阅读日期：20190427

发表日期：20181106

发表杂志：

影响因子：

平均影响因子：

标签：

|  |  |
| --- | --- |
| 文章名 InterPro in 2019: improving coverage, classification and access to protein sequence annotations  摘要 ABSTRACT | |
| 中文译名 | 2019年InterPro:提高蛋白质序列注释的覆盖率、分类和访问 |
| 网址 | (<https://www.ebi.ac.uk/interpro/beta/>) |
| 关键词 |  |
| 摘要 |  |
| 启发 |  |
| 摘录 |  |
| 相关论文 | 32. Velankar,S., Dana,J.M., Jacobsen,J., van Ginkel,G., Gane,P.J., Luo,J., Oldfeld,T.J., O’Donovan,C., Martin,M.-J. and Kleywegt,G.J. (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.*, **41**, D483–D489 |

题目：

The InterPro database (http://www.ebi.ac.uk/ interpro/) classifies protein sequences into families and predicts the presence of functionally important domains and sites. Here, we report recent developments with InterPro (version 70.0) and its associated software, including an 18% growth in the size of the database in terms on new InterPro entries, updates to content, the inclusion of an additional entry type, refined modelling of discontinuous domains, and the development of a new programmatic interface and website. These developments extend and enrich the information provided by InterPro, and provide greater flexibility in terms of data access. We also show that InterPro’s sequence coverage has kept pace with the growth of UniProtKB, and discuss how our evaluation of residue coverage may help guide future curation activities.

记录

INTRODUCTION

1,现在数据多，太多的标识不完。

2，现在最大的蛋白质标注是InterPro。由14个成员数据库得来。

UniProt Knowledgebase (UniProtKB) (1)

CATH-Gene3D (2), the Conserved Domains Database (CDD) (3), HAMAP (4), PANTHER (5),  
Pfam (6), PIRSF (7), PRINTS (8), ProDom (9), PROSITE  
Patterns (10), PROSITE Profles (10), SMART (11), the  
Structure–Function Linkage Database (SFLD) (12), SUPERFAMILY (13) and TIGRFAMs (14)

通过这些数据库计算得到蛋白质的位置特性。

3这些小的数据库都有各的专业性，有的是一片区域，有的是颗粒性的数据。只有一个子集有这样的特性，可以精确到配合基的特性，那就是催化剂类型的蛋白。他们是CATH-Gene3D, HAMAP, Pfam, PIRSF, PROSITE, CDD and SFLD,

4还提供一些有关序列的特性。

5InterPro和它相关的软件被广泛应用。数据库每两个月更新一次

RESULT

数据库更新了很多。不同的版本针对不同数据库的变更进行了更新。

新的网站

(<https://www.ebi.ac.uk/interpro/beta/>)

API

1主要 有六个API端：Entries, Proteins, Structures, Sets, Proteomes and Taxonomies.

