

分子生物计算 (Perl 语言编程)

天津医科大学
生物医学工程与技术学院

2018-2019 学年上学期 (秋)
2016 级生信班

第一章 绪论

伊现富 (Yi Xianfu)

天津医科大学 (TJMU)
生物医学工程与技术学院

2018 年 11 月



教学提纲

- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题

教学提纲

1 课程安排

2 生物信息学

3 生物学

4 计算机科学

5 编程

- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题





后 9 周，每周一/三，
上午后两节（10:00-11:40）/下午后两节（15:30-17:10），
西楼 610

授课内容

- 教材内容：第一章……～第十章
- 补充知识：Markdown, Git, ...



- 上课期间以小组形式（建议以宿舍为单位）编写程序
- 任务主题不限（兴趣主导，生信方面的最好）
- 理论课的最后一/两次课（2~4 学时）进行报告
- 小组成员人人都要参与
- 既参与编程又进行报告者加分
- 作为平时成绩的一大部分



目的

按照用户的需求生成随机密码。

需求

- 指定生成密码的元素（默认：混合使用数字、大小写字母、符号）
- 指定生成密码的长度（默认：12 个字符）
- 指定生成密码的个数（默认：1 个）
- 指定密码中元素的比例/字符数（默认：无比例/字符数要求）
- 指定需要排除的元素（比如：0 和 O, 1 和 I；默认：无）
- 挖掘更多人性化需求……

代码仓库

- [Yixf-Education/project_Perl](#)

目的

按照用户的需求生成随机密码。

需求

- 指定生成密码的元素（默认：混合使用数字、大小写字母、符号）
- 指定生成密码的长度（默认：12个字符）
- 指定生成密码的个数（默认：1个）
- 指定密码中元素的比例/字符数（默认：无比例/字符数要求）
- 指定需要排除的元素（比如：0和O, 1和I；默认：无）
- 挖掘更多人性化需求……

代码仓库

- [Yixf-Education/project_Perl](#)

目的

按照用户的需求生成随机密码。

需求

- 指定生成密码的元素（默认：混合使用数字、大小写字母、符号）
- 指定生成密码的长度（默认：12个字符）
- 指定生成密码的个数（默认：1个）
- 指定密码中元素的比例/字符数（默认：无比例/字符数要求）
- 指定需要排除的元素（比如：0和O, 1和I；默认：无）
- 挖掘更多人性化需求……

代码仓库

- [Yixf-Education/project_Perl](#)

后 9 周，每周二，上午前两节（8:00-9:40），教一楼 304

实验内容

- 紧跟理论课进度
- 幻灯片与教材上的程序/实例



理论课：80%

- ① 平时表现：5%
- ② 课堂测验：5%（期中）+5%（期末）
- ③ 自主学习：15%
- ④ 闭卷考试：50%

实验课：20%

- ① 平时表现：10%
- ② 实验报告：10%



教学提纲

1 课程安排

2 生物信息学

3 生物学

4 计算机科学

5 编程

6 Bioinformatics Career Survey
2008

7 R

8 书籍

9 回顾与总结

- 总结

- 思考题



生物信息学 (bioinformatics)

"If you can't do bioinformatics, you can't do biology."

Lincoln Stein (Biologist and former CSHL Professor)

```
CCGCT);GTATTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACC
A#print STDERR 'blast args: ', Dumper( \@args ), $/;A
CACCCACTAGGATACCAACAAACCTACCCACCCCTAACAGTACATACTACATC
C#y $pcf = DNALC::Pipeline::Config->new->cf('PIPELINE
Cmy $blast_script = File::Spec->catfile($pcf->{EXE_PA
Amy $rc = system($blast_script, @args);CAATCAACCTATA
Cprint STDERR "blast rc = $rc\n";TTAACAGTACATACTACATC
CTGTTCTTCATGGGAAGCAGATTGGGTACCACCCAAAGTATTGACCACCCA
C# 0 == successTACATTACTGCCAGCCACCATGAATATTGTACGGTACC
A# 2 == success, no resultsAAGCAAGTACAGCAATCAACCTATA
Cif ((0 == $rc || 2 == $rc) && -f $out_file) {GTACATC
CTGTTmy $alignment = '';TTGGGTACCACCCAAAGTATTGACCACCCA
CCGCTif ($fh->open($out_file)) {CCATGAATATTGTACGGTACC
ATATCAAAAwhile (<$fh>) {TACAAGCAAGTACAGCAATCAACCTATA
CACCCACTAGGAT$alignment .= $_;CCCTTAACAGTACATACTACATC
CTGTTCTT)ATGGGAAGCAGATTGGGTACCACCCAAAGTATTGACCACCCA
CCGCTATGT$fh->close;TACTGCCAGCCACCATGAATATTGTACGGTACC
ATATC)AAACCCCTCCCCATGCTTACAAGCAAGTACAGCAATCAACCTATA
CACCC$blast = DNALC::Pipeline::Phylogenetics::Blast->
CACCCACTAGGATproject_id => $self->project->id,GTACATC
```

Nothing in Biology Makes Sense Except in the Light of Evolution
and Bioinformatics.

生物信息学（bioinformatics）利用应用数学、信息学、统计学和计算机科学的方法研究生物学的问题。

生物信息学的**研究材料和结果**就是各种各样的生物学数据，其**研究工具**是计算机，**研究方法**包括对生物学数据的搜索（收集和筛选）、处理（编辑、整理、管理和显示）及利用（计算、模拟）。

目前主要的研究方向有：序列比对、序列组装、基因识别、基因重组、蛋白质结构预测、基因表达、蛋白质反应的预测，以及建立进化模型。



生物学技术往往生成大量的嘈杂数据。与数据挖掘类似，生物信息学利用数学工具从大量数据中提取有用的生物学信息。生物信息学所要处理的典型问题包括：重新组装在霰弹枪测序法测序过程中被打散的 DNA 序列，从蛋白质的氨基酸序列预测蛋白质结构，利用 mRNA 微阵列或质谱仪的数据检验基因调控的假说。

某些人将计算生物学作为生物信息学的同义词处理；但是另外一些人认为计算生物学和生物信息学应当被当作不同的条目处理，因为生物信息学更侧重于生物学领域中计算方法的使用和发展，而计算生物学强调应用信息学技术对生物学领域中的假说进行检验，并尝试发展新的理论。



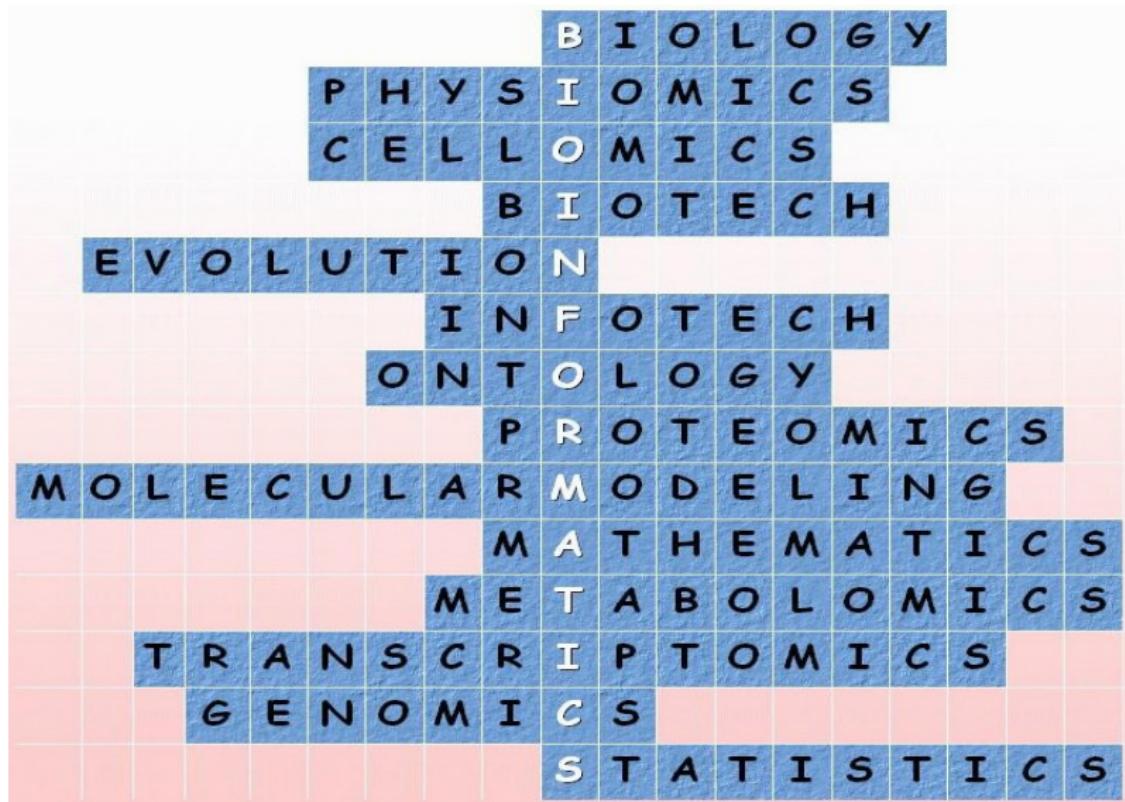
生物信息学可以定义为对分子生物学中两类信息流的研究：

第一类信息流源于分子生物学的中心法则：DNA 序列被转录为 mRNA 序列，后者被翻译为蛋白质序列。蛋白质序列继而折叠为具有功能的三维结构。按照达尔文演化理论，这些功能被生物体的环境所选择，从而驱动群体中 DNA 序列的进化。因此，第一类的生物信息学应用关注于中心法则中任一阶段的信息传递，包括 DNA 序列中基因的组织与控制、确定 DNA 中的转录单位、从序列预测蛋白质结构以及分子功能分析。

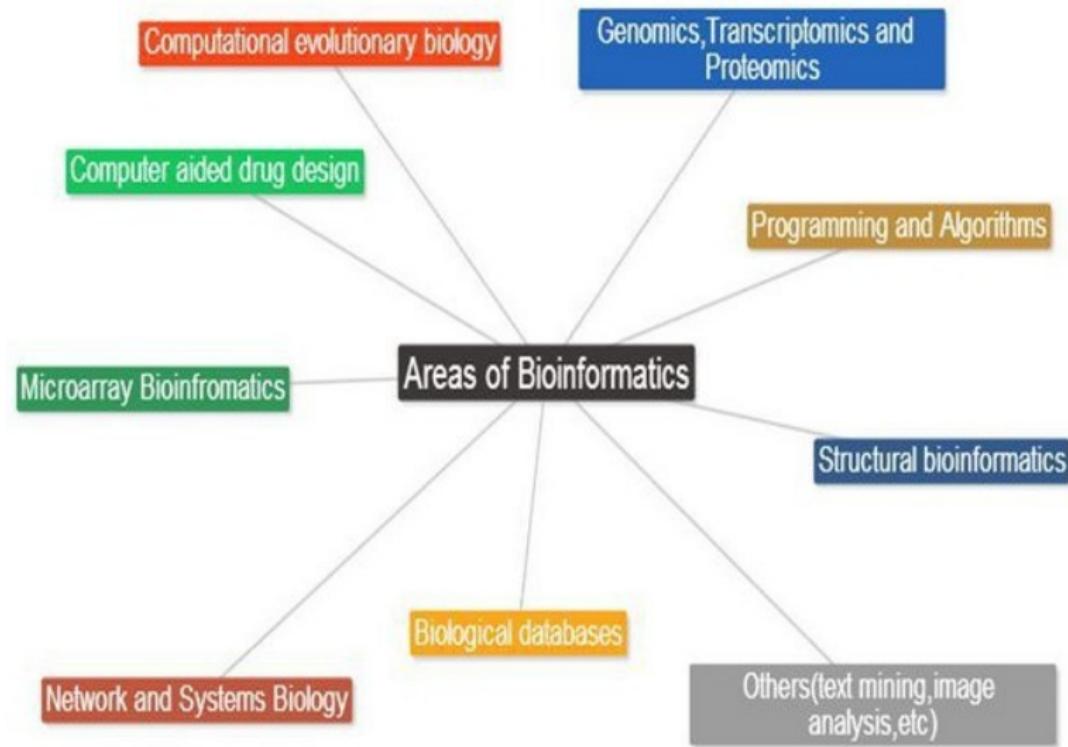
第二类信息流是基于科学方法：提出关于生物学活动的假设，设计实验以验证这些假设，评估结果与假设的相容性，然后根据实验数据对原假设作扩展或修正。第二类的生物信息学应用关注于这一流程中的信息传递，包括产生假设、设计实验、通过数据库将实验结果组织起来、检验数据与模型的相容性以及修正假设的各个系统。

