

## Introduction

大多是蛋白序列都是由相对少数目的始祖蛋白domain家族(ancestral protein domain families)所组成。因此，比对序列到已知的domain家族优于比对到已知的序列。蛋白序列分析就好比语言识别，不是将数字化的语音和已经说过的话来比较，而是使用机器学习的技术，对大量的，不同口音的语音训练得到统计模型，将输入与统计模型比较。同样，对于每个蛋白domain家族，它们典型的包含了上千个已知同源序列，这些同源序列能够比对成为深度的多重序列比对结构。这些序列比对，揭示了domain(domain)的结构和功能的特殊的进化保守的模式。

**HMMER软件可构建蛋白和DNA序列domain家族的概率模型，profile hidden Markov models, profile HMMs, or just profiles**，然后使用这些profile来注释新的序列，在序列数据库中搜索额外的同源序列，同时实现深度的多重序列比对(Pfam就是使用HMMs构建的蛋白domain模型数据库)。

蛋白domain同源家族的多重序列比对揭示了位置特异性的进化保守模式。关键的残基可能高度保守存在与某明确位置；某些位置可能能够接受明确的置换同时保有生化性质，例如疏水性，电荷或大小；某些位置可能进化成为近中性或多样性；某些位置可能比其他位置更容易出现插入和缺失。profile是一个位置特异性的打分模型，用来描述哪些特征更容易被观察到和多重序列比对中每个位置发生插入和删除的频率(a profile is a position-specific scoring model that describes which symbols are likely to be observed and how frequently insertions/deletions occur at each position(column) of a multiple sequence alignment)。

成对比较例如BLAST，使用的是BLOSUM和PAM矩阵，针对氨基酸置换仅有210种参数(20 X 20)，同时这些参数针对所有比对序列都一样使用。而profile针对一个约200氨基酸的蛋白domain的每一个位置拥有至少22个参数，同时这上千个参数用于评估一个特殊的蛋白家族比对，而不是所有的(每个位置约22个参数选择，是因为存在20种残基置换打分，加上gap open和gap extend罚分)。

## Problems HMMER is designed for

处理特殊序列家族，关注构建代表性的多重序列比对，HMMER hmmbuild程序帮助根据比对情况构建profile，hmmsearch程序用于在序列数据库中搜索profile系统性的查找更多的同源性。

HMMER3针对单个序列比对，不是多重序列比对，HMMER使用BLOSUM置换矩阵来构建特殊的profile HMMs。针对单个序列的蛋白数据库搜索，HMMER3拥有两种程序，phmmer和jackhmmer。phmmer优于BLASTP，jackhmmer优于PSI-BLAST(Position specific iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 in which a profile is automatically constructed from the first set of BLAST alignments.)。

HMMER3的hmmsearch程序可以在搜索profile数据库的同时，将序列解析成它的组成domain。Pfam数据库的构建是通过区分稳定收录的"seed"比对(少量的代表性序列)和"full"比对(所有检测到的同源性)来实现的。HMMER可用于构建seed模型，并且在数据库中搜索同源性，同时hmmalign程序能够通过将每条序列和seed序列比对自动地输出全比对情况。

HMMER使用的是组装算法，而不是最优比对算法。组装算法考虑了所有可能的比对，根据它们的相对相似性进行权重。HMMER在这展示比对情况时，能够明确展示比对的不确定性，它能注释比对的概率可信程度，或者每个比对的残基的概率可信程度。一些下游的分析可依靠该比对，例如系统发育树，可以区分可信的比对的残基。

HMMER2采用以下两种算法，由于运行速度考虑，默认采用Viterbi最佳比对算法。

Full probabilistic inference(the HMM Forward/Backward algorithms, ensemble log-odds sequence score)

Optimal alignment scores(the HMM Viterbi algorithm)

## Other implementations of profile HMM methods and position-specific scoring matrix methods

Software	URL
HH-suite	<a href="http://www.soding-conzentrion.lmu.de/software-and-servers-2">www.soding-conzentrion.lmu.de/software-and-servers-2</a>