2蛋白质数据来自UniPortKB

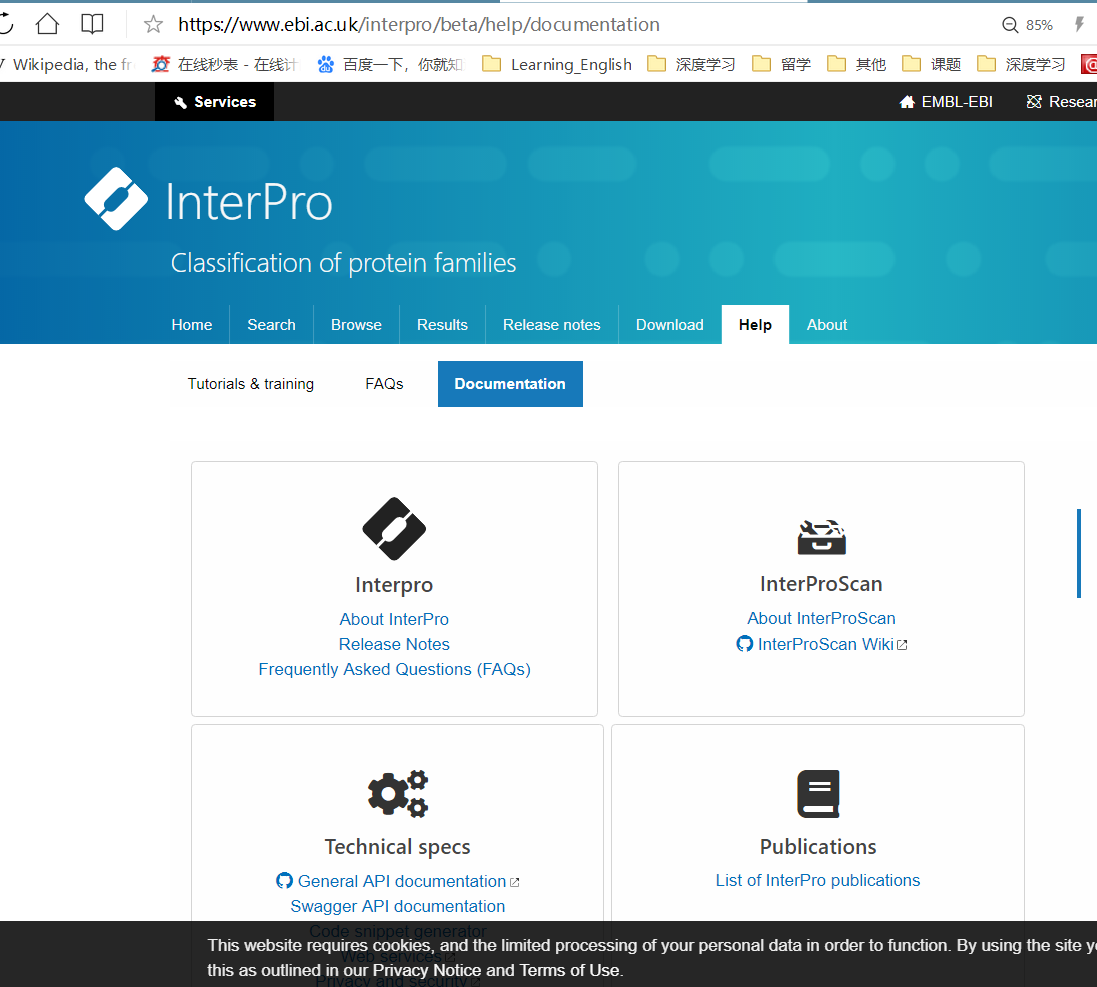
3Structure来自由PDBe提供的SIFTS mapping

4提供一个新的条目：同源超家族

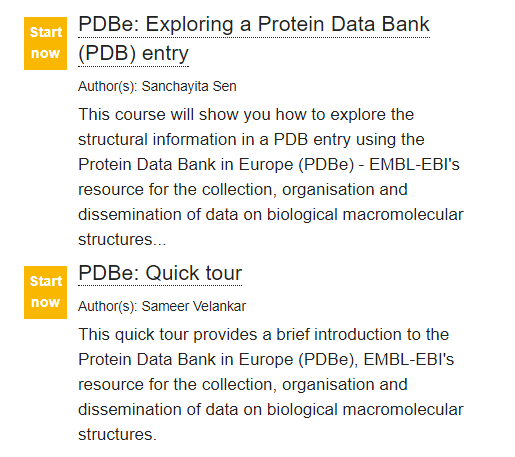
5不连续的区域。它们在3D结构中是一篇区域。它们是从C端开始不连续，还是从N端开始不连续，将会在输出文件中给出。（没懂）

Online documentation for the API is available at

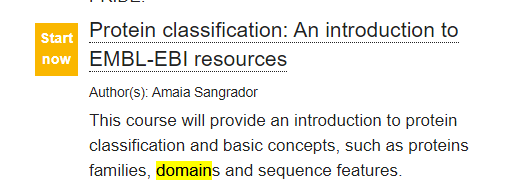
<https://www.ebi.ac.uk/interpro/beta/help/documentation/>