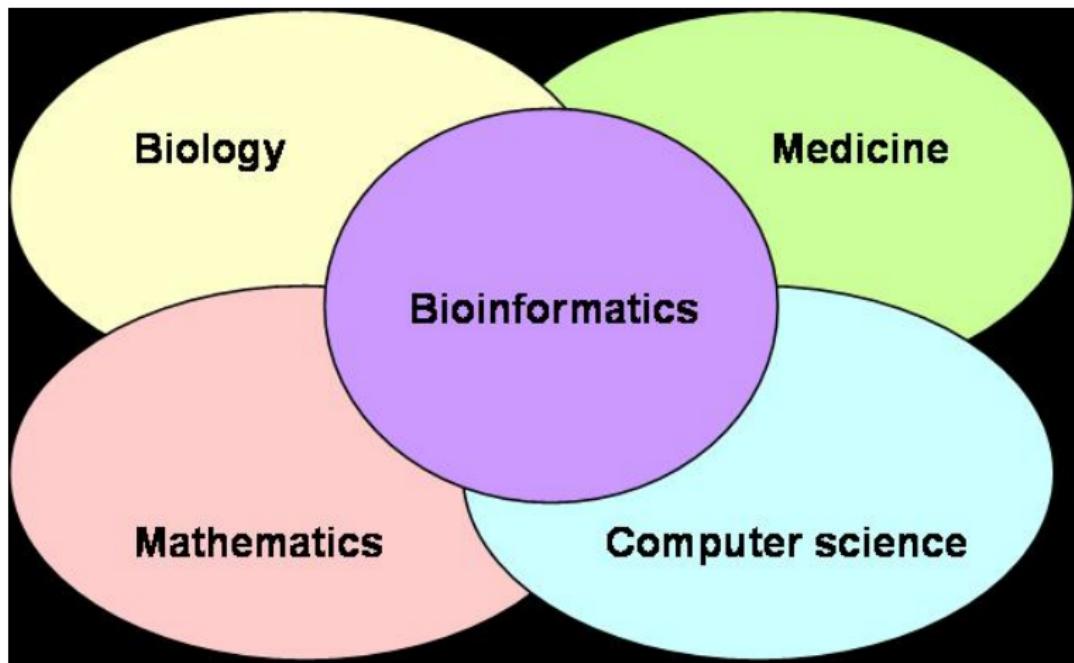


生物信息学 | 是什么？| 学科角度

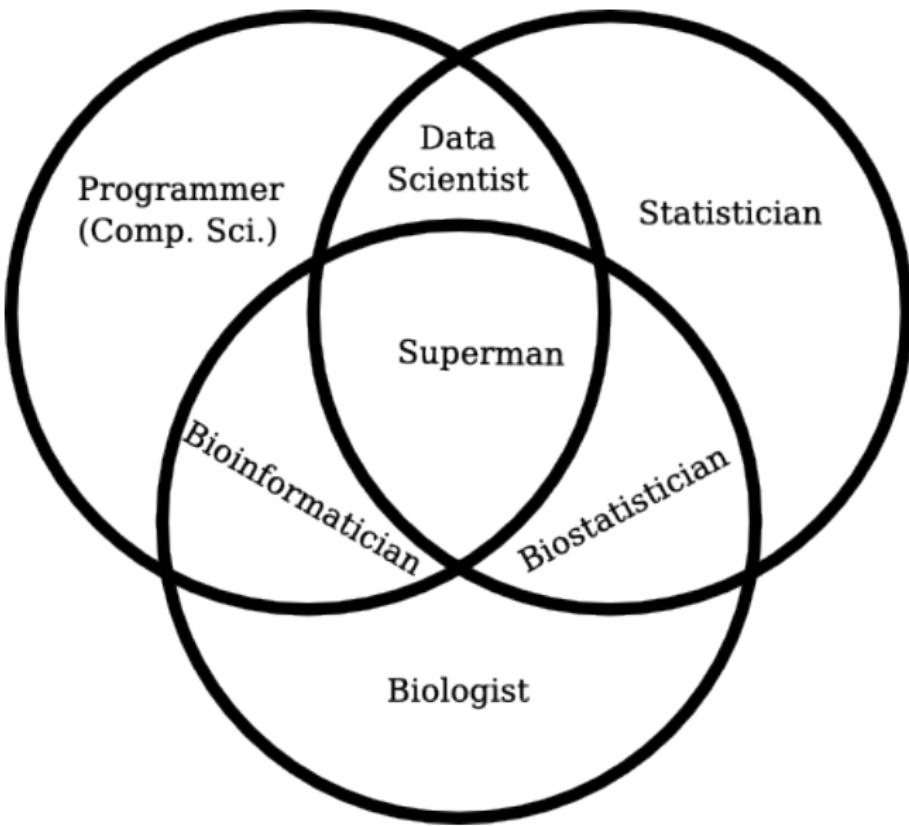


生物信息学 | 是什么？| 学科角度

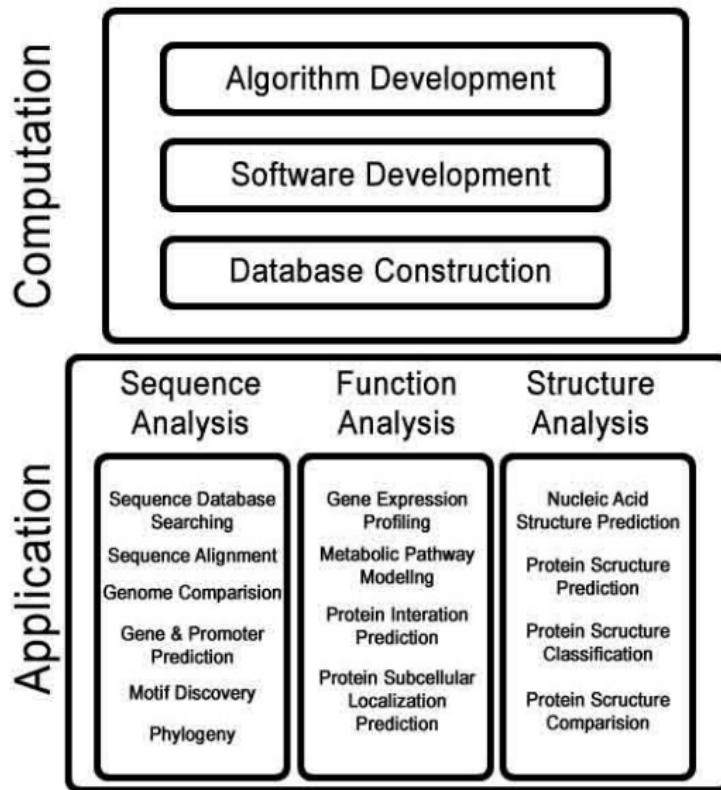




生物信息学 | 是什么？| 技术角度



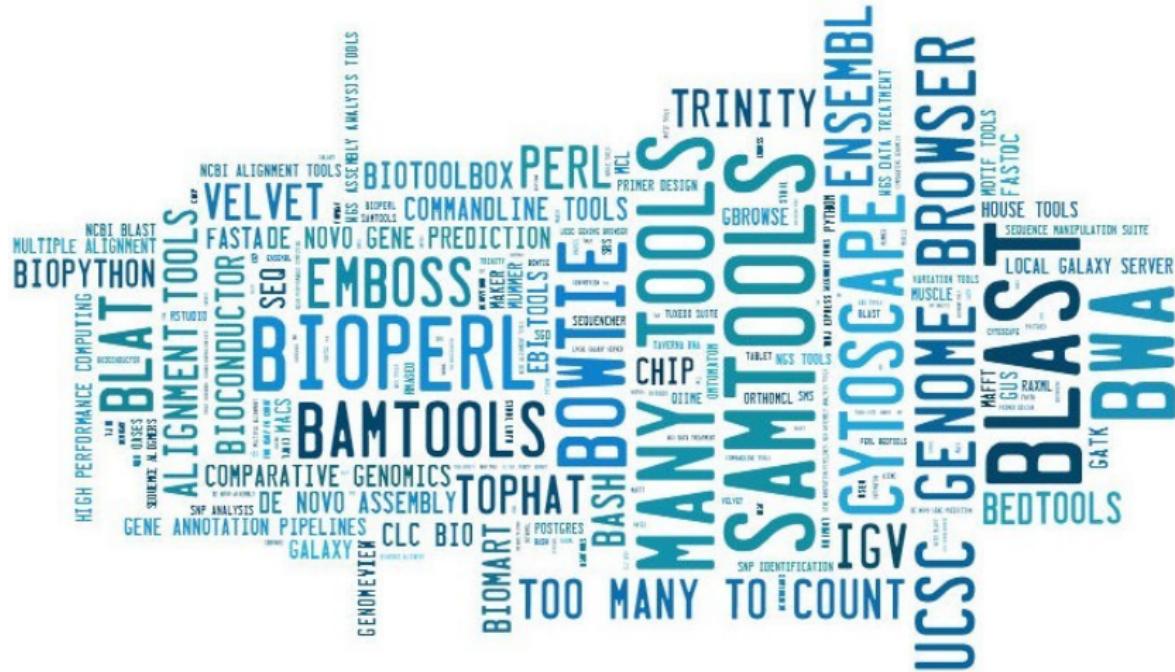
生物信息学 | 做什么？



生物信息学 | 资源 (数据库与工具)



生物信息学 | 资源 (技术与工具)



Cancer informatics Gene regulation
Personalized medicine Protein modeling
Computational biology Gene expression analysis
Image analysis Genomics and proteomics
Comparative genomics Gene expression databases
Epidemic models Computational drug discovery

Bioinformatics

Sequence analysis Bio-ontologies and semantics
Evolution and phylogenetics Structure prediction
Cheminformatics Next generation sequencing
Computational intelligence Transcriptomics
Biomedical engineering Amino acid sequencing
Structural bioinformatics Medical informatics
Microarrays
Visualization





生物信息学家 = 生命科学中的黑客！



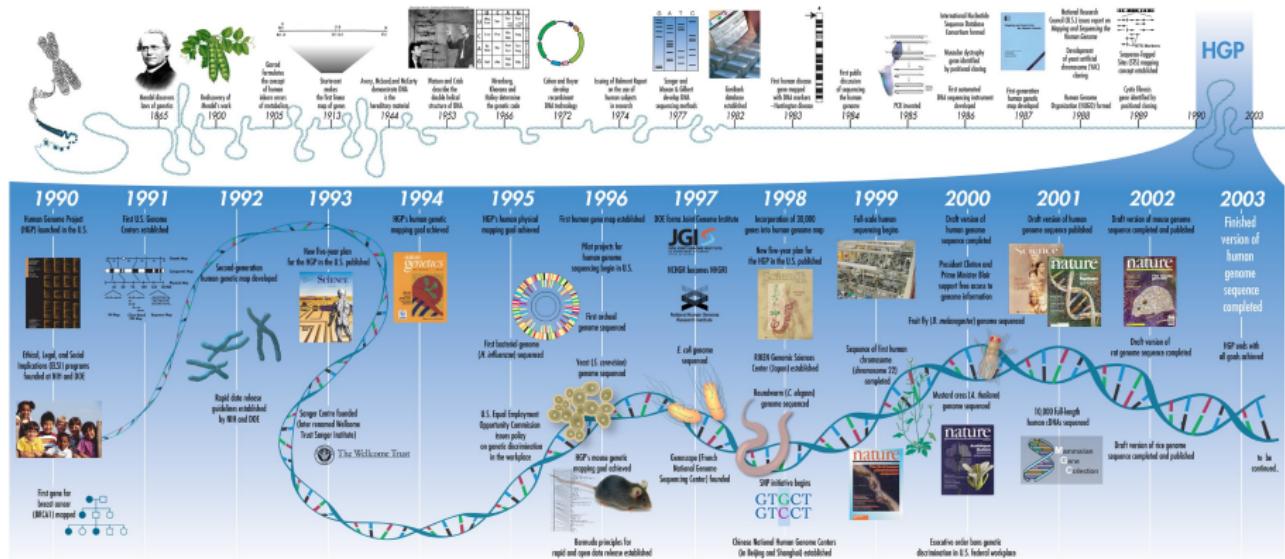
教学提纲

- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

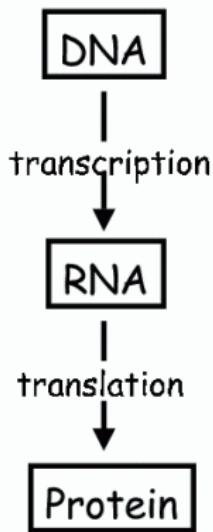
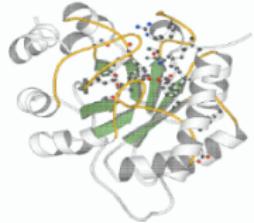
- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题



生物学 | 历史



Central Dogma: DNA → RNA → Protein

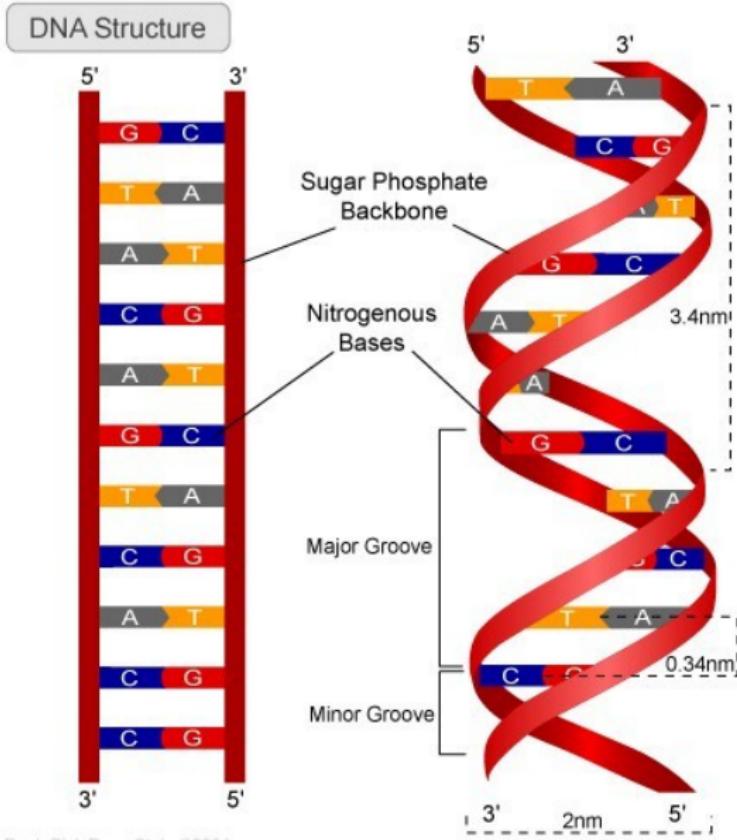


CCTGAGCCAACTATTGATGAA

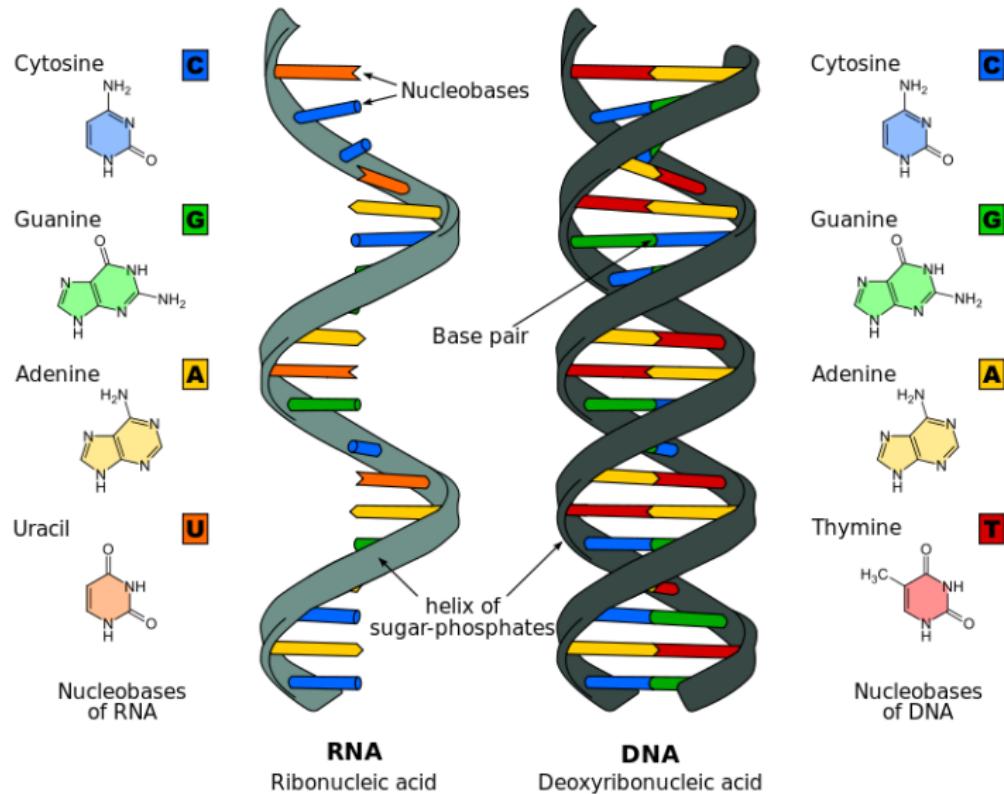
CCUGAGCCAACUAUUGAUUGAA

PEPTIDE

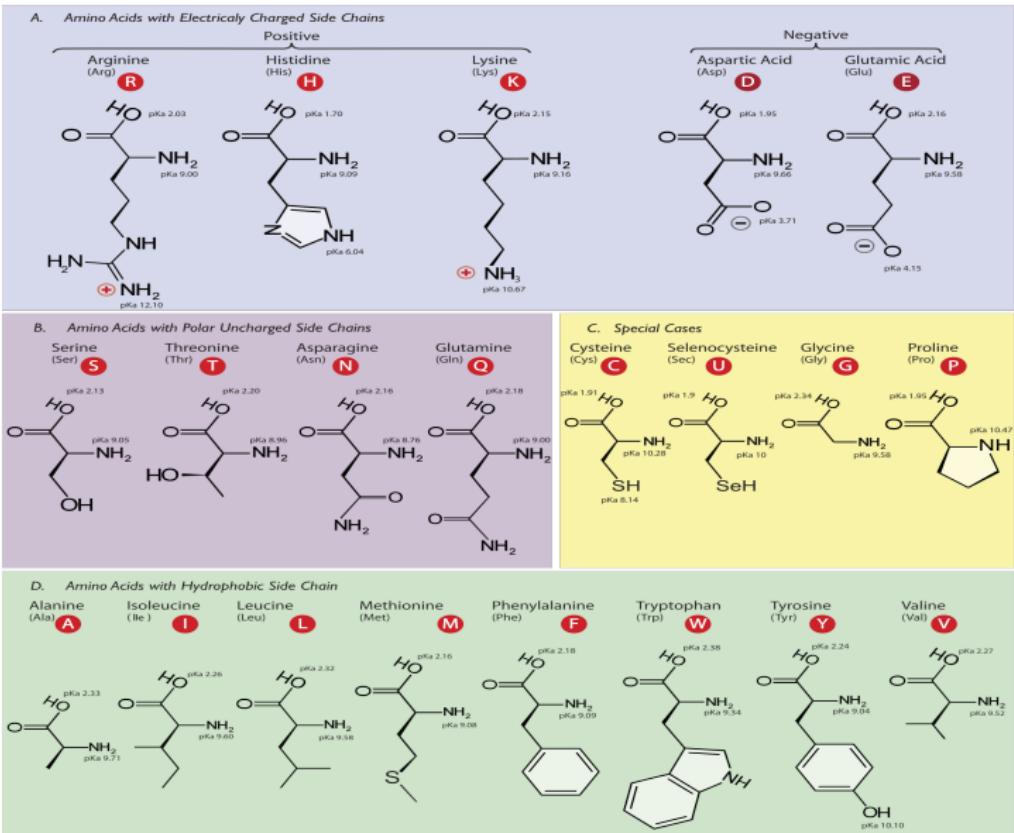
生物学 | DNA



生物学 | RNA



生物学 | 蛋白质 | 氨基酸



必需氨基酸

甲携来一本亮色书（甲硫氨酸、缬氨酸、赖氨酸、异亮氨酸、苯丙氨酸、亮氨酸、色氨酸、苏氨酸）

20 种氨基酸

苏缬亮异亮，苯丙属芳香。
还有色赖蛋，缺一人遭殃。

(以上是必需氨基酸)

丙组丝甘半，天谷建酸胺。
精酪加一脯，二十氨基酸。

(以上是非必需氨基酸)



必需氨基酸

甲携来一本亮色书（甲硫氨酸、缬氨酸、赖氨酸、异亮氨酸、苯丙氨酸、亮氨酸、色氨酸、苏氨酸）

20 种氨基酸

苏缬亮异亮，苯丙属芳香。

还有色赖蛋，缺一人遭殃。

(以上是必需氨基酸)