## Usage

HMMER 3.2.1 (June 2018); <http://hmmer.org/>

HMMER自动检测输入文件格式，同时自动检测输入的序列或者包含的是核酸还是蛋白。

单比对序列格式包括：fasta, uniprot, genebank, ddbj和embl

多重比对文件格式包括：stockholm, afa(aligned FASTA), clustal, clustallike(MUSCLE, etc. ), a2m, phylip(interleaved), phylips(sequential), psiblast和selex

## 使用profile搜索序列数据库(searching a sequence database with a profile, protein)

### 1. 使用hmmbuild构建profile

```
hmmbuild globins4.hmm globins4.sto
```

单个hmmbuild命名足够将Pfam seed alignment flatfile(i.e. Pfam-A.seed)转换成profile flatfile(Pfam.hmm)

hmmbuild屏幕输出：

```
carlos@hughesmedbook: /P/15.34.00/Data_analysis/hmmer_test/tutorial1
$hmmbuild globins4.hmm globins4.sto
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.2.1 (June 2018); http://hmmer.org/
# Copyright (C) 2018 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# -----
# input alignment file:      globins4.sto
# output HMM file:          globins4.hmm
# -----
# idx name                  nseq  alen  mlen  eff_nseq  re/pos  description
#-----
1      globins4              4    171   149    0.96   0.589
# CPU time: 0.13u 0.00s 00:00:00.13 Elapsed: 00:00:00.13
carlos@hughesmedbook: /P/15.34.00/Data_analysis/hmmer_test/tutorial1
```

globins4比对包含了4个序列，共171个比对列(alen)。HMMER将其转换为149个consensus positions的profile(mlen)，这意味着它定义了包含22个gap的比对列；由于这4个序列互相之间相似，它们仅能够代表0.96的"effective"总序列数目；该profile具有每个位置的相对熵值(re/pos; average score), 0.598bits。

新生成的profile文件保存为globins4.hmm:

```
1 HMMER3/f [3.2.1 | June 2018]
2 NAME globins4
3 LENG 149
4 ALPH amino
5 RF no
6 MM no
7 CONS yes
8 CS no
9 MAP yes
0 DATE Fri May 31 15:40:02 2019
1 NSEQ 4
2 EFFN 0.964844
3 CKSUM 2027839109
4 STATS LOCAL MSV -9.9014 0.70957
5 STATS LOCAL VITERBI -10.7224 0.70957
6 STATS LOCAL FORWARD -4.1637 0.70957
7 HMM
8
9 COMPO
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
262
```

多个较低分值的domain，另外加入target 序列包含了多个一致性重复序列，那么这些序列bits score之和也可能看起来显著性，因此：

假如两个E-value的显著性都远小于1，那么target 序列可能是query序列的同源序列；假如全序列的E-value显著，而最佳domain的E-value不显著，那么target 序列很可能就是多重domain家族，但是要留意出现重复序列。

列表为#dom的两列为target 序列包含明确数目domain的评估值，第一个exp为根据HMMER's 统计模型的期待domain值；第二个为最终识别注释比对到target 序列的domain数目，这也是后面将会展示的比对数目，假如出现很大差异时，可能时target 序列出现了高度重复；最后两列为target序列及可选的描叙信息。

```
Domain annotation for each sequence (and alignments):
>> 7LESS_DROME RecName: Full=Protein sevenless; EC=2.7.10.1;
```

#	score	bias	c-Evalue	i-Evalue	hmmfrom	hmm to	alifrom	ali to	envfrom	env to	acc
1 ?	-1.5	0.0	0.18	0.18	60	72 ..	396	408 ..	395	410 ..	0.86
2 !	40.8	0.0	1.2e-14	1.2e-14	1	83 [.	439	520 ..	439	521 ..	0.95
3 !	14.8	0.0	1.5e-06	1.5e-06	12	84 ..	836	913 ..	826	914 ..	0.74
4 !	5.0	0.0	0.0017	0.0017	9	36 ..	1209	1236 ..	1203	1258 ..	0.83
5 !	22.4	0.0	6.7e-09	6.7e-09	13	79 ..	1313	1380 ..	1305	1385 ..	0.81
6 ?	0.6	0.0	0.04	0.04	55	72 ..	1753	1769 ..	1720	1769 ..	0.87
7 !	47.3	0.9	1.1e-16	1.1e-16	1	84 [.	1800	1890 ..	1800	1891 ..	0.91
8 !	17.0	0.0	1.6e-07	1.6e-07	5	72 ..	1004	1066 ..	1001	1076 ..	0.91