在其中的教学课程中，找到了如何使用PDBe的相关教学。

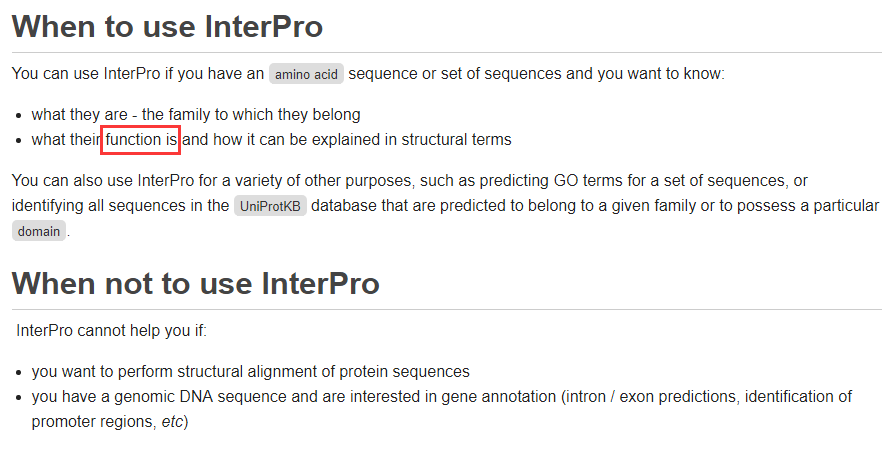


还有这个：有关域的介绍。



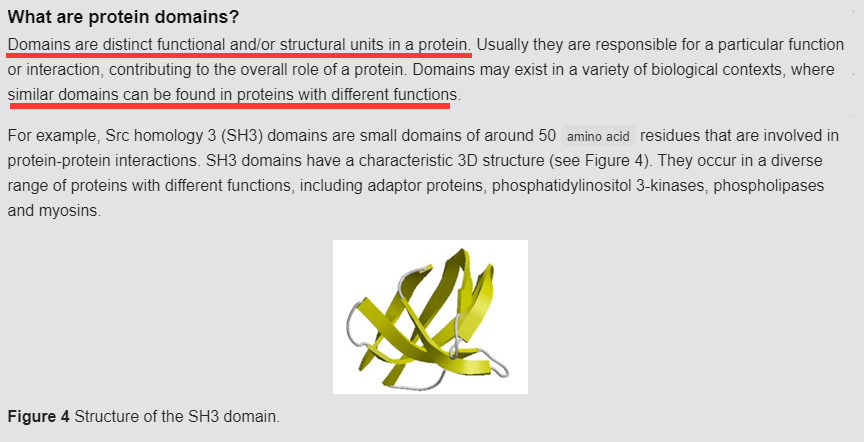
学习记录：

1，InterPro的主要功能是蛋白质家族分类。其中有一个功能是去除冗余数据。

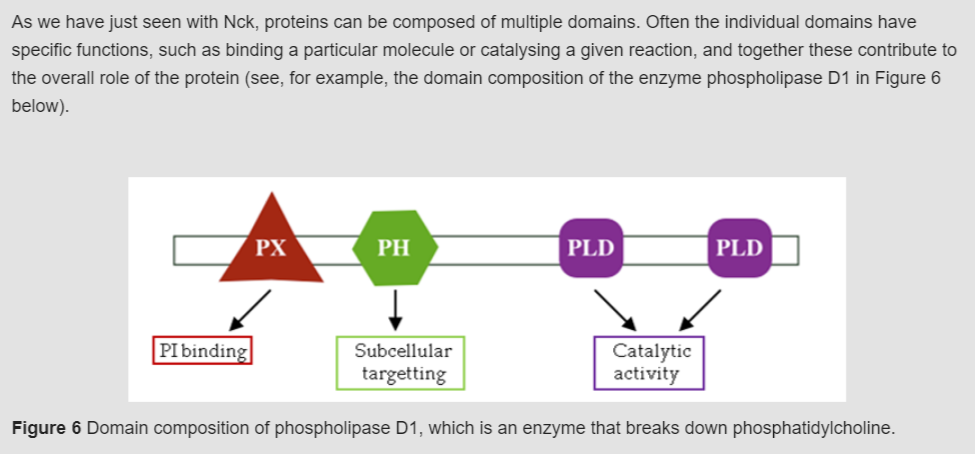
2，

3，At its heart在其核心。

4，什么是蛋白质域？