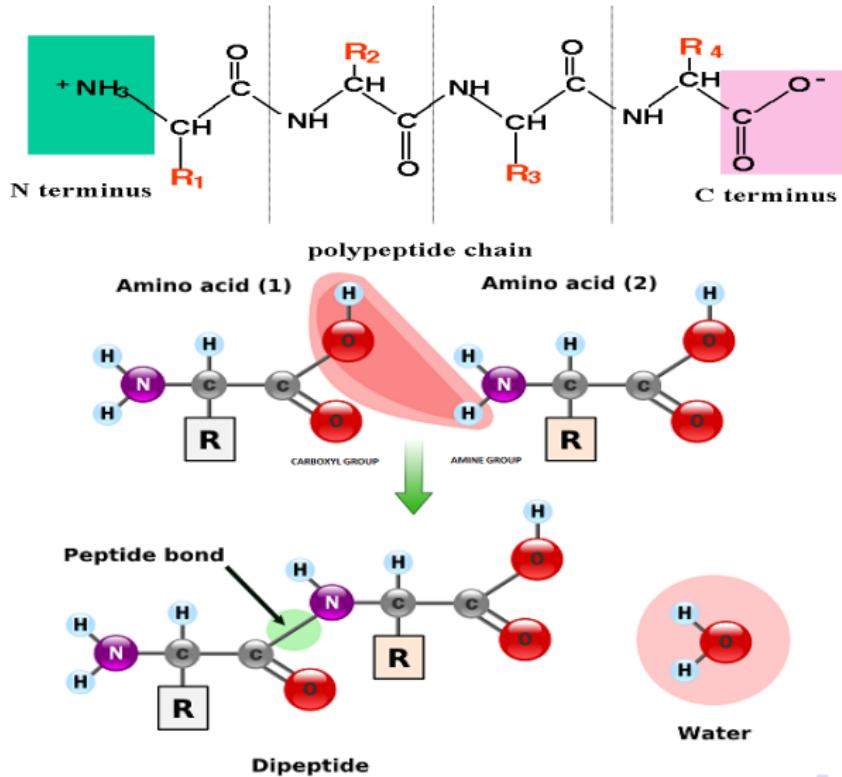
丙组丝甘半，天谷建酸胺。

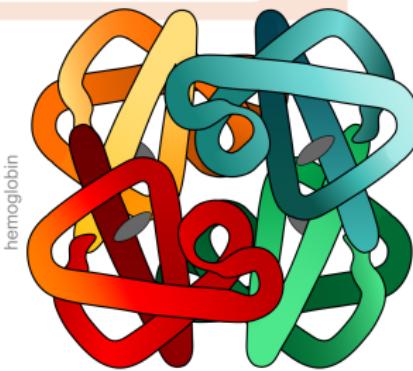
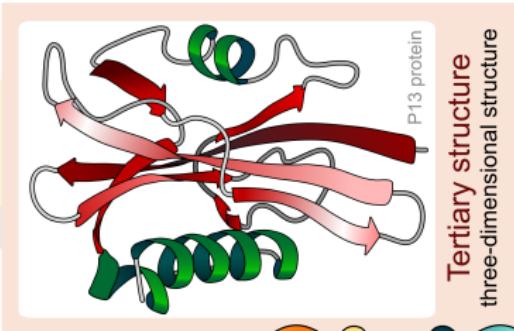
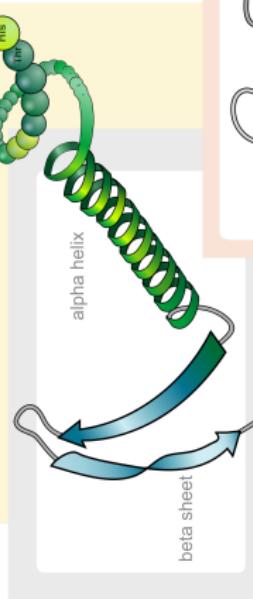
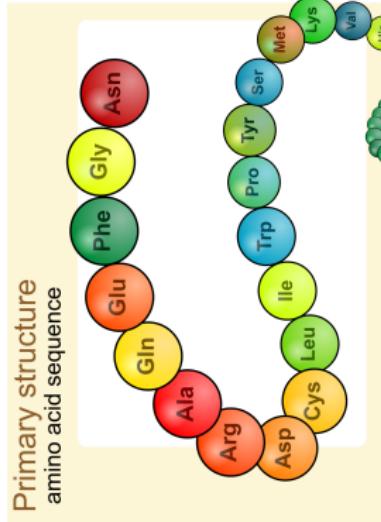
精酪加一脯，二十氨基酸。

(以上是非必需氨基酸)



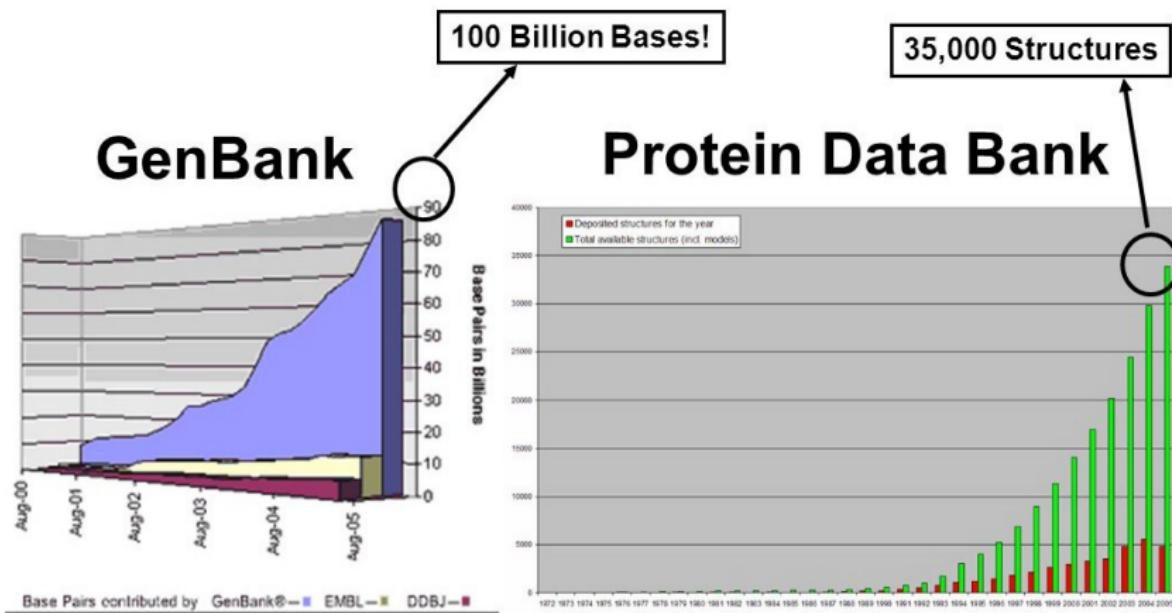
Peptide = chain of amino acids





Quaternary structure
complex of protein molecules



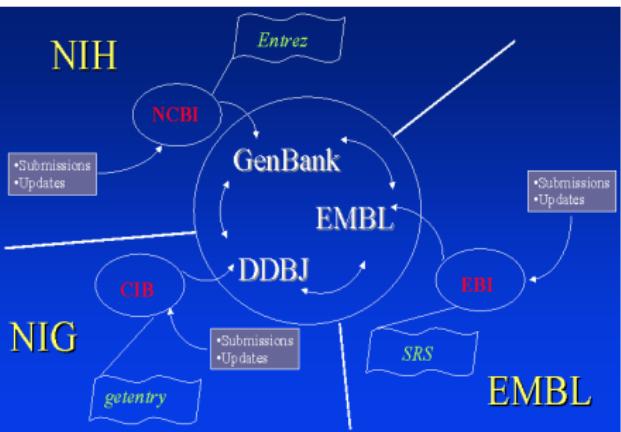
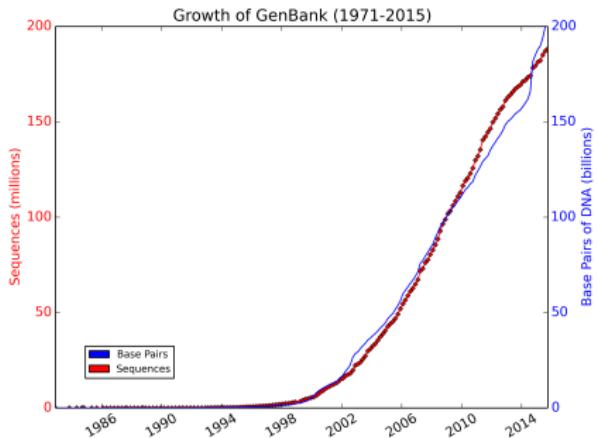


www.ncbi.nlm.nih.gov/Genbank

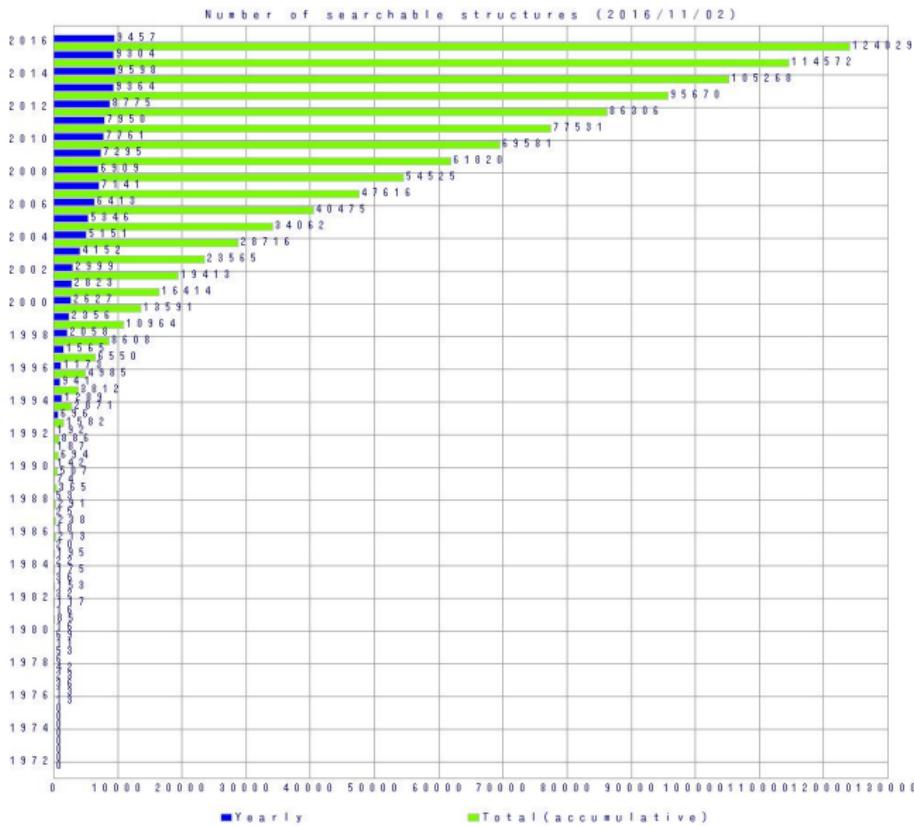
www.rcsb.org/pdb/holdings.html



生物学 | 数据库 | GenBank



生物学 | 数据库 | PDB



生物学 | 数据格式

Historically, bioinformatics has always used text files to store data.

Homo sapiens CD99 molecule, mRNA (cDNA clone MGC:19734 IMAGE:3606974), complete cds

Genbank: BC0101092

PARTA: CDS

DEFINITION Homo sapiens CD99 molecule, mRNA [cDNA clone MGC:19734 IMAGE:3606974], complete cds.

ACCESSION BC0101092

VERSION BC0101092.2 GI:33991438

KEYWORDS MGC;

SOURCE Homo sapiens [human]

ORGANISM Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo;

REFERENCE 1 (bases to 12551)

AUTHORS Stramberg Klaesner A, Altheil J, Hopkins R, Dastcherikoff S, Stapleton S, Scheetz T, Carninci P, Abramson R, McKernan K

COMMENT [S1-A] [March 2002] Predicted gene model.

LINE 285

ALPM [protein]

CDS [CDS]

PROTEIN [protein]

EXON [exon]

INTRON [introns]

TRANSLATE [translate]

TRANSMIT [transmit]

