## Instruction

蛋白一般拥有一个或者多个功能区域，常被称为"domains"。不同蛋白中不同的domains的组合会产生多样的功能特征。识别蛋白中存在的domain能给认识蛋白功能提供依据。 Pfam就是一个这些保守进化单元的数据库。

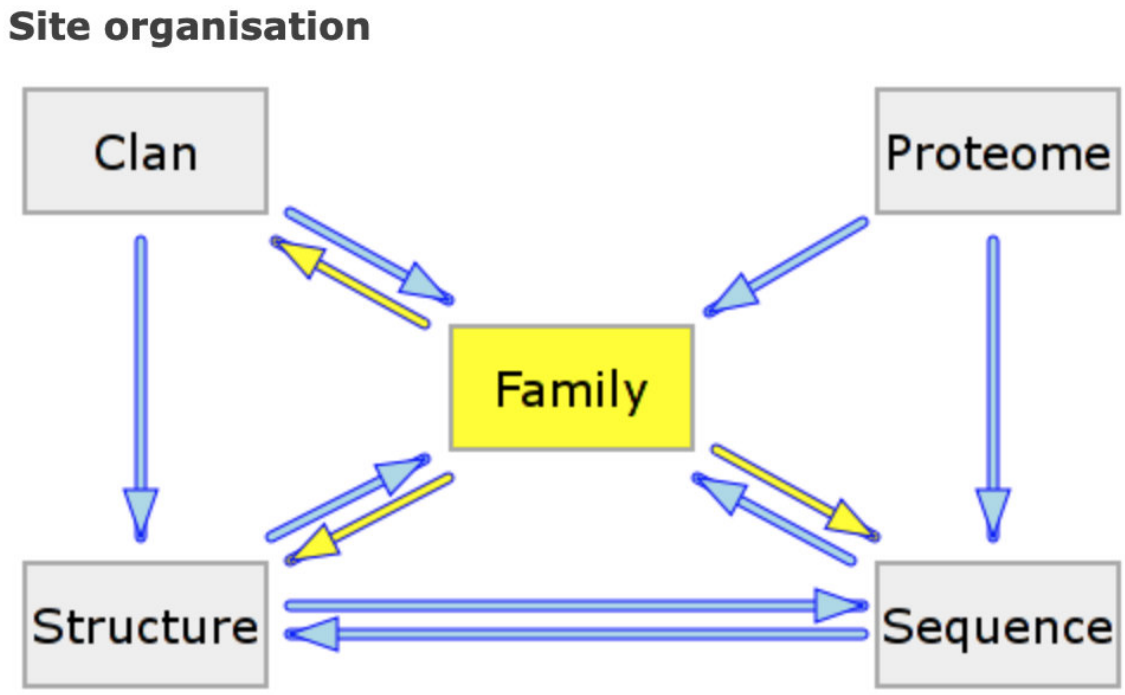
每一个Pfam收录的信息都代表了一套比对的序列，也就是一个profile。profile HMM(hidden Markov model)根据属于一个家族的一套少量的代表性的比对序列 ('seed'比对) 训练而来，训练得到的模型被用于在更大的数据库 (例如, UniProKB) 中彻底搜索发现所有同源序列。这些和模型显著性相似的序列比对到profile HMM是为了提供完整的比对信息。

相关的Pfam收录信息可以聚集成一个集，标记为"Clans"。这些"Clans"通常属于很大并且多样的超家族，这样单个profile HMM就可以获得所有序列成员。

在Pfam中比对序列可以帮助将功能信息从一条经实验验证过的序列推导到其他属于相同收录的没有经过实验验证的序列上，从而针对每一条收录提供综合性的注释。

利用Pfam，可以在models中搜索蛋白或核酸序列；搜索已有的家族和clans；针对任何指定的家族或收录回溯文本注释信息；查看一个家族或clans的多重序列比对；查看一个clan中家族之间的关系；查看家族内容下的蛋白结构信息；根据分类学延伸查看家族；使用关键词搜索数据库。

Pfam收录通过以下六种中的一种进行分类：family，相关蛋白区域集合；domain，结构单元；repeat，短的结构单元单独存在时不稳定，当多个拷贝存在时能形成稳定结构；motifs，球形结构域外的更小的单元；coiled-coil，主要包含coiled-coil motifs的区域，通常包含alpha-helices；disordered，保守区域，展示或预测包含有偏差的序列组成，或内部是无序(non-globular)。相关的Pfam收录聚集成clans，这之间的关系可能定义为序列，结构或porfile HMM相似性。




其他数据库

PROSITE
: This originally was based around regular expression patterns but now also includes profiles.

PRINTS
: This is based around protein "finger-prints" of a series of small conserved motifs making up a domain.

SMART
: This is a database concentrating on extracellular modules and signaling domains.

ADDA
: This is an automatic algorithm for domain decomposition and clustering of protein domain families.

InterPro
: Combines information from Pfam, Prints, SMART, Prosite and PRODOM.

CDD
: The Conserved Domain Database is derived from Pfam and SMART databases.

## Usage

1. 下载对应的数据库

`ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/`

```
-rwxr-xr-x  1 carlos  wheel   3.9G Jun  2 10:16 Pfam-A.regions.tsv
-rwxr-xr-x  1 carlos  wheel   666M Jun  1 17:05 Pfam-A.hmm.h3p
-rwxr-xr-x  1 carlos  wheel   313M Jun  1 17:05 Pfam-A.hmm.h3f
-rwxr-xr-x  1 carlos  wheel   566M Jun  1 17:05 Pfam-A.hmm.h3m
-rwxr-xr-x  1 carlos  wheel   1.2M Jun  1 17:05 Pfam-A.hmm.h3i
-rwxr-xr-x  1 carlos  staff   1.6K Jun  1 16:35 md5_checksums
-rwxr-xr-x@ 1 carlos  wheel    22K Oct  3  2018 relnotes.txt
-rwxr-xr-x  1 carlos  wheel   2.8M Sep  4  2018 Pfam-A.hmm.dat
-rwxr-xr-x@ 1 carlos  wheel    19K Sep  4  2018 userman.txt
-rwxr-xr-x  1 carlos  wheel   1.0M Aug 30  2018 Pfam-A.clans.tsv
-rwxr-xr-x  1 carlos  wheel   1.3G Aug 30  2018 Pfam-A.hmm
-rwxr-xr-x@ 1 carlos  wheel   155M Aug 30  2018 Pfam-A.seed.gz
-rwxr-xr-x  1 carlos  wheel    40K Aug 29  2018 active_site.dat
```

2. 使用hmmperspress处理下载数据，为hmmscan搜索准备HMM数据库

`hmmpress Pfam-A.hm`

3. 使用pfam_scan.pl在Pfam HMMs数据中搜索fasta输入文件

`pram_scan.pl -fasta <fasta_file> -dir <directory location of Pfam files>`

-align：显示HMM-sequence比对情况

-e_seq： 指定hmmscan搜索evalue阈值

-b_seq： 指定hmmscan搜索score阈值

-e_dom： 指定hmmscan搜索domain阈值

-cpu： 指定CPUs数目

-translate [mode]：输入为DNA, 搜索前先转录(six-frame)，mode指定为all或orf，表示搜索所有ORFs或仅搜索ORFs长度大于20的序列

-outfile： 输出文件，默认STDOUT

`pfam_scan.pl -fasta globins45.fa -dir Pfam_database -outfile test`

`pfam_scan.pl -fasta globins45.fa -dir Pfam_database -outfile test_align –align`