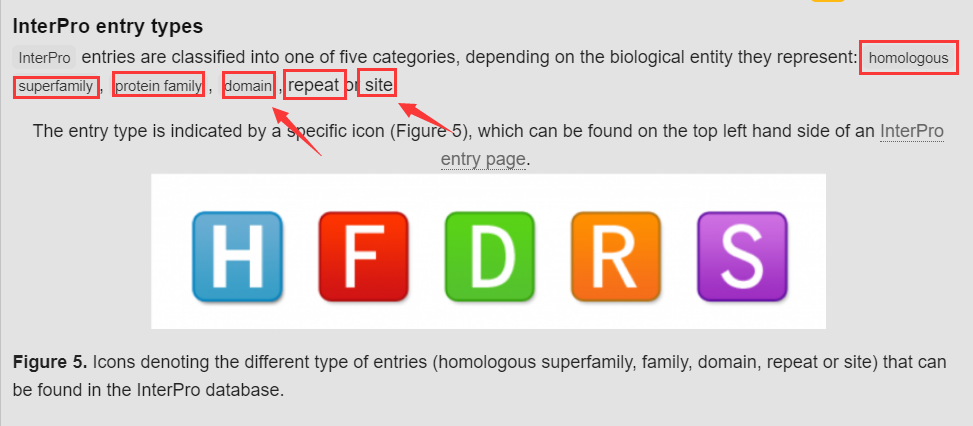


同一蛋白的一个域可以有多种功能。

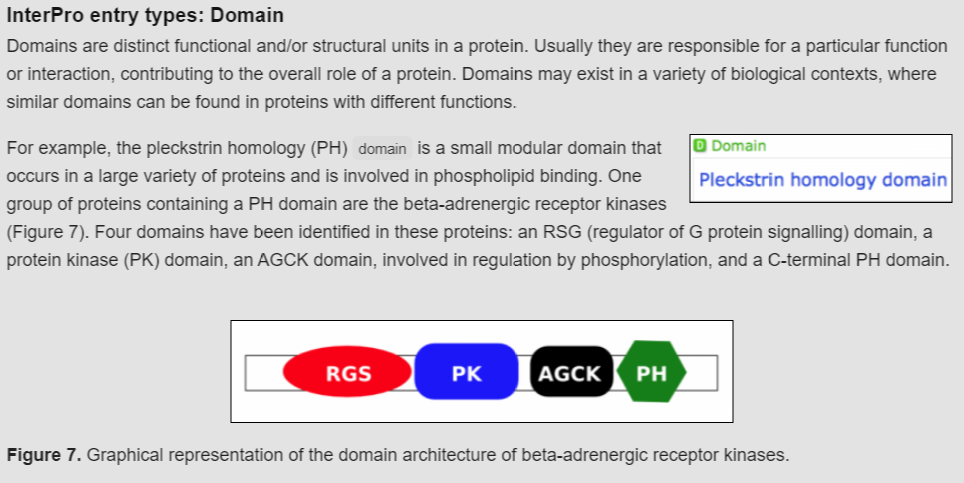


不同的域有不同的功能。

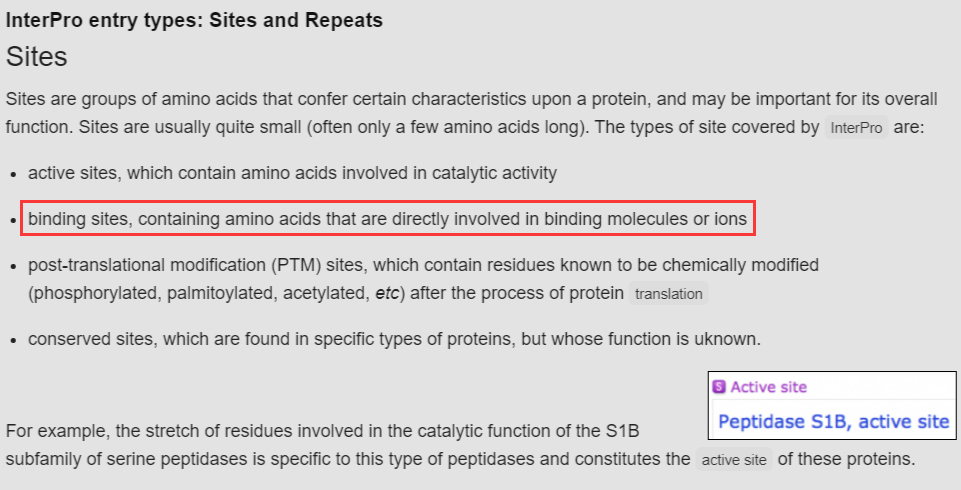
5InterPro entry types。这个数据库里有这样六种条目的数据。



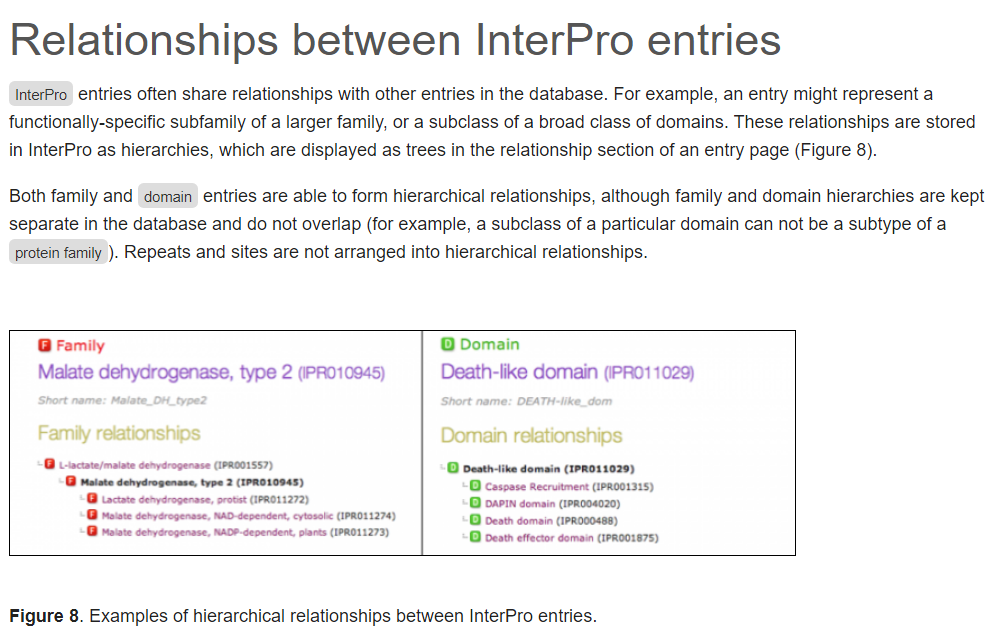
（其中有关蛋白质家族的介绍跳过。）



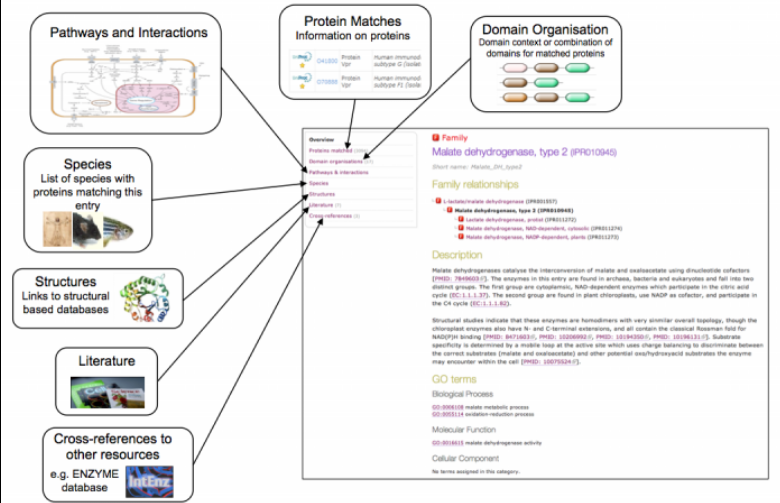