EXON 1 1-2551

TRANSLATE 1-2551

TRANSMIT 1-2551

EXON 2 256-275 276-285 286-295 296-305 306-315 316-325 326-335 336-345 346-355 356-365 366-375 376-385 386-395 396-405 406-415 416-425 426-435 436-445 446-455 456-465 466-475 476-485 486-495 496-505 506-515 516-525 526-535 536-545 546-555 556-565 566-575 576-585 586-595 596-605 606-615 616-625 626-635 636-645 646-655 656-665 666-675 676-685 686-695 696-705 706-715 716-725 726-735 736-745 746-755 756-765 766-775 776-785 786-795 796-805 806-815 816-825 826-835 836-845 846-855 856-865 866-875 876-885 886-895 896-905 906-915 916-925 926-935 936-945 946-955 956-965 966-975 976-985 986-995 996-1005 1006-1015 1016-1025 1026-1035 1036-1045 1046-1055 1056-1065 1066-1075 1076-1085 1086-1095 1096-1105 1106-1115 1116-1125 1126-1135 1136-1145 1146-1155 1156-1165 1166-1175 1176-1185 1186-1195 1196-1205 1206-1215 1216-1225 1226-1235 1236-1245 1246-1255 1256-1265 1266-1275 1276-1285 1286-1295 1296-1305 1306-1315 1316-1325 1326-1335 1336-1345 1346-1355 1356-1365 1366-1375 1376-1385 1386-1395 1396-1405 1406-1415 1416-1425 1426-1435 1436-1445 1446-1455 1456-1465 1466-1475 1476-1485 1486-1495 1496-1505 1506-1515 1516-1525 1526-1535 1536-1545 1546-1555 1556-1565 1566-1575 1576-1585 1586-1595 1596-1605 1606-1615 1616-1625 1626-1635 1636-1645 1646-1655 1656-1665 1666-1675 1676-1685 1686-1695 1696-1705 1706-1715 1716-1725 1726-1735 1736-1745 1746-1755 1756-1765 1766-1775 1776-1785 1786-1795 1796-1805 1806-1815 1816-1825 1826-1835 1836-1845 1846-1855 1856-1865 1866-1875 1876-1885 1886-1895 1896-1905 1906-1915 1916-1925 1926-1935 1936-1945 1946-1955 1956-1965 1966-1975 1976-1985 1986-1995 1996-2005 2006-2015 2016-2025 2026-2035 2036-2045 2046-2055 2056-2065 2066-2075 2076-2085 2086-2095 2096-2105 2106-2115 2116-2125 2126-2135 2136-2145 2146-2155 2156-2165 2166-2175 2176-2185 2186-2195 2196-2205 2206-2215 2216-2225 2226-2235 2236-2245 2246-2255 2256-2265 2266-2275 2276-2285 2286-2295 2296-2305 2306-2315 2316-2325 2326-2335 2336-2345 2346-2355 2356-2365 2366-2375 2376-2385 2386-2395 2396-2405 2406-2415 2416-2425 2426-2435 2436-2445 2446-2455 2456-2465 2466-2475 2476-2485 2486-2495 2496-2505 2506-2515 2516-2525 2526-2535 2536-2545 2546-2555 2556-2565 2566-2575 2576-2585 2586-2595 2596-2605 2606-2615 2616-2625 2626-2635 2636-2645 2646-2655 2656-2665 2666-2675 2676-2685 2686-2695 2696-2705 2706-2715 2716-2725 2726-2735 2736-2745 2746-2755 2756-2765 2766-2775 2776-2785 2786-2795 2796-2805 2806-2815 2816-2825 2826-2835 2836-2845 2846-2855 2856-2865 2866-2875 2876-2885 2886-2895 2896-2905 2906-2915 2916-2925 2926-2935 2936-2945 2946-2955 2956-2965 2966-2975 2976-2985 2986-2995 2996-3005 3006-3015 3016-3025 3026-3035 3036-3045 3046-3055 3056-3065 3066-3075 3076-3085 3086-3095 3096-3105 3106-3115 3116-3125 3126-3135 3136-3145 3146-3155 3156-3165 3166-3175 3176-3185 3186-3195 3196-3205 3206-3215 3216-3225 3226-3235 3236-3245 3246-3255 3256-3265 3266-3275 3276-3285 3286-3295 3296-3305 3306-3315 3316-3325 3326-3335 3336-3345 3346-3355 3356-3365 3366-3375 3376-3385 3386-3395 3396-3405 3406-3415 3416-3425 3426-3435 3436-3445 3446-3455 3456-3465 3466-3475 3476-3485 3486-3495 3496-3505 3506-3515 3516-3525 3526-3535 3536-3545 3546-3555 3556-3565 3566-3575 3576-3585 3586-3595 3596-3605 3606-3615 3616-3625 3626-3635 3636-3645 3646-3655 3656-3665 3666-3675 3676-3685 3686-3695 3696-3705 3706-3715 3716-3725 3726-3735 3736-3745 3746-3755 3756-3765 3766-3775 3776-3785 3786-3795 3796-3805 3806-3815 3816-3825 3826-3835 3836-3845 3846-3855 3856-3865 3866-3875 3876-3885 3886-3895 3896-3905 3906-3915 3916-3925 3926-3935 3936-3945 3946-3955 3956-3965 3966-3975 3976-3985 3986-3995 3996-4005 4006-4015 4016-4025 4026-4035 4036-4045 4046-4055 4056-4065 4066-4075 4076-4085 4086-4095 4096-4105 4106-4115 4116-4125 4126-4135 4136-4145 4146-4155 4156-4165 4166-4175 4176-4185 4186-4195 4196-4205 4206-4215 4216-4225 4226-4235 4236-4245 4246-4255 4256-4265 4266-4275 4276-4285 4286-4295 4296-4305 4306-4315 4316-4325 4326-4335 4336-4345 4346-4355 4356-4365 4366-4375 4376-4385 4386-4395 4396-4405 4406-4415 4416-4425 4426-4435 4436-4445 4446-4455 4456-4465 4466-4475 4476-4485 4486-4495 4496-4505 4506-4515 4516-4525 4526-4535 4536-4545 4546-4555 4556-4565 4566-4575 4576-4585 4586-4595 4596-4605 4606-4615 4616-4625 4626-4635 4636-4645 4646-4655 4656-4665 4666-4675 4676-4685 4686-4695 4696-4705 4706-4715 4716-4725 4726-4735 4736-4745 4746-4755 4756-4765 4766-4775 4776-4785 4786-4795 4796-4805 4806-4815 4816-4825 4826-4835 4836-4845 4846-4855 4856-4865 4866-4875 4876-4885 4886-4895 4896-4905 4906-4915 4916-4925 4926-4935 4936-4945 4946-4955 4956-4965 4966-4975 4976-4985 4986-4995 4996-5005 5006-5015 5016-5025 5026-5035 5036-5045 5046-5055 5056-5065 5066-5075 5076-5085 5086-5095 5096-5105 5106-5115 5116-5125 5126-5135 5136-5145 5146-5155 5156-5165 5166-5175 5176-5185 5186-5195 5196-5205 5206-5215 5216-5225 5226-5235 5236-5245 5246-5255 5256-5265 5266-5275 5276-5285 5286-5295 5296-5305 5306-5315 5316-5325 5326-5335 5336-5345 5346-5355 5356-5365 5366-5375 5376-5385 5386-5395 5396-5405 5406-5415 5416-5425 5426-5435 5436-5445 5446-5455 5456-5465 5466-5475 5476-5485 5486-5495 5496-5505 5506-5515 5516-5525 5526-5535 5536-5545 5546-5555 5556-5565 5566-5575 5576-5585 5586-5595 5596-5605 5606-5615 5616-5625 5626-5635 5636-5645 5646-5655 5656-5665 5666-5675 5676-5685 5686-5695 5696-5705 5706-5715 5716-5725 5726-5735 5736-5745 5746-5755 5756-5765 5766-5775 5776-5785 5786-5795 5796-5805 5806-5815 5816-5825 5826-5835 5836-5845 5846-5855 5856-5865 5866-5875 5876-5885 5886-5895 5896-5905 5906-5915 5916-5925 5926-5935 5936-5945 5946-5955 5956-5965 5966-5975 5976-5985 5986-5995 5996-6005 6006-6015 6016-6025 6026-6035 6036-6045 6046-6055 6056-6065 6066-6075 6076-6085 6086-6095 6096-6105 6106-6115 6116-6125 6126-6135 6136-6145 6146-6155 6156-6165 6166-6175 6176-6185 6186-6195 6196-6205 6206-6215 6216-6225 6226-6235 6236-6245 6246-6255 6256-6265 6266-6275 6276-6285 6286-6295 6296-6305 6306-6315 6316-6325 6326-6335 6336-6345 6346-6355 6356-6365 6366-6375 6376-6385 6386-6395 6396-6405 6406-6415 6416-6425 6426-6435 6436-6445 6446-6455 6456-6465 6466-6475 6476-6485 6486-6495 6496-6505 6506-6515 6516-6525 6526-6535 6536-6545 6546-6555 6556-6565 6566-6575 6576-6585 6586-6595 6596-6605 6606-6615 6616-6625 6626-6635 6636-6645 6646-6655 6656-6665 6666-6675 6676-6685 6686-6695 6696-6705 6706-6715 6716-6725 6726-6735 6736-6745 6746-6755 6756-6765 6766-6775 6776-6785 6786-6795 6796-6805 6806-6815 6816-6825 6826-6835 6836-6845 6846-6855 6856-6865 6866-6875 6876-6885 6886-6895 6896-6905 6906-6915 6916-6925 6926-6935 6936-6945 6946-6955 6956-6965 6966-6975 6976-6985 6986-6995 6996-7005 7006-7015 7016-7025 7026-7035 7036-7045 7046-7055 7056-7065 7066-7075 7076-7085 7086-7095 7096-7105 7106-7115 7116-7125 7126-7135 7136-7145 7146-7155 7156-7165 7166-7175 7176-7185 7186-7195 7196-7205 7206-7215 7216-7225 7226-7235 7236-7245 7246-7255 7256-7265 7266-7275 7276-7285 7286-7295 7296-7305 7306-7315 7316-7325 7326-7335 7336-7345 7346-7355 7356-7365 7366-7375 7376-7385 7386-7395 7396-7405 7406-7415 7416-7425 7426-7435 7436-7445 7446-7455 7456-7465 7466-7475 7476-7485 7486-7495 7496-7505 7506-7515 7516-7525 7526-7535 7536-7545 7546-7555 7556-7565 7566-7575 7576-7585 7586-7595 7596-7605 7606-7615 7616-7625 7626-7635 7636-7645 7646-7655 7656-7665 7666-7675 7676-7685 7686-7695 7696-7705 7706-7715 7716-7725 7726-7735 7736-7745 7746-7755 7756-7765 7766-7775 7776-7785 7786-7795 7796-7805 7806-7815 7816-7825 7826-7835 7836-7845 7846-7855 7856-7865 7866-7875 7876-7885 7886-7895 7896-7905 7906-7915 7916-7925 7926-7935 7936-7945 7946-7955 7956-7965 7966-7975 7976-7985 7986-7995 7996-8005 8006-8015 8016-8025 8026-8035 8036-8045 8046-8055 8056-8065 8066-8075 8076-8085 8086-8095 8096-8105 8106-8115 8116-8125 8126-8135 8136-8145 8146-8155 8156-8165 8166-8175 8176-8185 8186-8195 8196-8205 8206-8215 8216-8225 8226-8235 8236-8245 8246-8255 8256-8265 8266-8275 8276-8285 8286-8295 8296-8305 8306-8315 8316-8325 8326-8335 8336-8345 8346-8355 8356-8365 8366-8375 8376-8385 8386-8395 8396-8405 8406-8415 8416-8425 8426-8435 8436-8445 8446-8455 8456-8465 8466-8475 8476-8485 8486-8495 8496-8505 8506-8515 8516-8525 8526-8535 8536-8545 8546-8555 8556-8565 8566-8575 8576-8585 8586-8595 8596-8605 8606-8615 8616-8625 8626-8635 8636-8645 8646-8655 8656-8665 8666-8675 8676-8685 8686-8695 8696-8705 8706-8715 8716-8725 8726-8735 8736-8745 8746-8755 8756-8765 8766-8775 8776-8785 8786-8795 8796-8805 8806-8815 8816-8825 8826-8835 8836-8845 8846-8855 8856-8865 8866-8875 8876-8885 8886-8895 8896-8905 8906-8915 8916-8925 8926-8935 8936-8945 8946-8955 8956-8965 8966-8975 8976-8985 8986-8995 8996-9005 9006-9015 9016-9025 9026-9035 9036-9045 9046-9055 9056-9065 9066-9075 9076-9085 9086-9095 9096-9105 9106-9115 9116-9125 9126-9135 9136-9145 9146-9155 9156-9165 9166-9175 9176-9185 9186-9195 9196-9205 9206-9215 9216-9225 9226-9235 9236-9245 9246-9255 9256-9265 9266-9275 9276-9285 9286-9295 9296-9305 9306-9315 9316-9325 9326-9335 9336-9345 9346-9355 9356-9365 9366-9375 9376-9385 9386-9395 9396-9405 9406-9415 9416-9425 9426-9435 9436-9445 9446-9455 9456-9465 9466-9475 9476-9485 9486-9495 9496-9505 9506-9515 9516-9525 9526-9535 9536-9545 9546-9555 9556-9565 9566-9575 9576-9585 9586-9595 9596-9605 9606-9615 9616-9625 9626-9635 9636-9645 9646-9655 9656-9665 9666-9675 9676-9685 9686-9695 9696-9705 9706-9715 9716-9725 9726-9735 9736-9745 9746-9755 9756-9765 9766-9775 9776-9785 9786-9795 9796-9805 9806-9815 9816-9825 9826-9835 9836-9845 9846-9855 9856-9865 9866-9875 9876-9885 9886-9895 9896-9905 9906-9915 9916-9925 9926-9935 9936-9945 9946-9955 9956-9965 9966-9975 9976-9985 9986-9995 9996-10005

Z 0146

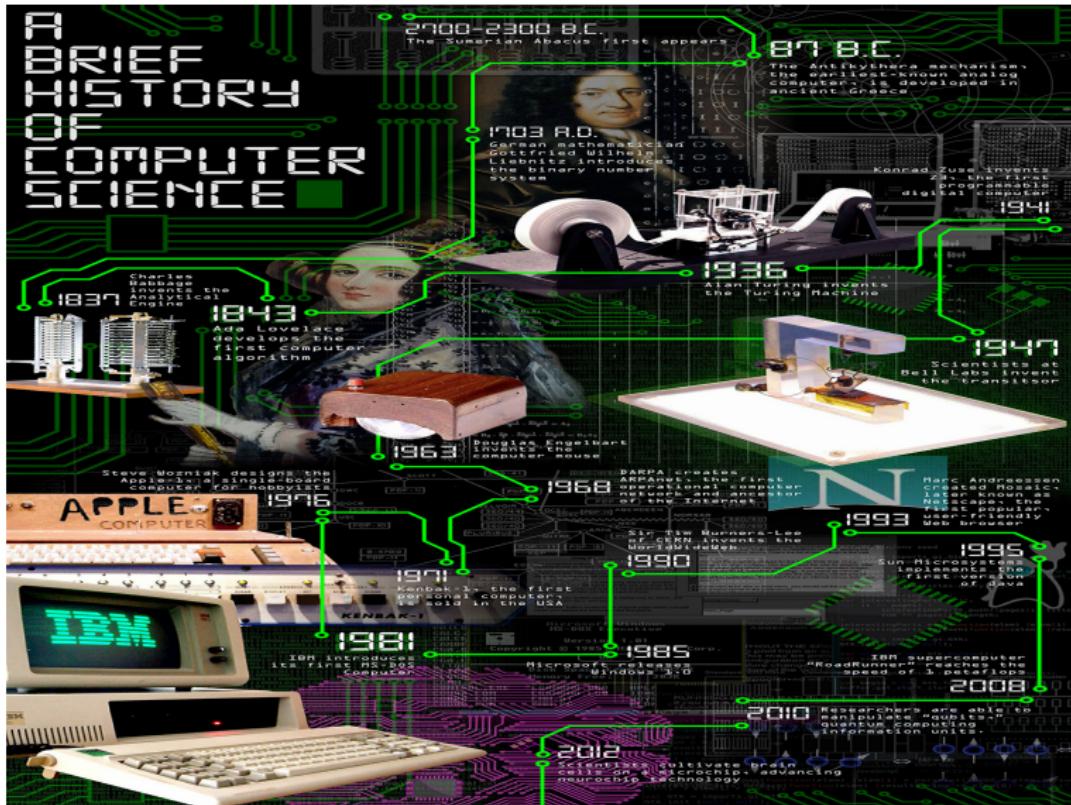
HMM profile

教学提纲

- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题





in vivo

In vivo 为拉丁文 “在活体内” 之意。在科学文献中，*in vivo* 常指进行于完整且存活的个体内的组织的实验，以区别在生物体上移除下来的组织或死亡的组织上进行的实验（对应的拉丁文为 *in vitro*）。

in vitro

In vitro 是拉丁语中 “在玻璃里” 的意思，意指进行或发生于试管内的实验与实验技术。更广义的意思，则指活生物体之外的环境中的操作。

in silico

In silico 是指 “在硅之中”，也就是说 “进行于电脑中，或是经由电脑模拟” 之意，此用语是衍生自另外两个在生物学上常用的短语：*in vivo*（生物活体内）及 *in vitro*（生物活体外）。

in vivo

In vivo 为拉丁文 “在活体内” 之意。在科学文献中，*in vivo* 常指进行于完整且存活的个体内的组织的实验，以区别在生物体上移除下来的组织或死亡的组织上进行的实验（对应的拉丁文为 *in vitro*）。

in vitro

In vitro 是拉丁语中 “在玻璃里” 的意思，意指进行或发生于试管内的实验与实验技术。更广义的意思，则指活生物体之外的环境中的操作。

in silico

In silico 是指 “在硅之中”，也就是说 “进行于电脑中，或是经由电脑模拟” 之意，此用语是衍生自另外两个在生物学上常用的短语：*in vivo*（生物活体内）及 *in vitro*（生物活体外）。

in vivo

In vivo 为拉丁文 “在活体内” 之意。在科学文献中，*in vivo* 常指进行于完整且存活的个体内的组织的实验，以区别在生物体上移除下来的组织或死亡的组织上进行的实验（对应的拉丁文为 *in vitro*）。

in vitro

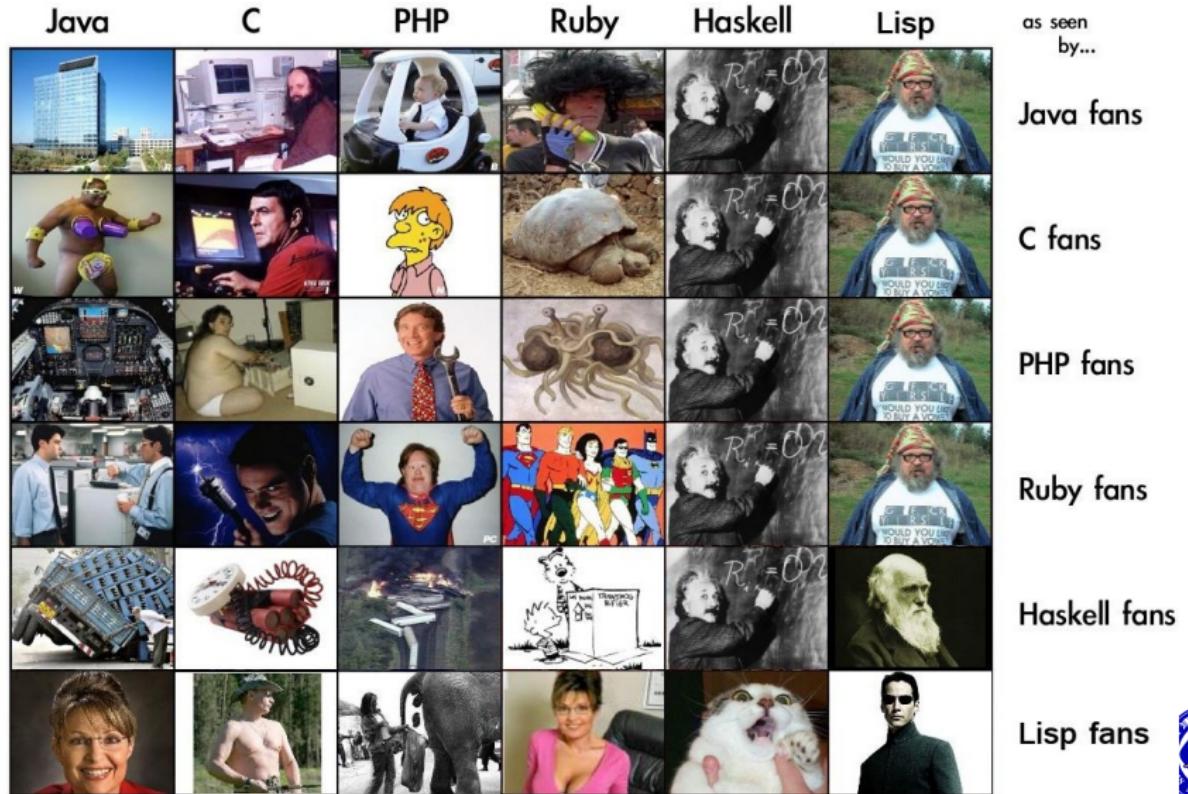
In vitro 是拉丁语中 “在玻璃里” 的意思，意指进行或发生于试管内的实验与实验技术。更广义的意思，则指活生物体之外的环境中的操作。

in silico

In silico 是指 “在硅之中”，也就是说 “进行于电脑中，或是经由电脑模拟” 之意，此用语是衍生自另外两个在生物学上常用的短语：*in vivo*（生物活体内）及 *in vitro*（生物活体外）。



计算机科学 | 编程语言 | 崇拜/鄙视链



	C++		JavaScript
	Java/C#		PHP(Without MySQL)
	Ruby		Pascal
	Perl		Lisp
	Visual Basic		Haskell
	Python		C