第三部分包含了每个序列的domain注释情况：

domain按照出现顺序依次排列，而不是根据显著性排列；！和？表示给domain是否满足per-sequence和per-domain inclusion thresholds，该阈值用于决定该匹配是否应认为是"true"，一般要求per-sequence E-value小于等于0.01，per-domain E-value小于等于0.01，reporting E-value常设为10.0；bit score和bias值为针对序列的分值，但只是针对one domain's envelope(现定义envelope bouding，然后再里面计算single best dom值)；接下累的c-Evalue表示conditional E-value，用于计算每个domain的统计显著性；第二个E-value为independent E-value，表示在整个数据库搜索中该序列的显著性，因此：

假如independent E-value远小于1，为显著性，表明该domain自身足够显著，以至于整个序列都可以认为是显著同源；

假如该序列已经存在一个或者多个高分值的domian，足够决定该序列为query序列的同源性序列，那么就可以查看conditonal E-value来寻找稍微低分值的domain。

接下列的列信息对应了hmm起点终点，target序列起点终点，对应的符号"... "表明比对发生在序列内部，"[]"表示比对超过了query或target末端，"[."和".]"对应了左侧和右侧的超出；envfrom和envto定义了target 序列的domain的envelope位置，推荐使用envelope位置来注释domian在target序列的位置；最后一列信息比对中每个残基的精确度。

```
34 Alignments for each domain:
35 == domain 1 score: -1.5 bits; conditional E-value: 0.18
36      EESSSTTEEEEEE CS
37      fn3 60 ltlgkpgteYevr 72
38      l+ L p+t+Y++r
39 7LESS_DROME 396 LEALIPYTQYRFR 408
40      67799*****8 PP
41
42 == domain 2 score: 40.8 bits; conditional E-value: 1.2e-14
43      TSBCEEEEEESSEEEEEEE-CSSSSSTCEEEEEEEETTSSSTEEEEEEESTCEEEEEESSTTEEEEEEEEEETTEEEE CS
44      fn3 1 saPsnlsvsevtstsltvsWeppkdgpgpitgYeveyrekgeeewneftvprtttsvtltgkpgteYevrVqavnggggep 83
45      saP ++ ++ l v+W p + +gpi+gY+++++++ + e+ vp+ s+ +++L++gt+Y++ + +n++gegp
46 7LESS_DROME 439 SAPVIEHLMGLDDSHLAVHWHWPGRFTNGPIEGYRLRLSSSEGNA-TSEQLVPAGRGSYIFSQIQAGTNYTLALSMINKQGEPP 520
47      5777788889999*****9998.*****997 PP
48
```

第四部分包含了domain的比对情况：

fn3开头的行为query profile的比对序列，大写字母表时非常保守位置，点(.)表示target序列相对于profile出现了插入；加号(+)表示阳性值，可以解释为保守置换，7LESS\_DROME开头的行为target序列，短横(-)比阿诗相对于profile，target出现了删除情况；最下面的行代表了每个比对残基的posterior概率(essentially the except accuracy)，0为0-5%，1为5-15%，9位85-95%，\*为95-100%，可以用这些值来判断哪部分比对是被较好的认定的。

```
106 Internal pipeline statistics summary:
107 -----
108 Query model(s):                      1 (85 nodes)
109 Target sequences:                    1 (2554 residues searched)
110 Passed MSV filter:                   1 (1); expected 0.0 (0.02)
111 Passed bias filter:                  1 (1); expected 0.0 (0.02)
112 Passed Vit filter:                   1 (1); expected 0.0 (0.001)
113 Passed Fwd filter:                   1 (1); expected 0.0 (1e-05)
114 Initial search space (Z):            1 [actual number of targets]
115 Domain search space (domZ):          1 [number of targets reported over threshold]
116 # CPU time: 0.00u 0.00s 00:00:00.00 Elapsed: 00:00:00.00
117 # Mc/sec: 26.30
118 //
119 [ok]
```