有关蛋白质域的介绍。图示如上。



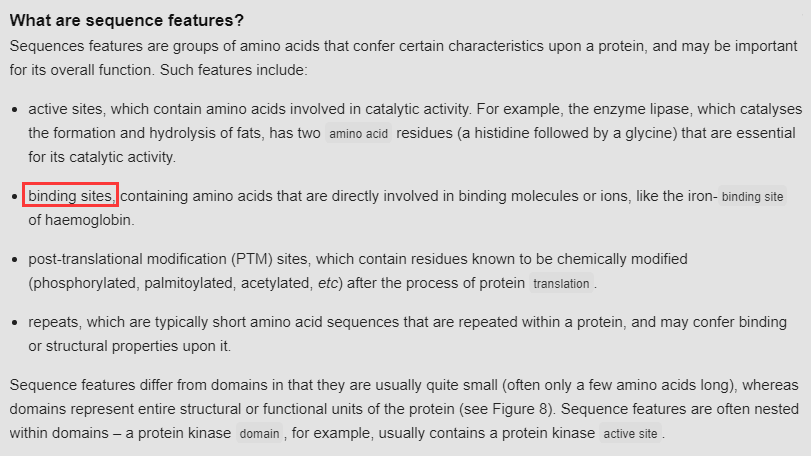
我们是否可以直接从这里找到作用位点？或做验证？或做训练集？



各种条目是分等级，且相互之间有联系的。

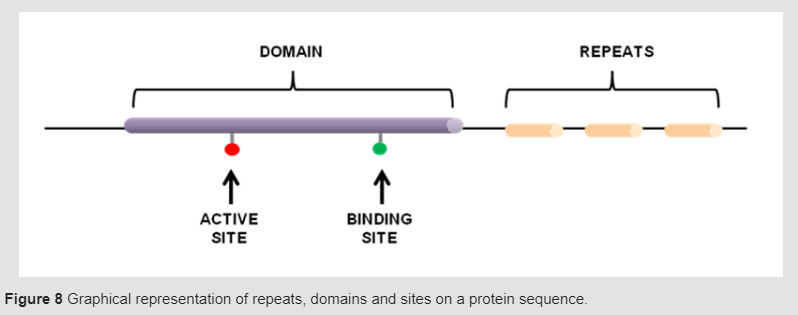


边栏中也有许多有用的链接。

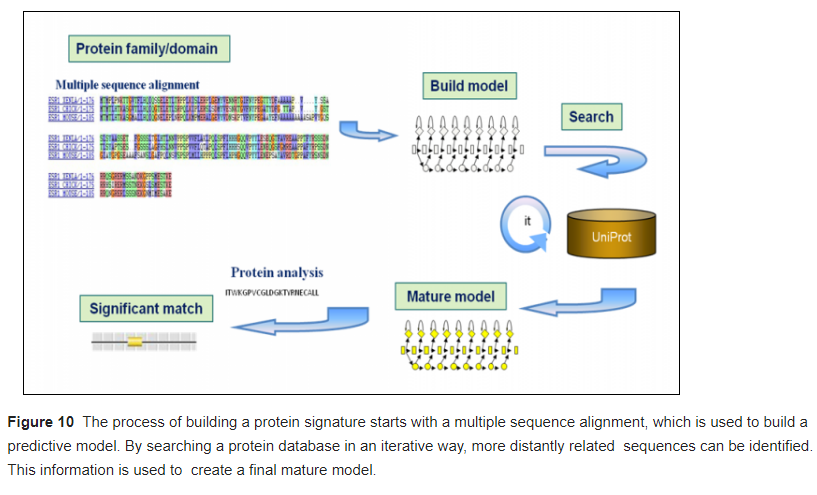


蛋白质绑定位点信息是直接给出的。。。。（那为啥我去测试出来的蛋白质序列没有预测出给定位点？）

