Introduction

大多是蛋白序列都是由相对小数目的始祖蛋白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(Pfram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hmmscan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 Forwad/Backwaid 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	www.soeding.genzentrum.lmu.de/software-and-servers-2

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。

未比对序列格式包括：fasta, uniport, 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：

```
$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
```

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墒值(re/pos; average score)，0.598bits。

新生成的profile文件保存为globins4.hmm：

```
HMMER3/f [3.2.1 | June 2018]
NAME  globins4
LENG  149
ALPH  amino
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Fri May 31 15:40:02 2019
NSEQ  4
EFFN  0.964844
CKSUM 2027839109
STATS LOCAL MSV       -9.9014  0.70957
STATS LOCAL VITERBI  -10.7224  0.70957
STATS LOCAL FORWARD   -4.1637  0.70957
HMM          A        C        D        E        F        G        H        I        K
             m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.36553  4.52577  2.96709  2.70473  3.20818  3.02239  3.41069  2.90041  2.55332  2
          2.68640  4.42247  2.77497  2.73145  3.46376  2.40504  3.72516  3.29302  2.67763  2
          0.57544  1.78073  1.31293  1.75577  0.18968  0.00000        *
      1   1.70038  4.17733  3.76164  3.36686  3.72281  3.29583  4.27570  2.40482  3.29230  2
```

2. 使用hmmsearch搜索序列数据库

hmmsearch接受fasta文件作为target 数据库输入，也可接受EMBL/UniProtKB文本格式，Genebank格式，同时会自动检测输入格式

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

```
hmmsearch fn3.hmm 7LESS_DROME > fn3.out
```

查看输出fn3.out

第一部分包含了使用程序情况及选项

```
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence    Description
    ------- ------ -----    ------- ------ -----   ---- --  --------    -----------
    6.2e-56  173.1   1.6    1.1e-16   47.3   0.9    9.5  9  7LESS_DROME  RecName: Full=Prote
```

第二部分包含了经过排列后的最佳hits信息(根据E-value，最显著性排序)

第一列的E-value是最重要信息，表明期待的假阳性率(non-homologous sequences)，一般选择小于le-3的hits；第二列为sequence bit score(log-odds score for the complete sequence)，该值根据profile和target sequence得来；第三个bias，为用于序列bit score的偏差性序列组成的修正项，例如这里score为173.1，bias为1.6，意味着原始score为174.7，减去1.6的序列组成偏向性修正值，调整后得173.1，仅bias和score值处于相同数量级时，需要格外注意；接下来的三个也是E-value，score和bias，是针对序列中的最佳比对domain的值，而不是所有识别domain的值之和(前三个)。图中可见，所有最佳之和为173.1，而最佳值为47.3，这可能意味值该序列为多重domain家族成员，因为它包含了

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):
1 >> 7LESS_DROME  RecName: Full=Protein sevenless;          EC=2.7.10.1;
2    #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
3  ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
4    1 ?   -1.5   0.0      0.18      0.18      60      72 ..     396     408 ..     395     410 .. 0.86
5    2 !   40.8   0.0   1.2e-14   1.2e-14       1      83 [.     439     520 ..     439     521 .. 0.95
6    3 !   14.8   0.0   1.5e-06   1.5e-06      12      84 ..     836     913 ..     826     914 .. 0.74
7    4 !    5.0   0.0    0.0017    0.0017       9      36 ..    1209    1236 ..    1203    1258 .. 0.83
8    5 !   22.4   0.0   6.7e-09   6.7e-09      13      79 ..    1313    1380 ..    1305    1385 .. 0.81
9    6 ?    0.6   0.0      0.04      0.04      55      72 ..    1753    1769 ..    1720    1769 .. 0.87
0    7 !   47.3   0.9   1.1e-16   1.1e-16       1      84 [.    1800    1890 ..    1800    1891 .. 0.91
     8 !   17.9   0.0   1.6e-07   1.6e-07       5      73 ..    1904    1966 ..    1901    1976 .. 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envelpe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 ltgLkpgteYevr 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                   TSBCEEEEEEEESSSEEEEEEE-CSSSSSTECEEEEEEEETTSSSTEEEEEEESTCSEEEEESSSTTEEEEEEEEEEETTEEEE CS
44           fn3   1 saPsnlsvsevtstsltvsWeppkdgggpitgYeveyrekgseeewneftvprtttsvtltgLkpgteYevrVqavngggegp 83
45                   saP   ++    ++ l v+W p +  +gpi+gY++++++++++  + e+ vp+   s+ +++L++gt+Y++ +  +n++gegp
46   7LESS_DROME 439 SAPVIEHLMGLDDSHLAVHWHPGRFTNGPIEGYRLRLSSSEGNA-TSEQLVPAGRGSYIFSQLQAGTNYTLALSMINKQGEGP 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_NUMAN 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profie数据库(searching a profile database with a query sequence, protein)

相对于在单个proifle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hmmpress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

```
hmmpress minifam
```

同时产生对应二进制文件

```
$hmmpress 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

```
Alignments for each domain:
== domain 1  score: -1.5 bits;  conditional E-value: 0.37
                 EESSSTTEEEEEE CS
         fn3  60 ltgLkpgteYevr 72
                 l+ L p+t+Y++r
7LESS_DROME 396 LEALIPYTQYRFR 408
                 67799*******8 PP

== domain 2  score: 40.8 bits;  conditional E-value: 2.3e-14
                 TSBCEEEEEEESSSEEEEEEE-CSSSSSTECEEEEEEEETTSSSTEEEEEEEESTCSEEEEESSSTTEEEEEEEEEEETTEEEE CS
         fn3   1 saPsnlsvsevtstsltvsWeppkdgggpitgYeveyrekgseeewneftvprtttsvtltgLkpgteYevrVqavngggegp 83
                 saP   ++    ++ l v+W p +  +gpi+gY++++++++++  + e+ vp+   s+ +++L++gt+Y++ +  +n++gegp
7LESS_DROME 439 SAPVIEHLMGLDDSHLAVHWHPGRFTNGPIEGYRLRLSSSEGNA-TSEQLVPAGRGSYIFSQLQAGTNYTLALSMINKQGEGP 520
                 5777788889999**************************9998.**********************************997 PP
```

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://hmmer.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比对时后才会出现。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 excepts 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，然后再再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porfile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。