最后一部分为比对统计情况，

对应为query profile有85个consensus positions(nodes, mlen)；target序列有1条，共2554个残基；接下列是4个打分算法用于增加敏感度和计算需求，对应其期待值和实际过滤情况。

## 使用phmmer搜索单个蛋白序列(single sequence protein queries using phmmer, protein)

phmmer和hmmsearch一样，只是仅需要提供query序列而不是query profile

```
phmmer HBB_HUMAN globins45.fa
```

在globins45.fa文件中搜索和HBB\_HUMAN同源的domain，输出和hmmsearch一样

## 使用jackhmmer迭代搜索单个蛋白序列(iterative protein searches using jackhmmer)

jackhmmer在数据库中迭代搜索单个query序列，类似PSI-BLAST，第一轮搜索和phmmer一样，搜索所有比对domain，然后对这些序列构建profile，最后使用该profile搜索有数据库,迭代数目一致持续到没有新的序列发现，或指定迭代数目(-N，默认5)。

```
jackhmmer HBB_HUMAN globins45.fa
```

不同于phmmer的是jackhmmer标记"new"序列为"+", 同时"lost"序列为"-". "new"序列为序列通过了当前轮的阈值，但是没有通过上一轮阈值，"lost"相反。

jackhmmer一般会搜索很多轮才能完全搜索完，因此会有多个输出。

## 使用单个query序列搜索profile数据库(searching a profile database with a query sequence, protein)

相对于在单个profile中搜索多个序列，也可以在包含了不同doman的profile中搜索单个序列，进行注释。profile数据可能是Pfam，SMART或TIGRFams。

## 1. 构建profile数据库文件

profile数据可文件就是多个单个profile文件的组合，可以通过建立单独的profile 文件或将这些profile文件组合起来，或者将Stockholm比对文件组合起来，然后使用hmmbuild构建profile。

```
hmmbuild globins4.hmm globins4.sto
```

```
hmmbuild fn3.hmm fn3.sto
```

```
hmmbuild Pkinase.hmm Pkinase.sto
```

```
cat globins4.hmm fn3.hmm Pkinase.hmm > minifam
```

另外，所有比对文件中，只有Stockholm格式能够组合序列比对到相同的文件中，同时要求每个Stockholm文件都含#=GF ID，然后再使用hmmbuild构建profile数据库。对于单个比对，hmmbuild使用当前文件名称或者指定参数-n提供。

## 2. 使用hmmcompress压缩并索引flatfile

假如搜索，hmmscan可依靠二进压缩索引的flatfiles，因此，首先压缩索引profile文件

```
hmmcompress minifam
```

同时产生对应二进制文件

```
$hmmcompress minifam
Working... done.
Pressed and indexed 3 HMMs (3 names and 2 accessions).
Models pressed into binary file: minifam.h3m
SSI index for binary model file: minifam.h3i
Profiles (MSV part) pressed into: minifam.h3f
Profiles (remainder) pressed into: minifam.h3p
```