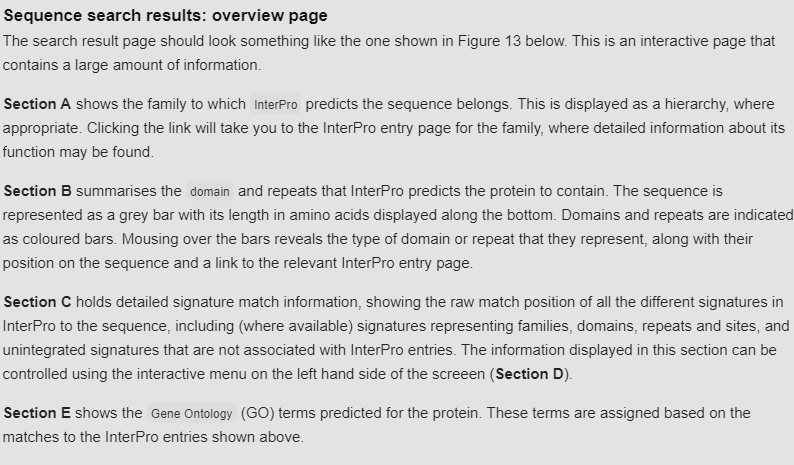
预测出的“域”中也并不能和位点有很大联系。我放到excle里比较了一下。但只比较了1条。

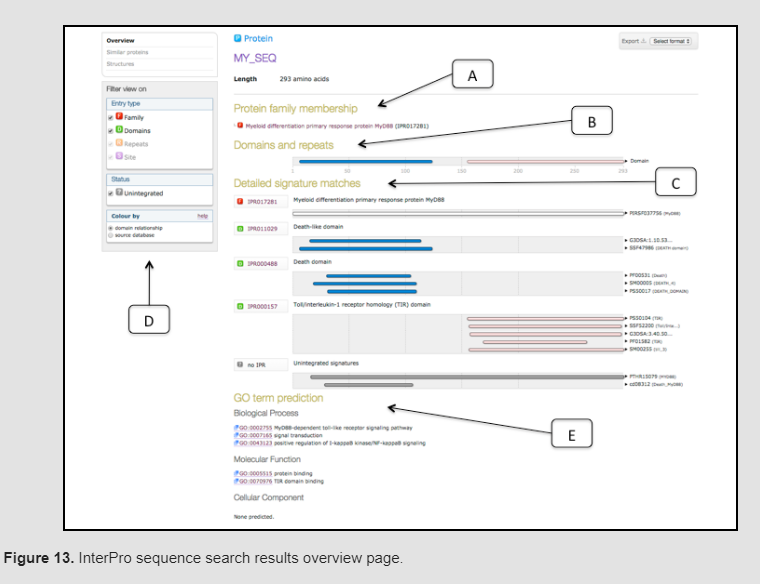


绿色点点是绑定位点。



上图说明了数据库标识蛋白质特征的流程。





测试一个序列及其相关的信息。

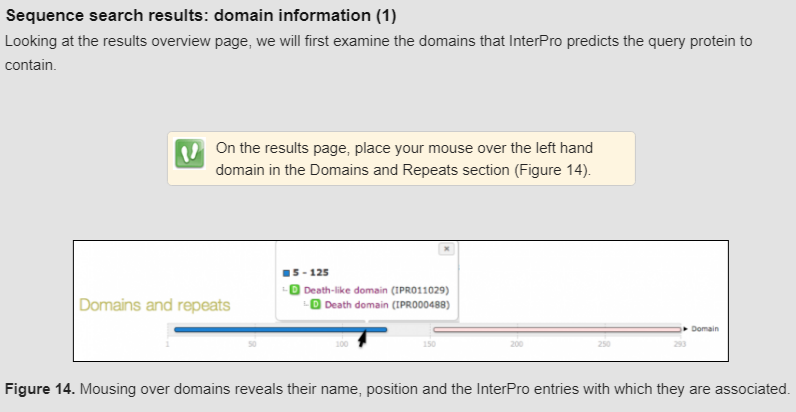
随机选取我数据条目中的一条蛋白，结合教程、InterPro\PDB\Uniprot进行查看。