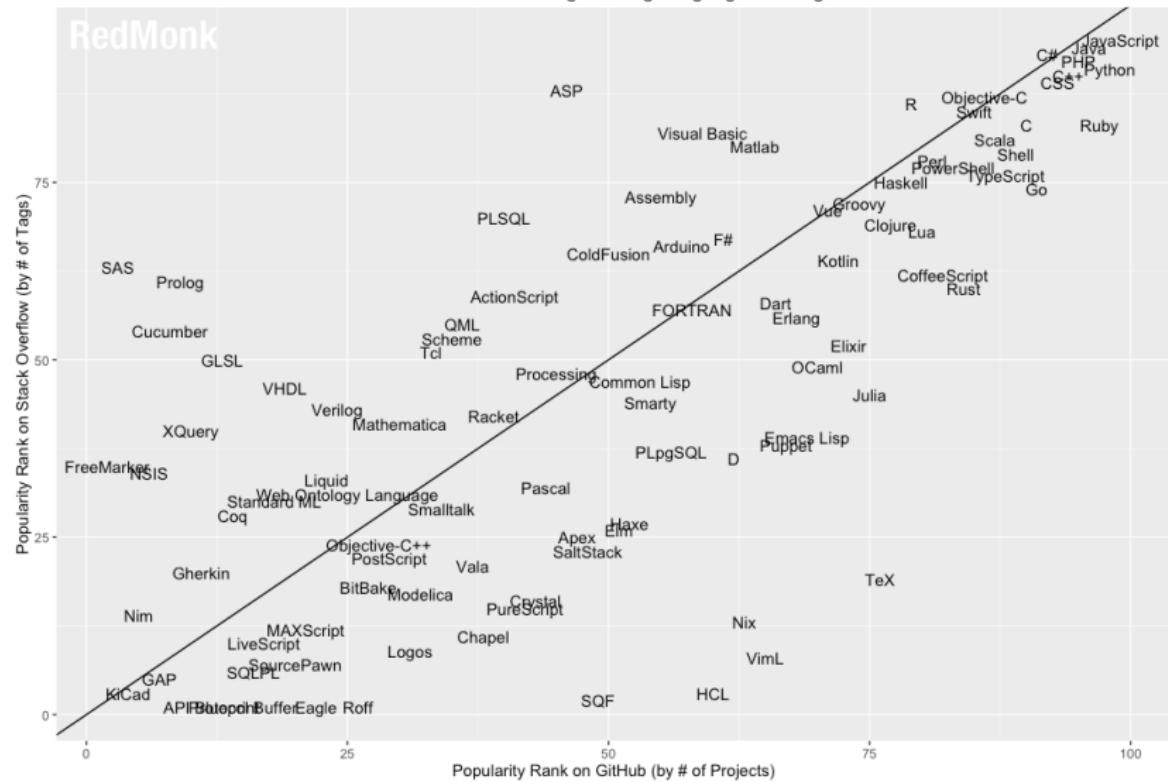
语言	宗教	超级英雄	《哈利波特》
C	犹太教	—	伏地魔
C++	伊斯兰教	—	西弗勒斯·斯内普
Java	正统基督教	万磁王	洛雷斯·乌姆里奇
Lisp	佛教	Xavier 教授	—
Perl	巫毒教	—	罗恩·韦斯莱
PHP	Cafeteria 基督教	小丑王	德拉科·马尔福
Python	人文主义	蝙蝠侠	哈利·波特
Ruby	新异教主义	钢铁侠	—
shell	—	—	鲁伯·海格



语言	女人	武器	船
C	霸道女总裁	M1 式加兰德步枪	核潜艇
C++	—	双截棍	—
Java	娇妻贤内助	M240 通用弹夹式自动机枪	大货船
Lisp	女博士	剃须刀	—
Perl	—	燃烧弹	拖船
PHP	—	水管子	竹筏
Python	万人迷	双管枪	—
Ruby	—	宝刀	摩托艇
shell	女公务员	锤子	—



RedMonk Q318 Programming Language Rankings



计算机科学 | 编程语言 | 排名 | 2018

Oct 2018	Oct 2017	Change	Programming Language	Ratings	Change
1	1		Java	17.801%	+5.37%
2	2		C	15.376%	+7.00%
3	3		C++	7.593%	+2.59%
4	5	▲	Python	7.156%	+3.35%
5	8	▲	Visual Basic .NET	5.884%	+3.15%
6	4	▼	C#	3.485%	-0.37%
7	7		PHP	2.794%	+0.00%
8	6	▼	JavaScript	2.280%	-0.73%
9	-	▲	SQL	2.038%	+2.04%
10	16	▲	Swift	1.500%	-0.17%
11	13	▲	MATLAB	1.317%	-0.56%
12	20	▲	Go	1.253%	-0.10%
13	9	▼	Assembly language	1.245%	-1.13%
14	15	▲	R	1.214%	-0.47%
15	17	▲	Objective-C	1.202%	-0.31%
16	12	▼	Perl	1.168%	-0.80%
17	11	▼	Delphi/Object Pascal	1.154%	-1.03%
18	10	▼	Ruby	1.108%	-1.22%
19	19		PL/SQL	0.779%	-0.63%
20	18	▼	Visual Basic	0.652%	-0.77%



计算机科学 | 编程语言 | 排名 | 历史

Programming Language	2018	2013	2008	2003	1998	1993	1988
Java	1	2	1	1	17	-	-
C	2	1	2	2	1	1	1
C++	3	4	3	3	2	2	4
Python	4	7	6	11	24	13	-
C#	5	5	7	8	-	-	-
Visual Basic .NET	6	11	-	-	-	-	-
PHP	7	6	4	5	-	-	-
JavaScript	8	9	8	7	21	-	-
Ruby	9	10	9	18	-	-	-
R	10	23	48	-	-	-	-
Objective-C	14	3	40	50	-	-	-
Perl	16	8	5	4	3	9	22
Ada	29	19	18	15	12	5	3
Lisp	30	12	16	13	8	6	2
Fortran	31	24	21	12	6	3	15



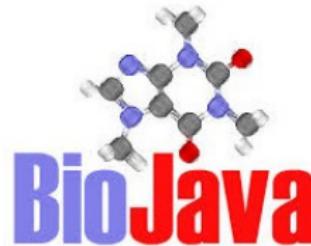
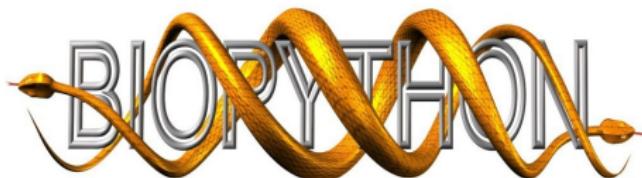
生物信息学常用编程语言

Perl (1987) 、 Python (1991) 、 Ruby (1995) 、 Java (1995)



生物信息学专用

BioPerl、Biopython、BioRuby、BioJava



教学提纲

- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题



不需要！

- 有现成的工具可以使用
- 五湖四海皆兄弟
- 有钱能使磨推鬼
-

需要！

- 没有现成的工具
- 兄弟们都在忙着混江湖
- 工资涨如龟速，物价飙如赤兔
-



不需要！

- 有现成的工具可以使用
- 五湖四海皆兄弟
- 有钱能使磨推鬼
-

需要！

- 没有现成的工具
- 兄弟们都在忙着混江湖
- 工资涨如龟速，物价飙如赤兔
-



- 有助于理解现有工具的配置与个性化修改
- 编写程序来批量运行现有程序
- 数据分析很简单，前期数据处理很难很难……（二八定律，80/20 法则，帕雷托法则）
- 增强研究工作的可重复性
- 数据越来越大，时间越来越少
- 人生苦短，学习编程
-



编程有助于逻辑思维的锻炼，系统观的形成，以及创造能力和解决问题能力的培养。但这些都是潜移默化的，需要有一个积累的过程。

- 试错力：编程的试错成本比较低，在一次次修复 bug 的过程中，慢慢地就不怕去试错了
- 创造力：编程可以把想象变为现实，不断去创造奇迹
- 学习能力：编程独特的操作和思考性，不仅会让编程者爱上学习，还会帮助养成严谨、认真的学习习惯
- 逻辑思维：一个人语言组织、领导能力的基础
- 科学思维：编程就是用科学的眼光去思考问题、分析问题，更全面地去看待这个世界
-



编程思维

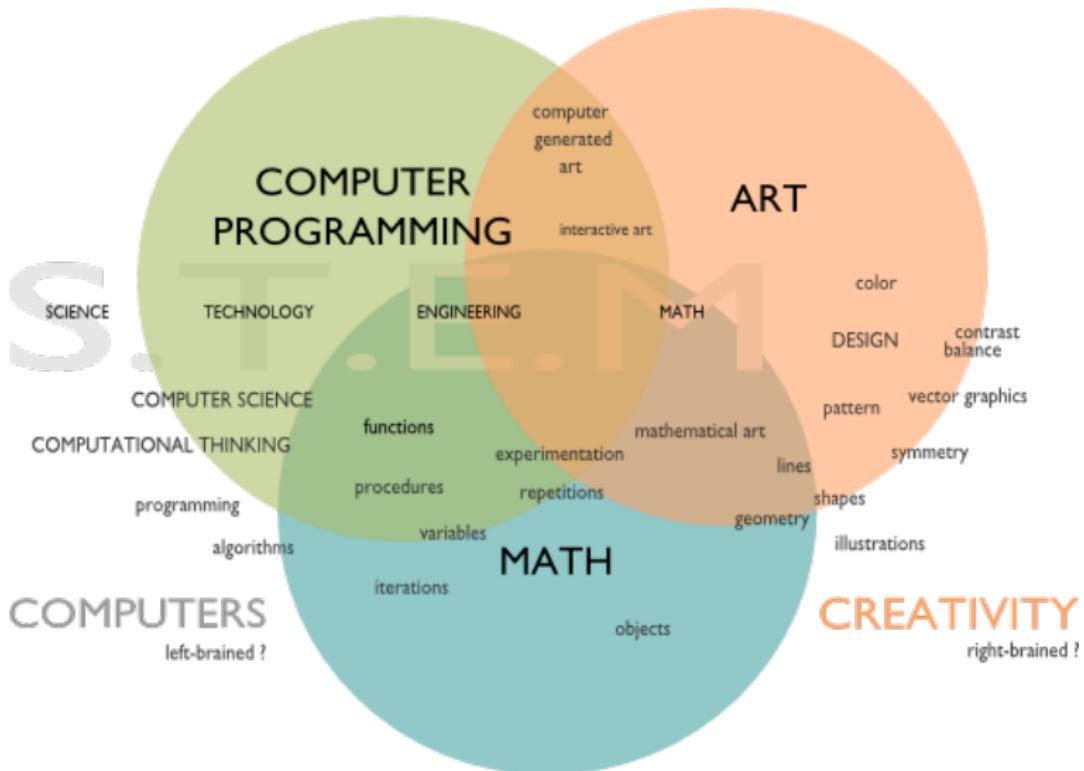
编程思维 (computational thinking) 就是 “理解问题-找出路径” 的思维过程，它由分解、模式识别、抽象、算法四个步骤组成。

通过这四个步骤，一个棘手的复杂问题先被拆解成一系列好解决的小问题；每一个小问题被单独检视、思考，搜索解决方法；然后，聚焦几个重要节点，忽视小细节，形成解决思路；最后，设计步骤，执行——问题解决！

通过编程做一个作品的过程，就是自己创造一个事物的过程。最起码首先要在心里大致构造出自己想要的作品模样，然后开始思考第一步该怎么做，第二步该怎么做。



编程 | 与艺术



每一位程序员都是一位艺术家 每一个程序都是一件艺术品 编写程序 = 艺术创作！



If you program enough,
it can change the way you look at the world ...



任务

- 同时精确测量粒子的位置和动量
- 寻找第四种可以镶嵌平面的凸六边形
- 寻找可以镶嵌平面的边数大于 6 的凸多边形
- 分析敲除所有必需基因后的基因表达

原理

- 海森堡不确定性原理
- 有且只有三种凸六边形可以镶嵌平面
- 边数大于 6 的凸多边形都无法镶嵌平面
- 都说了是必需基因……



任务

- 同时精确测量粒子的位置和动量
- 寻找第四种可以镶嵌平面的凸六边形
- 寻找可以镶嵌平面的边数大于 6 的凸多边形
- 分析敲除所有必需基因后的基因表达

原理

- 海森堡不确定性原理
- 有且只有三种凸六边形可以镶嵌平面
- 边数大于 6 的凸多边形都无法镶嵌平面
- 都说了是必需基因……



任务

- 同时精确测量粒子的位置和动量
- 寻找第四种可以镶嵌平面的凸六边形
- 寻找可以镶嵌平面的边数大于 6 的凸多边形
- 分析敲除所有必需基因后的基因表达

原理

- 海森堡不确定性原理
- 有且只有三种凸六边形可以镶嵌平面
- 边数大于 6 的凸多边形都无法镶嵌平面
- 都说了是必需基因……



任务

- 同时精确测量粒子的位置和动量
- 寻找第四种可以镶嵌平面的凸六边形
- 寻找可以镶嵌平面的边数大于 6 的凸多边形
- 分析敲除所有必需基因后的基因表达

原理

- 海森堡不确定性原理
- 有且只有三种凸六边形可以镶嵌平面
- 边数大于 6 的凸多边形都无法镶嵌平面
- 都说了是必需基因……



任务

- 破解 RSA 秘钥
- 暴力破解由 94 个字符（26 小写、26 大写、10 数字、32 标点）随机组合成的长 12 个字符的密码
- 蛋白质折叠中通过随机尝试找到总能量最低的构象状态

原因

- 对极大整数做因数分解非常困难（破解 RSA-2048（2048-bit）的密钥可能需要耗费传统电脑 10 亿年的时间，而量子计算机只需要 100 秒就可以完成。）【量子计算机 + 秀尔算法】/ 【黎曼猜想】
- 普通台式机，每秒运算 40 亿次，需要大约 30 万年以上才能破解利文索尔佯谬（Levinthal's paradox）：100 氨基酸，每个 2 种构象，每次尝试耗时 10^{-13} s，穷举需要 40 亿年【深度学习？】

任务

- 破解 RSA 秘钥
- 暴力破解由 94 个字符（26 小写、26 大写、10 数字、32 标点）随机组合成的长 12 个字符的密码
- 蛋白质折叠中通过随机尝试找到总能量最低的构象状态

原因

- 对极大整数做因数分解非常困难（破解 RSA-2048（2048-bit）的密钥可能需要耗费传统电脑 10 亿年的时间，而量子计算机只需要 100 秒就可以完成。）【量子计算机 + 秀尔算法】/【黎曼猜想】
- 普通台式机，每秒运算 40 亿次，需要大约 30 万年以上才能破解利文索尔佯谬（Levinthal's paradox）：100 氨基酸，每个 2 种构象，每次尝试耗时 10^{-13} s，穷举需要 40 亿年【深度学习？】

任务

- 破解 RSA 秘钥
- 暴力破解由 94 个字符（26 小写、26 大写、10 数字、32 标点）随机组合成的长 12 个字符的密码
- 蛋白质折叠中通过随机尝试找到总能量最低的构象状态

原因

- 对极大整数做因数分解非常困难（破解 RSA-2048（2048-bit）的密钥可能需要耗费传统电脑 10 亿年的时间，而量子计算机只需要 100 秒就可以完成。）【量子计算机 + 秀尔算法】/【黎曼猜想】
- 普通台式机，每秒运算 40 亿次，需要大约 30 万年以上才能破解
- 利文索尔佯谬（Levinthal's paradox）：100 氨基酸，每个 2 种构象，每次尝试耗时 10^{-13} s，穷举需要 40 亿年【深度学习？】

