90 年代，北京大学生物系蛋白质结构功能研究组，用蛋白质分子晶体 X-衍射的方法，分别测定了斑头雁和灰雁血红蛋白的结构，并进行了比较分析，证实了当年佩鲁茨的推测[7]。

5.2　*研究方法*

利用蛋白质结构显示和模拟软件 Swiss-PdbViewer[8]，我们可以对已经测定的斑头雁和灰雁氧合血红蛋白的空间结构进行比较分析。具体操作步骤大体如下：

从蛋白质结构数据库 PDB（http：//www.rcsb.org/）下载斑头雁和灰雁氧合血红蛋白三维空间结构数据文件 1A4F.pdb 和 1FAW.pdb。

（1）在 Swiss-PdbViewer 中打开灰雁血红蛋白数据文件 1FAW.pdb，选择其中 A 和 B 两条链（即 α 和 β-两个亚基），保存为新文件 1FAWab.pdb。

（2）打开新保存的文件 1FAWab.pdb，选择只显示主链模式；打开斑头雁血红蛋白数据文件 1A4F.pdb，也选择只显示主链模式。

（3）利用该软件包中的结构叠合工具 Magic Fit，可以发现，这两个蛋白质分子的结构总体十分相似。

（4）在控制面板中找到斑头雁 α-亚基 119 位的

丙氨酸和β-亚基55位亮氨酸，显示它们的侧链原子，测量它们之间的距离。

（5）在控制面板中找到灰雁α-亚基119位的脯氨酸和β-亚基55位亮氨酸，显示它们的侧链原子，测量它们之间的距离。

5.3 结果分析

上述斑头雁和灰雁血红蛋白三维结构的比较分析表明，斑头雁氧合血红蛋白1A4F α-亚基119位丙氨酸侧链的β碳原子（CB）与β-亚基55位亮氨酸侧链末端的两个δ碳原子（CD1和CD2）距离均在4 Å以上，最近距离为4.56 Å；而灰雁该位点侧链脯氨酸γ碳原子与β-亚基55位亮氨酸侧链末端的一个碳原子距离为3.79 Å。这一差别很可能影响血红蛋在结合和释放氧气过程中构象发生变化，从而影响其结合氧气能力，造成这两种鸟类不同的生活习性。

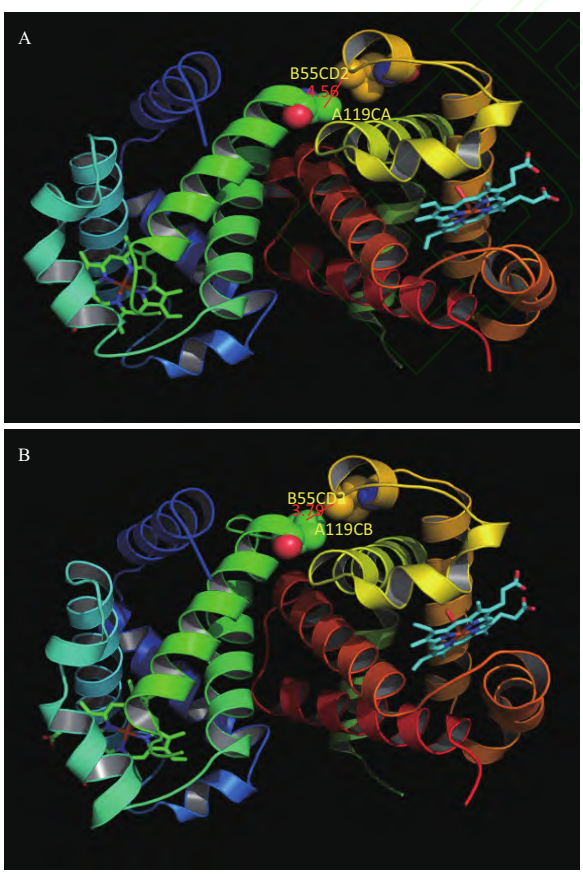

图9 斑头雁（A）和灰雁（B）血红蛋白结构比较

图9为利用PyMol分子结构显示软件绘制的分析结果。与Swiss-Pdbviewer相比，其图形显示和输出功能更强。

6 结语

以上我们以血红蛋白序列和结构为例，介绍“实用生物信息技术”课程教学种用到的几种生物信息方法。希望选修本课程的学生对本课程的教学有所了解，也希望对自学生物信息技术及其应用的读者有所启发。关于本课程的详细介绍和具体内容，读者可浏览本课程专用教学网站（http：//abc.cbi.pku.edu.cn/），参阅笔者生物信息学简报（Briefings in Bioinformatics）相关文章[1]。

参考文献

[1] Luo J. Teaching the ABCs of bioinformatics：a brief introduction to the Applied Bioinformatics Course [J]. Brief Bioinform, 2014, 15：1004-1013.

[2] Liu X, Wu J, Wang J, et al. WebLab：a data-centric, knowledge-sharing bioinformatic platform [J]. Nucleic Acids Res, 2009, 37：W33-39.

[3] Hardison RC. Evolution of hemoglobin and its genes [J]. Cold Spring Harb Perspect Med, 2012, 2：a011627.

[4] Tamura K, Stecher G, Peterson D, et al. MEGA6：Molecular Evolutionary Genetics Analysis version 6. 0 [J]. Mol Biol Evol, 2013, 30：2725-2729.

[5] Jessen TH, Weber RE, Fermi G, et al. Adaptation of bird hemoglobins to high altitudes：demonstration of molecular mechanism by protein engineering [J]. Proc Natl Acad Sci USA, 1991, 88：6519-6522.

[6] Perutz MF. Species adaptation in a protein molecule [J]. Mol Biol Evol, 1983, 1：1-28.

[7] Zhang J, Hua Z, Tame JR, et al. The crystal structure of a high oxygen affinity species of haemoglobin [J]. J Mol Biol, 1996, 255：484-493.

[8] Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer：a historical perspective [J]. Electrophoresis, 2009, 30：S162-173.

（责任编辑 马鑫）