【选取的序列】

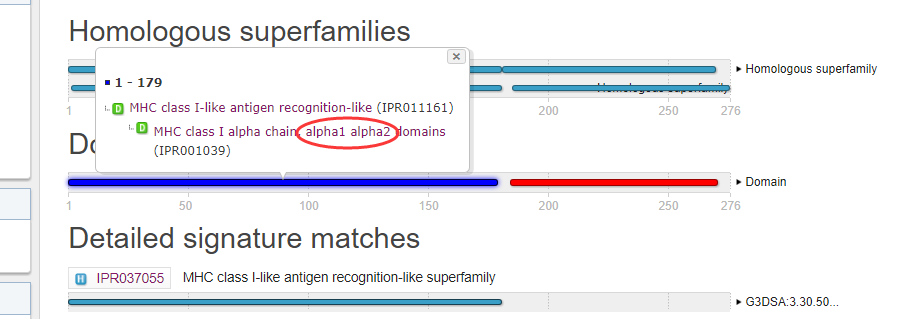
>4MNQ\_A

GSHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAASQRMEPRAPWIEQEGPEYWDGETRKVKAHSQTHRVDLGTLRGYYNQSEAGSHTVQRMYGCDVGSDWRFLRGYHQYAYDGKDYIALKEDLRSWTAADMAAQTTKHKWEAAHVAEQLRAYLEGTCVEWLRRYLENGKETLQRTDAPKTHMTHHAVSDHEATLRCWALSFYPAEITLTWQRDGEDQTQDTELVETRPAGDGTFQKWAAVVVPSGQEQRYTCHVQHEGLPKPLTLRWEP

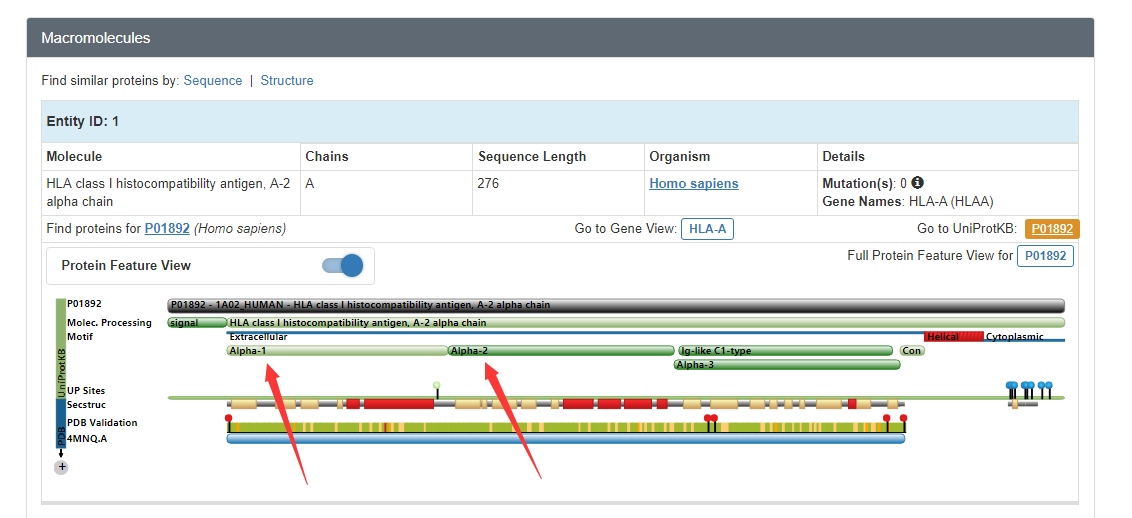
000000000000000000000000000000000010000111000100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000