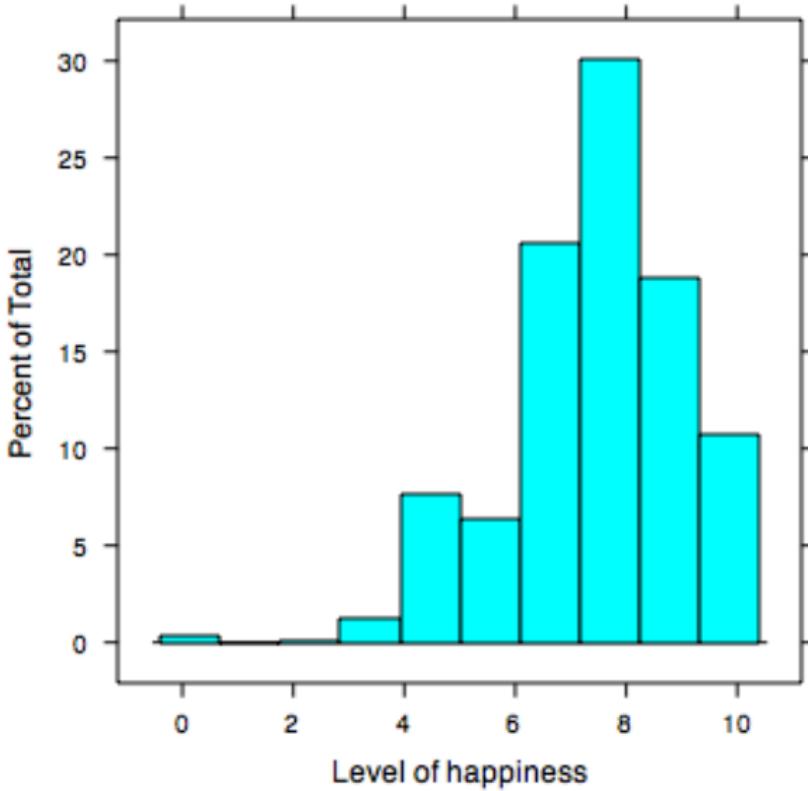
教学提纲

- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

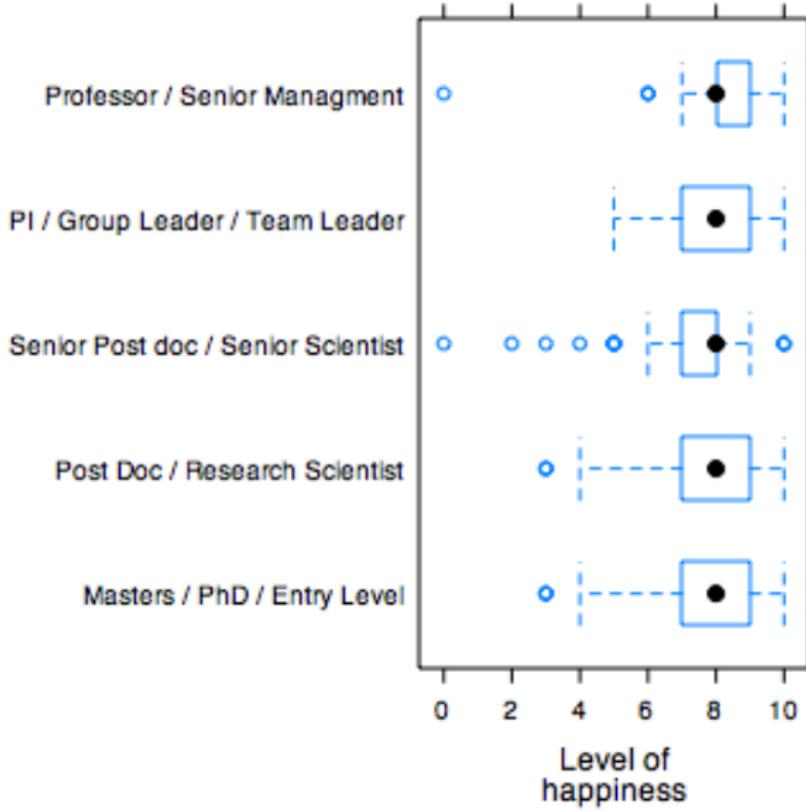
- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题