## 3. 使用hmmscan搜索profile数据库

```
hmmscan minifam 7LESS_DROME
```

输出文件的头文件和第一部分的和hmmsearch输出一样

hmmscan的search space的大小为profile数据库中profiles的数目(这里是3，对于Pfam搜索，为20000)。在hmmsearch中，search space的大小为target序列数据库中的序列数目。

```
21 Domain annotation for each model (and alignments):
22 >> fn3 Fibronectin type III domain
23 # score bias c-Evalue i-Evalue hmmfrom hmm to alifrom ali to envfrom env to acc
24 ---
25 1 ? -1.5 0.0 0.37 0.55 60 72 .. 396 408 .. 395 410 .. 0.86
26 2 ! 40.8 0.0 2.3e-14 3.5e-14 1 83 [. 439 520 .. 439 521 .. 0.95
27 3 ! 14.8 0.0 3.1e-06 4.6e-06 12 84 .. 836 913 .. 826 914 .. 0.74
28 4 ! 5.0 0.0 0.0035 0.0052 9 36 .. 1209 1236 .. 1203 1258 .. 0.83
29 5 ! 22.4 0.0 1.3e-08 2e-08 13 79 .. 1313 1380 .. 1305 1385 .. 0.81
30 6 ? 0.6 0.0 0.079 0.12 55 72 .. 1753 1769 .. 1720 1769 .. 0.87
31 7 ! 47.3 0.9 2.2e-16 3.2e-16 1 84 [. 1800 1890 .. 1800 1891 .. 0.91
32 8 ! 17.9 0.0 3.2e-07 4.8e-07 5 73 .. 1904 1966 .. 1901 1976 .. 0.91
33 9 ! 10.1 0.0 9e-05 0.00014 1 85 [] 1994 2107 .. 1994 2107 .. 0.87
34
```

domain部分为domain表格加上比对输出，和hmmsearch一样



```

35 Alignments for each domain:
36 == domain 1  score: -1.5 bits;  conditional E-value: 0.37
37             EESSSTTEEEEE CS
38             fn3  60 ltlkpgteYevr 72
39             l+ L p+tt+Y++r
40 7LESS_DROME 396 LEALIPYTQYRFR 408
41             67799*****8 PP
42
43 == domain 2  score: 40.8 bits;  conditional E-value: 2.3e-14
44             TSBCEEEEEESSSEEEEEEE-CSSSSSTECEEEEEEEETTSSSTEEEEEEESTCSEEEEESSSTTEEEEEEEEEETTEEEE CS
45             fn3  1 saPsnlsvsevtstsltsWeppkdgggpitgYeveyrekgeewneftvprtttsvtltgkpgteYevrVqavnggggegp 83
46             saP  ++      ++ l v+W p +  +gpi+gY+++++++ + e+ vp+   s+ +++L++gt+Y++ +  +n++gegp
47 7LESS_DROME 439 SAPVIEHLMGLDDSHLAVHWHHPGRFTNGPIEGYRLRLSSSEGNA-TSEQLVPAGRGSYIFSQIQAGTNYTLALSMINKQGE GP 520
48             5777788889999*****9998.*****997 PP
49

```

同hmmsearch结果

## 搜索DNA序列

HMMER原来是用于蛋白序列分析，hmmsearch和hmmscan能够用于判断这对query profile，全部的target序列是否同源。

nhmmer和nhmmscan程序用于在DNA序列中搜索DNAprofile。用于Dfam数据库(dfam.org)，该数据库提供了来自多个重要基因组的多个共有DNA重复单元的比对和profile。使用方法同hmmsearch和hmmscan。

### 1. 使用hmmbuild构建profile

hmmbuild既可以构建蛋白profile，也可以构建DNA profile

```
hmmbuild MADE1.hmm MADE2.sto
```

```

$hmmbuild MADE1.hmm MADE1.sto
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.2.1 (June 2018); http://hmmerr.org/
# Copyright (C) 2018 Howard Hughes Medical Institute.

```

需要注意的是输出行具有头"W"，只有在DNA/RNA比对后才会出现。该值代表长度的上限，nhmmer期待能够发现一个家族(this represents an upper bound on the length at which nhmmer expects to find an instance of the family)。该值常常大于mlen，尽管mlen和W的比值依赖seed比对中观察到的插入率。越大的W值，运行时间越长。

### 2. 使用nhmmer搜索DNA序列数据

nhmmer能够接受profile文件由hmmbuild构建或包含单个DNA序列或多重DNA比对的DNA序列，针对多重DNA比对文件，nhmmer先针对比对文件进行profile构建，默认保存后缀.hmm，然后再在profile中搜索target DNA序列

```
nhmmer MADE2.sto dna_target.fa
```

```
nhmmer MADE1.hmm dna_target.fa
```

输出和hmmsearch大部分相同。关键差异在于，每个hit不是target数据库中全序列，而是profile和target数据库子序列的局部比对。

```
10
11 Query:      MADE1 [M=80]
```

nhmmscan相对于hmmscan就如同nhmmer相对于hmmsearch。

---