2008 | happiness

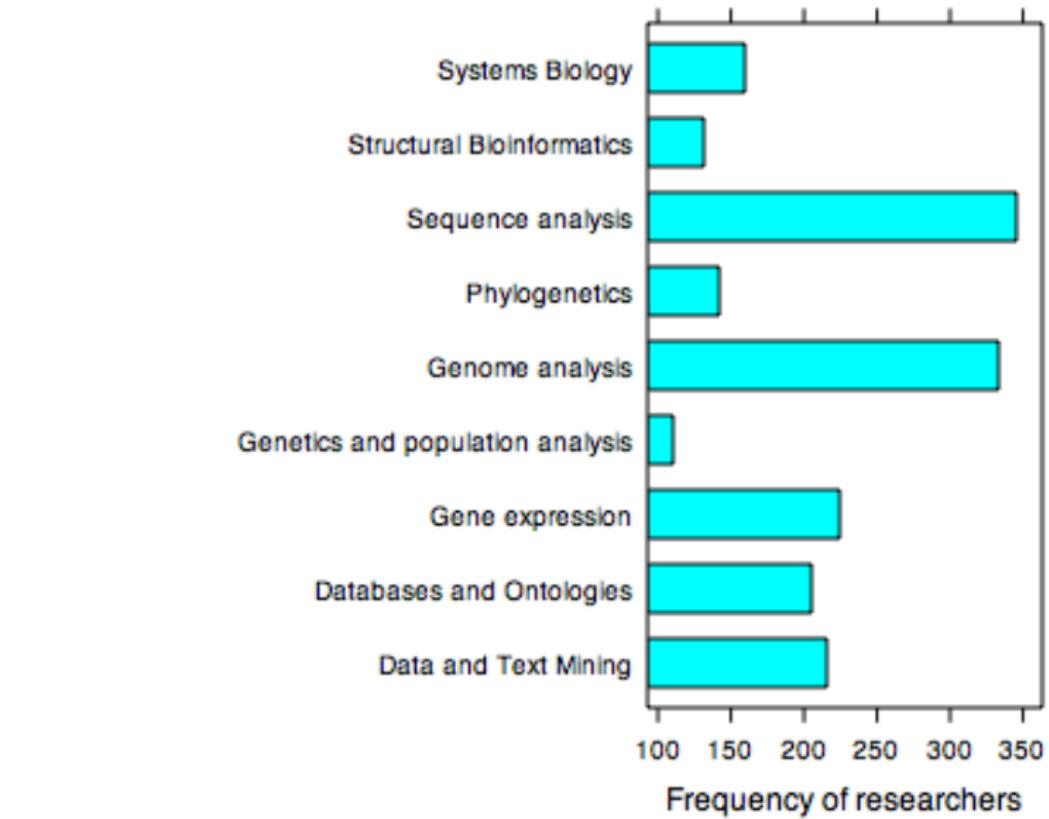


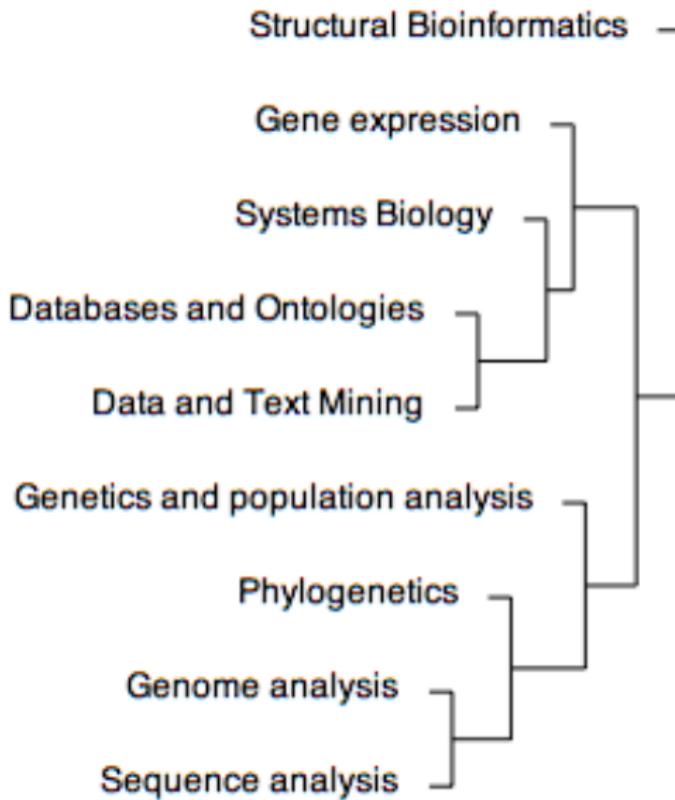
2008 | happiness vs. career position



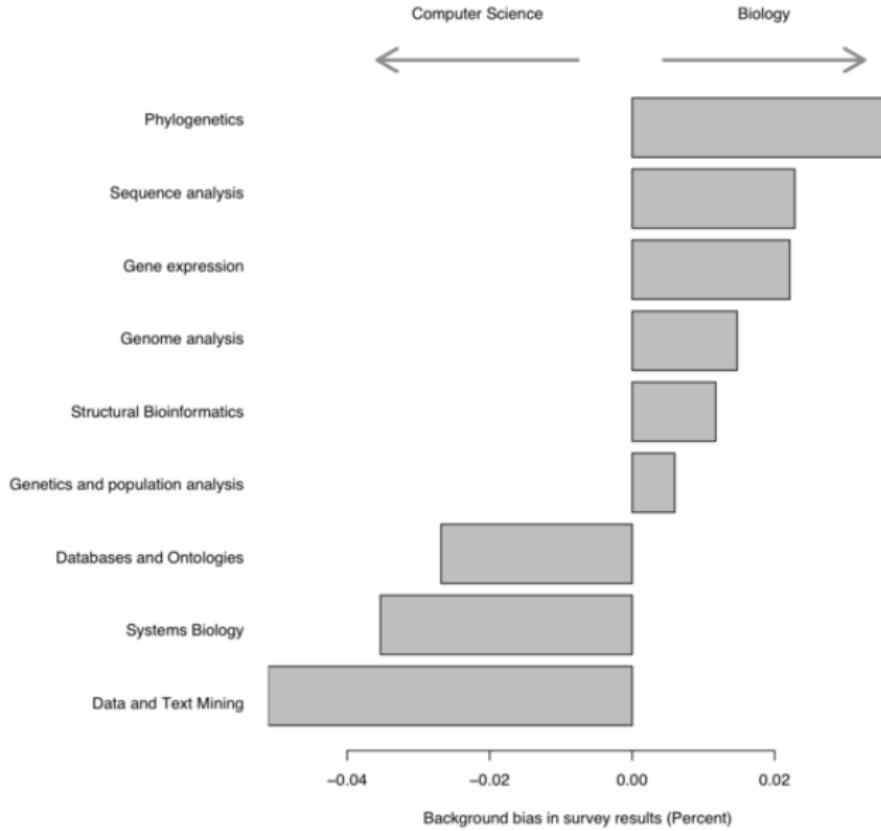
2008 | likes vs. dislikes



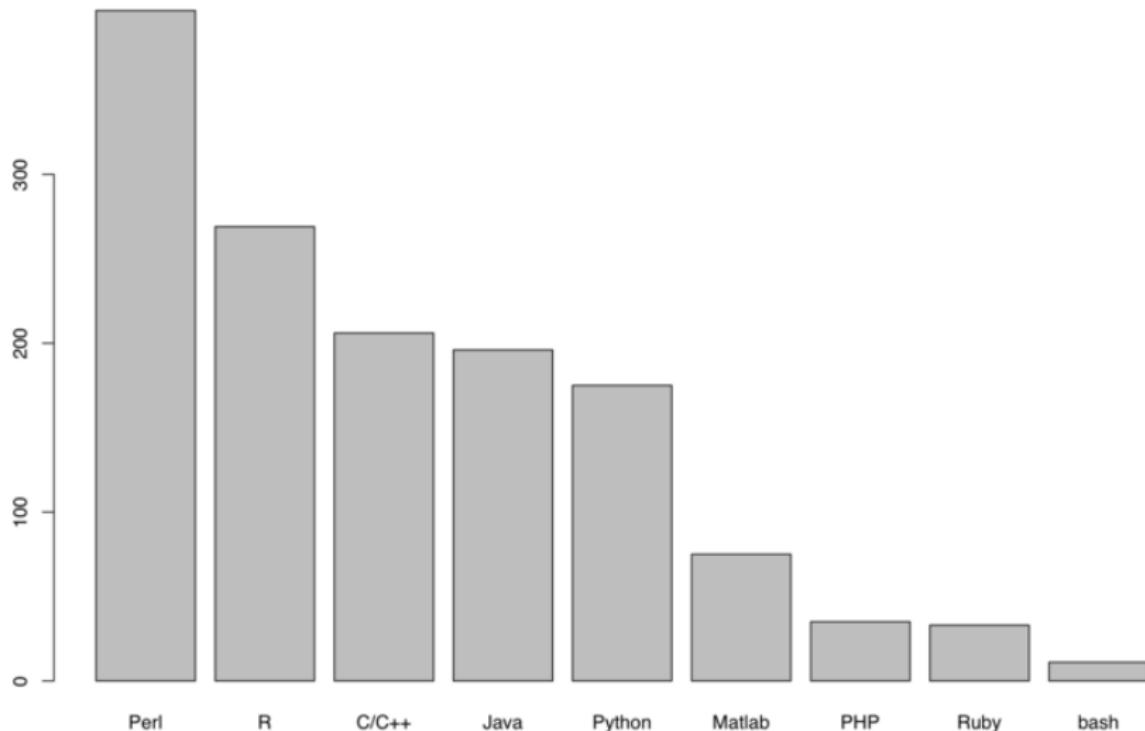




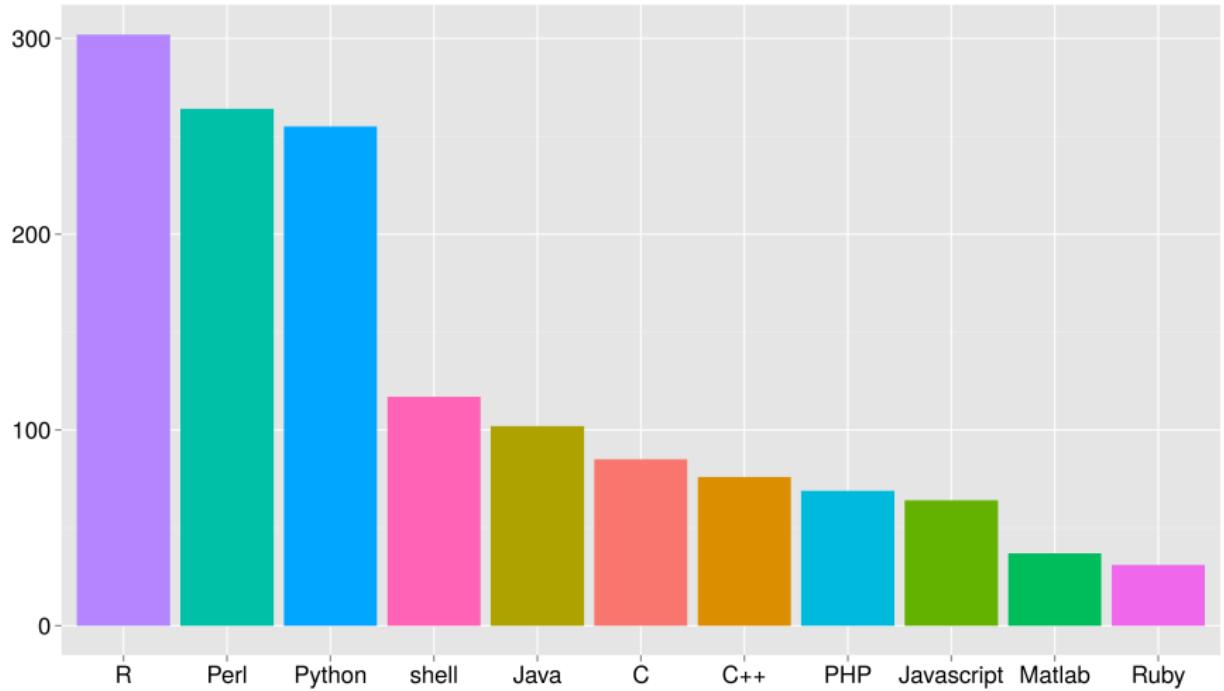
2008 | background

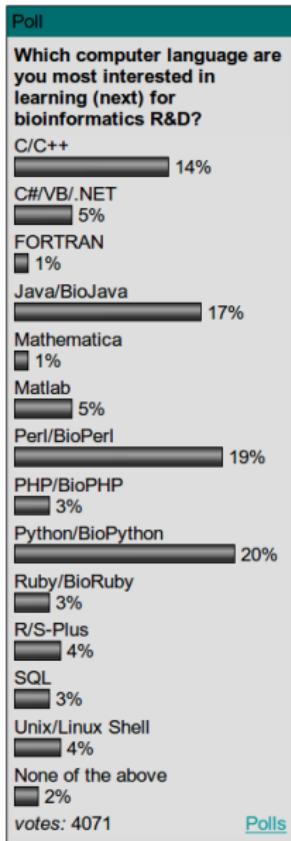


2008 | language

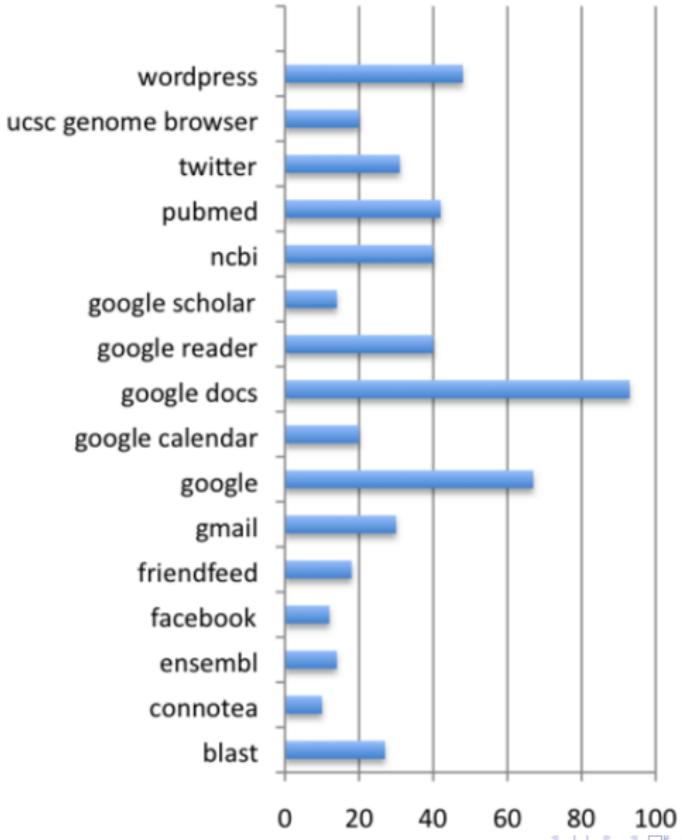


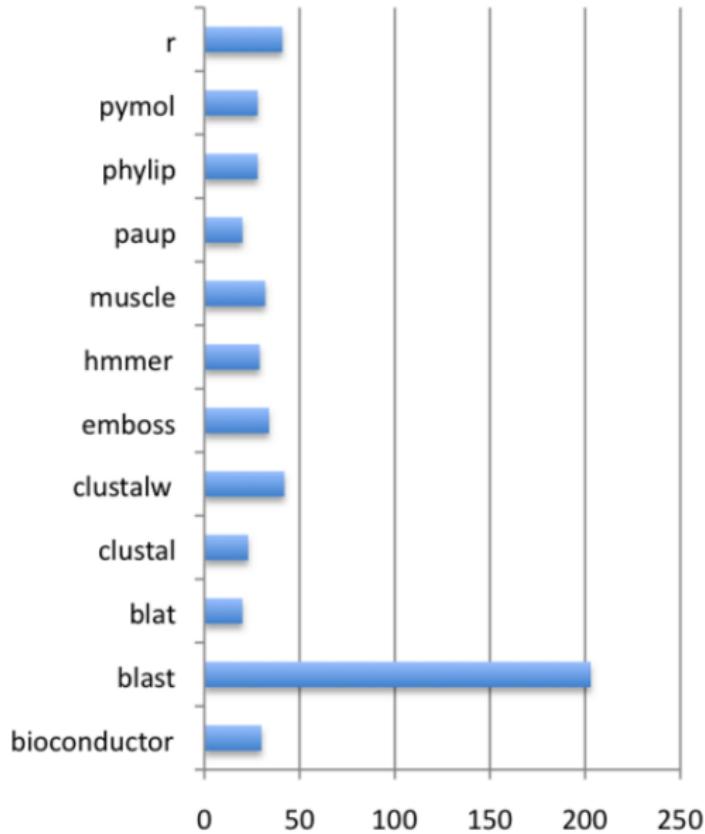
2012 | language





2008 | applications





industry vs. academic

- Academic Salary Mean/Median: \$36,520 / \$33,712
- Industry Salary Mean/Median: \$66,239 / \$64,235

Salary/Years Experience

- Academic Mean/Median: \$10,970 / \$8,333
- Industry Mean/Median: \$17,410 / \$12,000



industry vs. academic

- Academic Salary Mean/Median: \$36,520 / \$33,712
- Industry Salary Mean/Median: \$66,239 / \$64,235

Salary/Years Experience

- Academic Mean/Median: \$10,970 / \$8,333
- Industry Mean/Median: \$17,410 / \$12,000



- Bioinformatics Career Survey 2008
- Bioinformatics Career Survey 2008 Results
- bioinformatics-career-survey (GitHub)
- A comparison of common programming languages used in bioinformatics
- A comparison of bioinformatics programming languages
- Programming Languages of Bioinformatics
- A global perspective on evolving bioinformatics and data science training needs



教学提纲

- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题



R

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

R 语言，一种自由软体程式语言与操作环境，主要用于**统计分析、绘图、数据挖掘**。R 本来是由来自新西兰奥克兰大学的罗斯·伊哈卡和罗伯特·杰特曼开发（也因此称为 R），现在由“R 开发核心团队”负责开发。

- R 内建多种统计学及数字分析功能。R 的功能也可以透过安装套件 (Packages, 用户撰写的功能) 增强。
- R 的另一强项是绘图功能，制图具有印刷的素质，也可加入数学符号。
- 虽然 R 主要用于统计分析或者开发统计相关的软体，但也有人用作矩阵计算。其分析速度可媲美专用于矩阵计算的自由软件 GNU Octave 和商业软件 MATLAB。

Packages

R 的功能能够透过由用户撰写的套件增强。增加的功能有特殊的统计技术、绘图功能，以及编程介面和数据输出/输入功能。这些软件包是由 R 语言、LaTeX、Java 及最常用 C 语言和 Fortran 撰写。

CRAN

CRAN 为 Comprehensive R Archive Network (R 综合典藏网) 的简称。它除了收藏了 R 的执行档下载版、原始码和说明文件，也收录了各种用户撰写的软件包。现时，全球有超过一百个 CRAN 镜像站。

Bioconductor

生物信息学社群时常使用 R 进行分子生物学数据分析。Bioconductor 计划就是让 R 作为基因图谱分析工具。

Packages

R 的功能能够透过由用户撰写的套件增强。增加的功能有特殊的统计技术、绘图功能，以及编程介面和数据输出/输入功能。这些软件包是由 R 语言、LaTeX、Java 及最常用 C 语言和 Fortran 撰写。

CRAN

CRAN 为 Comprehensive R Archive Network (R 综合典藏网) 的简称。它除了收藏了 R 的执行档下载版、原始码和说明文件，也收录了各种用户撰写的软件包。现时，全球有超过一百个 CRAN 镜像站。

Bioconductor

生物信息学社群时常使用 R 进行分子生物学数据分析。Bioconductor 计划就是让 R 作为基因图谱分析工具。

Packages

R 的功能能够透过由用户撰写的套件增强。增加的功能有特殊的统计技术、绘图功能，以及编程介面和数据输出/输入功能。这些软件包是由 R 语言、LaTeX、Java 及最常用 C 语言和 Fortran 撰写。

CRAN

CRAN 为 Comprehensive R Archive Network (R 综合典藏网) 的简称。它除了收藏了 R 的执行档下载版、原始码和说明文件，也收录了各种用户撰写的软件包。现时，全球有超过一百个 CRAN 镜像站。

Bioconductor

生物信息学社群时常使用 R 进行分子生物学数据分析。Bioconductor 计划就是让 R 作为基因图谱分析工具。

RStudio

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).

RCode

A modern environment for R.



RStudio

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).

RCode

A modern environment for R.



Useful and popular packages: Load data

readr readr makes it easy to read many types of **tabular data** including: delimited files, fixed width files and web log files.

readxl Import **excel files** into R. readxl supports both the legacy .xls format and the modern xml-based .xlsx format.

XLConnect Help you read, write and format **Micorsoft Excel** files from R.

foreign Functions for reading and writing data stored by some versions of Epi Info, Minitab, S, SAS, SPSS, Stata, Systat and Weka and for reading and writing some dBase files.

httr A set of useful tools for working with **http** connections.

XML Read and create **XML** documents with R.

jsonlite Read and create **JSON** data tables with R.

Useful and popular packages: Manipulate data

tidyverse Tools for changing the layout of your data sets. Use the gather and spread functions to convert your data into the **tidy format**, the layout R likes best.

magrittr magrittr provides a mechanism for chaining commands with a new **forward-pipe operator**, `%>%`.

dplyr Essential shortcuts for subsetting, summarizing, rearranging, and joining together data sets for **fast data manipulation**.



Useful and popular packages: Manipulate data

- lubridate** Tools that make working with **dates and times** easier.
- stringr** Easy to learn tools for **regular expressions and character strings**.
- plyr** plyr is a set of tools for a common set of problems: you need to **split** up a big data structure into homogeneous pieces, **apply** a function to each piece and then **combine** all the results back together.
- DT** The R package DT provides an R interface to the JavaScript library DataTables. R data objects (matrices or data frames) can be displayed as tables on HTML pages, and DataTables provides filtering, pagination, sorting, and many other features in the tables.

Useful and popular packages: Visualize data

ggplot2 R's famous package for making beautiful graphics.

ggplot2 lets you use **the grammar of graphics** to build layered, customizable plots.

ggvis **Interactive, web based graphics** built with the grammar of graphics.

DiagrammeR Create **graph diagrams and flowcharts** using R.

htmlwidgets The htmlwidgets package provides a framework for easily creating R bindings to JavaScript libraries.



Useful and popular packages: Report results

- rmarkdown** rmarkdown lets you insert R code into a **markdown document**. R then generates a final document, in a wide variety of formats, that replaces the R code with its results.
- knitr** knitr is an elegant, flexible and fast dynamic report generation that combines R with TeX, Markdown, or HTML. For open access publishing, and **reproducible research** in statistics.
- xtable** The xtable function takes an R object (like a data frame) and returns the latex or HTML code you need to paste a pretty version of the object into your documents. Copy and paste, or pair up with R Markdown.

Useful and popular packages: Report results

Shiny Easily make **interactive, web apps** with R. A perfect way to explore data and share findings with non-programmers.

shinydashboard shinydashboard makes it easy to use Shiny to create dashboards.



Useful and popular packages: High performance

- data.table** An alternative way to organize data sets for **very, very fast** operations. Useful for big data.
- parallel** Use **parallel processing** in R to speed up your code or to crunch large data sets.
- foreach** Provides Foreach Looping Construct for R. Foreach is an idiom that allows for iterating over elements in a collection, without the use of an explicit loop counter. Using foreach without side effects also facilitates **executing the loop in parallel**.
- doParallel** Foreach Parallel Adaptor for the ‘parallel’ Package. Provides a parallel backend for the **%dopar%** function using the parallel package.

Useful and popular packages: Development

- devtools** An essential suite of tools for turning your code into an R package.
- roxygen2** A quick way to **document** your R packages. roxygen2 turns inline code comments into documentation pages and builds a package namespace.
- testthat** testthat provides an easy way to write **unit tests** for your code projects.



Easily install and load packages from the tidyverse

The **tidyverse** is a collection of R packages that share common philosophies and are designed to work together.

The **core tidyverse packages** that you are likely to use in almost every analysis:

- `readr`, for data import.
- `tidyr`, for data tidying.
- `dplyr`, for data manipulations.
- `ggplot2`, for data visualisation.
- `stringr`, for string manipulations.
- `purrr`, for functional programming.
- `tibble`, for tibbles, a modern re-imagining of data frames.
- `forcats`, a suite of useful tools that solve common problems with factors.

Easily install and load packages from the tidyverse

It also installs a selection of other tidyverse packages that you're likely to use frequently, but probably not in every analysis. This includes packages for:

- Working with specific types of vectors: hms (times), stringr (strings), lubridate (date/times),forcats (factors), blob (blob/binary data).
- Importing other types of data: DBI (databases), haven (SPSS, SAS and Stata files), httr (web apis), jsonlite (JSON), readxl (.xls and .xlsx files), rvest (web scraping), xml2 (XML).
- Programming and modelling: magrittr (%>%), glue (alternative to paste()), modelr (modelling within a pipeline), broom (turning models into tidy data).

Others

viridis Use the color scales in this package to make plots that are pretty, better represent your data, easier to read by those with colorblindness, and print well in grey scale.

argparse A command line parser to be used with Rscript to write "#!" shebang scripts that gracefully accept positional and optional **arguments** and automatically generate usage.

ggtree Visualization and annotation of **phylogenetic trees**.

ComplexHeatmap ComplexHeatmap package provides a highly flexible way to arrange multiple **heatmaps** and supports self-defined annotation graphics.

... ...



Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development.

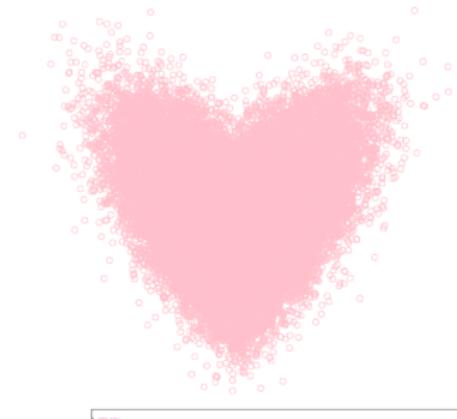


R | Example

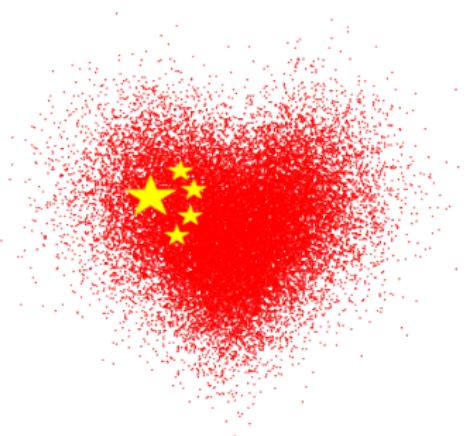
```
1 n <- 50000
2 r <- 0.7
3 r_e <- (1 - r * r) ^ 0.5
4 X <- rnorm(n)
5 Y <- X * r + r_e * rnorm(n)
6 Y <- ifelse(X>0, Y, -Y)
7 plot(X, Y, col="pink")
```



R | Example



LOVE
LOVE

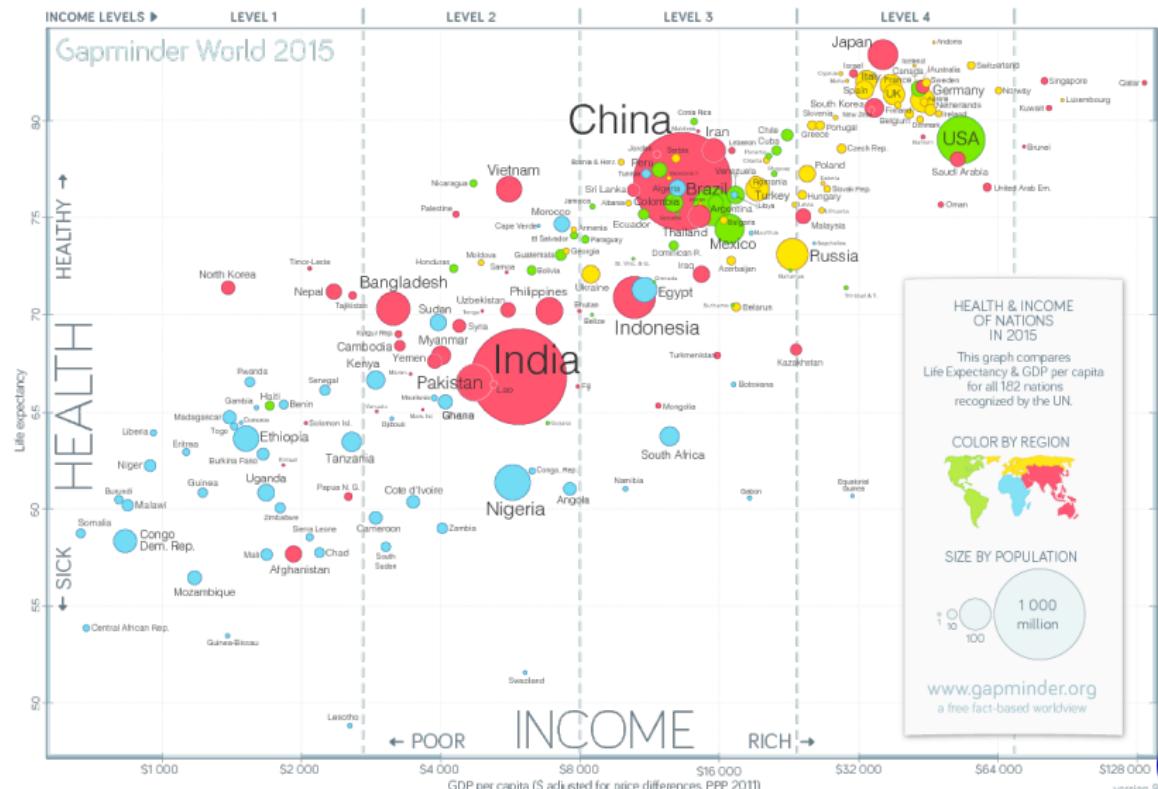


R | Example

```
1 xrange <- c(-15, 15); yrange <- c(0, 16)
2 plot(0, xlim=xrange, ylim=yrange, type="n")
3
4 yr <- seq(yrange[1], yrange[2], len=50)
5 offsetFn <- function(y) { 2 * sin(0 + y/3) }
6 offset <- offsetFn(yr)
7 leftE <- function(y) { -10 - offsetFn(y) }
8 rightE <- function(y) { 10 + offsetFn(y) }
9 xp <- c(leftE(yr), rev(rightE(yr)))
10 yp <- c(yr, rev(yr))
11 polygon(xp, yp, col="#ffeecc", border=NA)
12
13 h <- 9
14 xt <- seq(0, rightE(h), len=100)
15 yt <- log(1 + log(1 + log(xt + 1)))
16 yt <- yt - min(yt); yt <- h * yt/max(yt)
17 x <- c(leftE(h), rightE(h), rev(xt), -xt)
18 y <- c(h, h, rev(yt), yt)
19 polygon(x, y, col="red", border=NA)
```



R | Example



DATA SOURCES—INCOME: World Bank's GDP per capita, PPP 2011 (International \$) without additions by Gapminder. X-axis uses log scale to make a straight income line same distance on all levels. POPULATION: Numbers from UN Population Division. LIFE EXPECTANCY: <http://en.wikipedia.org/>—The interactive version of this chart is available at www.gapminder.org/world, which lets you animate historic data for hundreds of indicators. LICENSE: Our charts are freely available under Creative Commons Attribution License. Please copy them, modify them and even sell them, as long as you mention "Based on a free chart from gapminder.org".



R | Example | parallel

```
1 #!/usr/bin/Rscript
2
3 # print 1 to 10: use 20 seconds
4 system.time(
5   for (i in 1:10) {
6     Sys.sleep(2)
7     print(i)
8   }
9 )
```



R | Example | parallel

```
1 #!/usr/bin/Rscript
2 library(foreach)
3 library(doParallel)
4
5 # no_cores <- detectCores() - 1
6 no_cores <- 10
7 cl <- makeCluster(no_cores)
8 registerDoParallel(cl)
9
10 system.time(
11   x <- foreach (i=1:10) %dopar% {
12     Sys.sleep(2); print(i)
13   }
14 )
15
16 stopCluster(cl)
```



- The R Project for Statistical Computing
- RStudio
- Learn R
- RStudio Cheat Sheets
- R packages inspired by R and its community
- Quick list of useful R packages
- Bioconductor

教学提纲

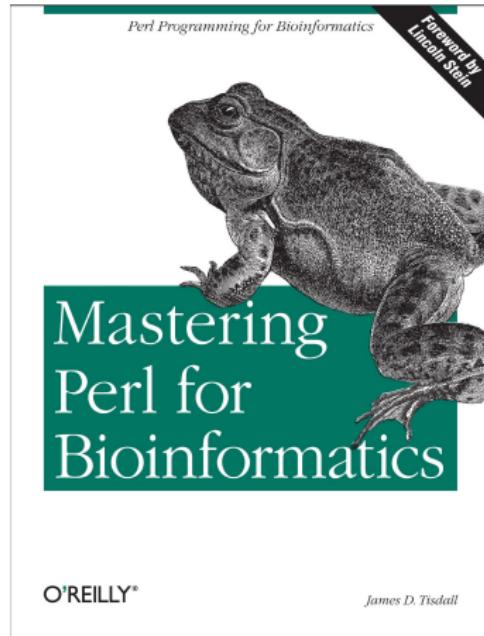
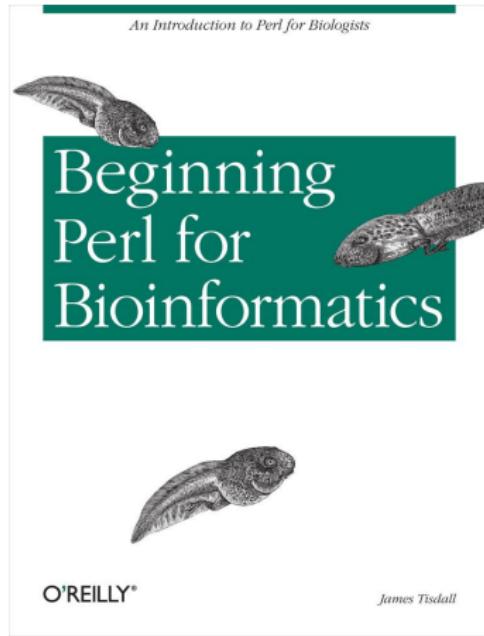
- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题



生物信息学角度

Beginning \Longrightarrow *Mastering Perl for Bioinformatics*



编程语言角度：代号

小骆驼 ⇒ 羊驼书 ⇒ 小羊驼母亲和她的孩子 ⇒ 大骆驼 ⇒ 黑豹书

英文名

Learning Perl ⇒ *Intermediate Perl* ⇒ *Mastering Perl* ⇒ *Programming Perl*
⇒ *Advanced Perl Programming*

中文名

《Perl 语言入门》⇒ 《Perl 进阶》⇒ 《精通 Perl》⇒ 《Perl 语言编程》
⇒ 《高级 Perl 编程》



编程语言角度：代号

小骆驼 ⇒ 羊驼书 ⇒ 小羊驼母亲和她的孩子 ⇒ 大骆驼 ⇒ 黑豹书

英文名

Learning Perl ⇒ *Intermediate Perl* ⇒ *Mastering Perl* ⇒ *Programming Perl*
⇒ *Advanced Perl Programming*

中文名

《Perl 语言入门》⇒ 《Perl 进阶》⇒ 《精通 Perl》⇒ 《Perl 语言编程》
⇒ 《高级 Perl 编程》



编程语言角度：代号

小骆驼 ⇒ 羊驼书 ⇒ 小羊驼母亲和她的孩子 ⇒ 大骆驼 ⇒ 黑豹书

英文名

Learning Perl ⇒ *Intermediate Perl* ⇒ *Mastering Perl* ⇒ *Programming Perl*
⇒ *Advanced Perl Programming*

中文名

《Perl 语言入门》⇒ 《Perl 进阶》⇒ 《精通 Perl》⇒ 《Perl 语言编程》
⇒ 《高级 Perl 编程》



编程语言角度：代号

小骆驼 ⇒ 羊驼书 ⇒ 小羊驼母亲和她的孩子 ⇒ 大骆驼 ⇒ 黑豹书

英文名

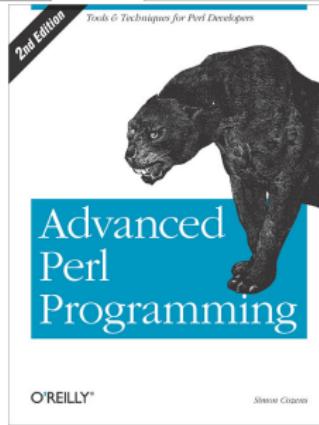
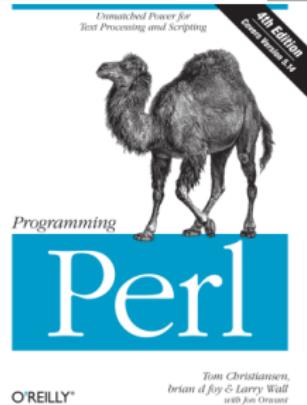
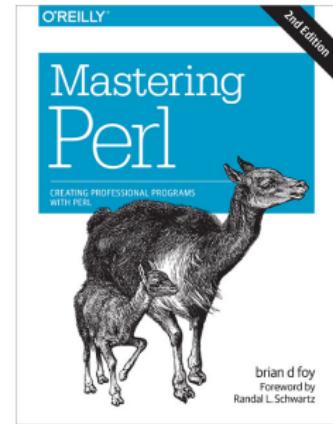
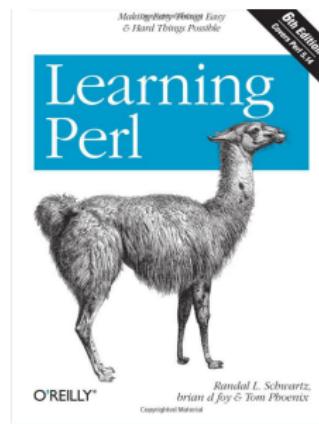
Learning Perl ⇒ *Intermediate Perl* ⇒ *Mastering Perl* ⇒ *Programming Perl*
⇒ *Advanced Perl Programming*

中文名

《Perl 语言入门》⇒ 《Perl 进阶》⇒ 《精通 Perl》⇒ 《Perl 语言编程》
⇒ 《高级 Perl 编程》



书籍 | Perl



中文名

- 《Perl 入门经典》
- 《高阶 Perl》
- 《Perl 高效编程》
- 《Perl 最佳实践》
- *Perl Cookbook*

英文名

- *Beginning Perl*
- *Higher-Order Perl*
- *Effective Perl Programming*
- *Perl Best Practices*
- *Perl Cookbook*



书籍 | Perl



中文名

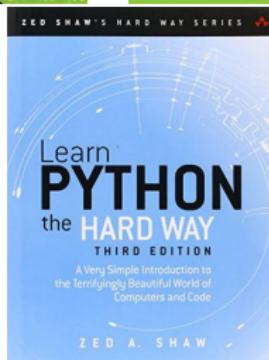
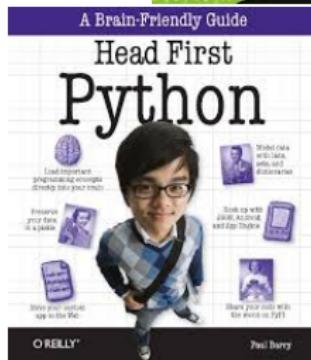
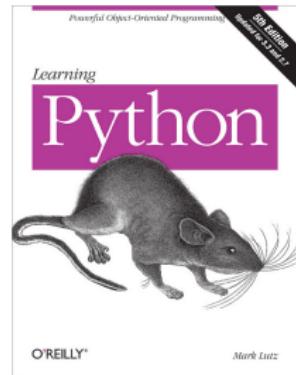
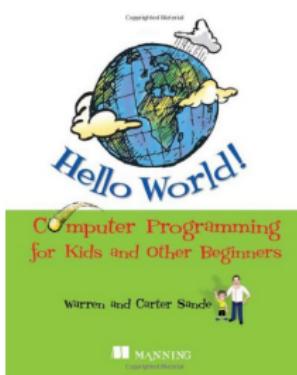
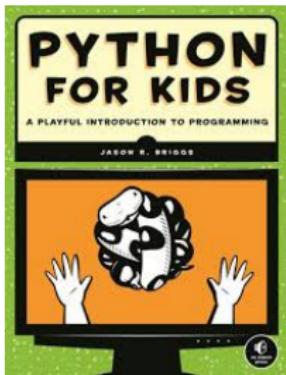
- 《趣学 Python 编程》
- 《父与子的编程之旅：与小卡特一起学 Python》
- 《Python 学习手册》
- 《深入浅出 Python》
- 《“笨办法”学 Python》
- 《像计算机科学家一样思考 Python》
- *Python Cookbook*

英文名

- *Python for Kids*
- *Computer Programming for Kids and Other Beginners*
- *Learning Python*
- *Head First Python*
- *Learn Python the Hard Way*
- *Think Python*
- *Python Cookbook*



书籍 | Python



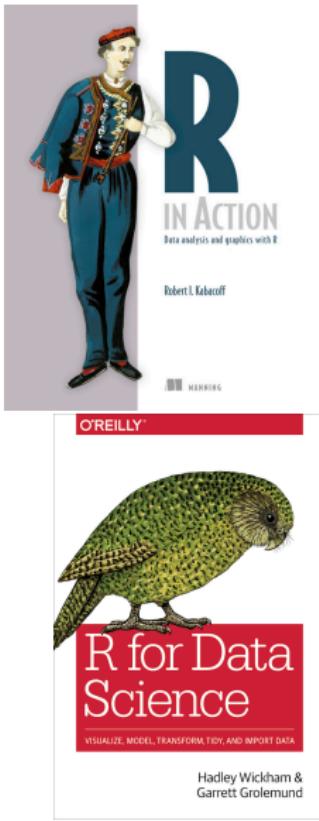
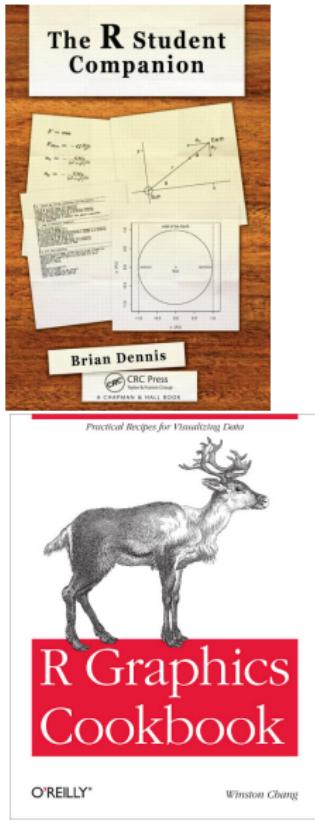
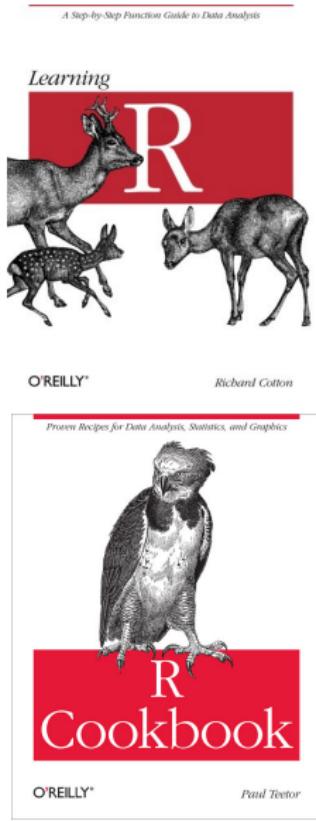
中文名

- 《学习 R》
- 《R 语言初学指南》
- 《R 语言实战》
- 《R 语言经典实例》
- 《R 数据可视化手册》
- 《R 数据科学》

英文名

- *Learning R*
- *The R Student Companion*
- *R in Action*
- *R Cookbook*
- *R Graphics Cookbook*
- *R for Data Science*





- 《Python 数据科学手册》
- 《Python 数据科学入门》
- 《Python 数据分析基础》
- 《Python 编程入门》
- 《Python 基础教程》
- 《Python 语言及其应用》
- 《Python 编程——从入门到实践》
- 《R 语言入门与实践》



计算机科学与程序设计

- 《Linux Shell 脚本攻略》
- 《我的第一本编程书》
- 《教孩子学编程》
- 《编程导论》
- 《深入浅出程序设计》
- 《程序是怎样跑起来的》
- 《写给大家看的算法书》
- 《啊哈！算法》
- 《算法的乐趣》
- 《算法帝国》
- 《算法笔记》
- 《轻松学算法》
- 《大话数据结构》
- 《程序员的数学》
- 《程序员的数学 2：概率统计》
- 《程序员的数学 3：线性代数》
- 《统计思维：程序员数学之概率统计》
- 《程序员的数学思维修炼》

数据科学

- 《数据科学入门》
- 《面向数据科学家的实用统计学》
- 《命令行中的数据科学》
- 《深入浅出数据分析》
- 《菜鸟侦探挑战数据分析》



统计学

- 《统计学》 (*Statistics for Engineers and the Sciences*)
- 《统计学及其应用》
- 《深入浅出统计学》
- 《从零开始学统计》
- 《从零开始读懂统计学》
- 《白话统计学》
- 《爱上统计学》
- 《赤裸裸的统计学》
- 《你一定爱读的极简统计学》
- 《介绍丛书: 统计学》
- 《漫画玩转统计学》
- 《漫画统计学》

统计学科普

- 《统计学的世界》
- 《统计学漫话》
- 《统计学七支柱》
- 《数字唬人》
- 《统计数字会撒谎》
- 《统计数据的真相》
- 《生活中的概率趣事》
- 《改变世界的 134 个概率统计故事》
- 《统计会犯错——如何避免数据分析中的统计陷阱》
- 《妙趣横生的统计学——培养大数据时代的统计思维》
- 《别拿相关当因果——因果关系简易入门》

教学提纲

- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题



教学提纲

- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题



知识点

- 生物信息学：交叉学科，多种技能，领域宽泛
- 生物学：DNA, RNA, 蛋白质
- 计算机科学：三个术语，编程语言
- R：readr、tidyr、dplyr、ggplot2 等常用包



教学提纲

- 1 课程安排
- 2 生物信息学
- 3 生物学
- 4 计算机科学
- 5 编程

- 6 Bioinformatics Career Survey
2008
- 7 R
- 8 书籍
- 9 回顾与总结
 - 总结
 - 思考题



- 生物信息学主要是由哪些学科交叉而来的？
- 生物信息学主要需要哪些方面的知识和技能？
- DNA 是由哪四种碱基组成的？RNA 与之有何不同？
- 列举常见的 20 种氨基酸，它们三字母和单字母的缩写分别是什么？
- 列举常见的存储 DNA 序列和蛋白质结构的数据库。
- *in vivo*、*in vitro* 和 *in silico* 分别代表什么含义？
- 列举常见的编程语言，在生物信息学中常用的编程语言，专用于生物信息学的工具集。
- 列举 R 的常用包并简介其主要用途。



Markdown

回顾、总结 Markdown 标记语言的基本语法：

- 标题
- 强调
- 列表
- 代码
- 引用
- 链接
-



Powered by



T_EX L^AT_EX X_ET_EX Beamer