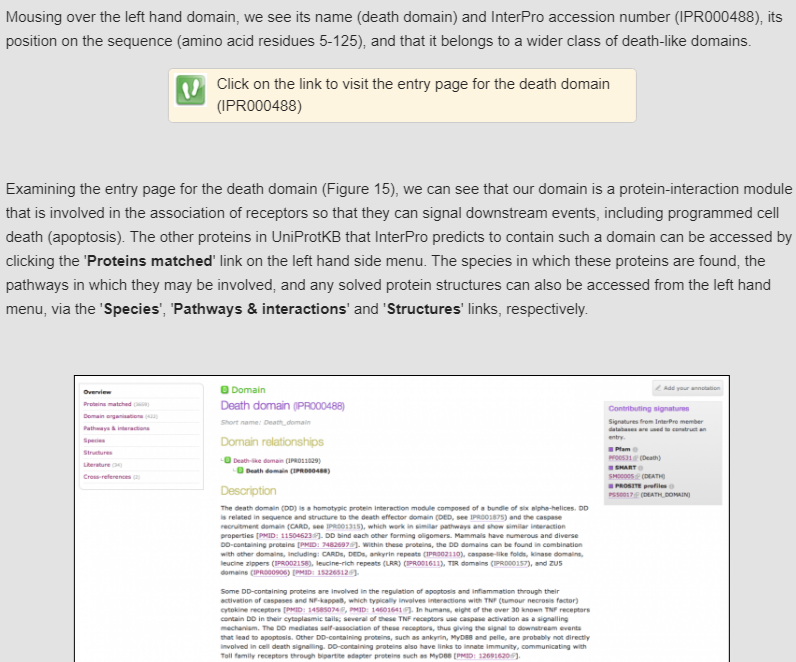
（当然我没有使用示例蛋白，所以我只能看看我的蛋白质是什么样的。）



Intro中的。

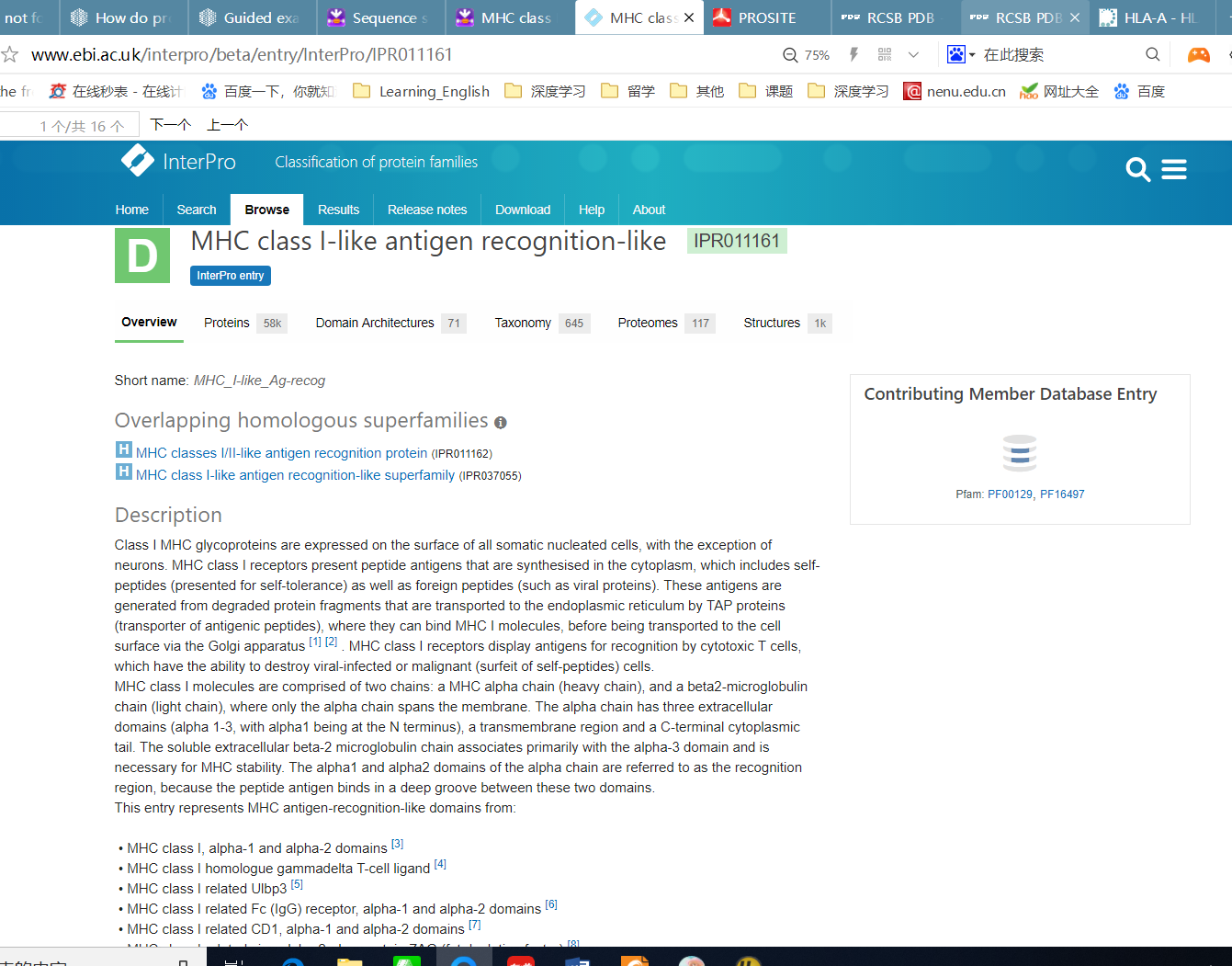


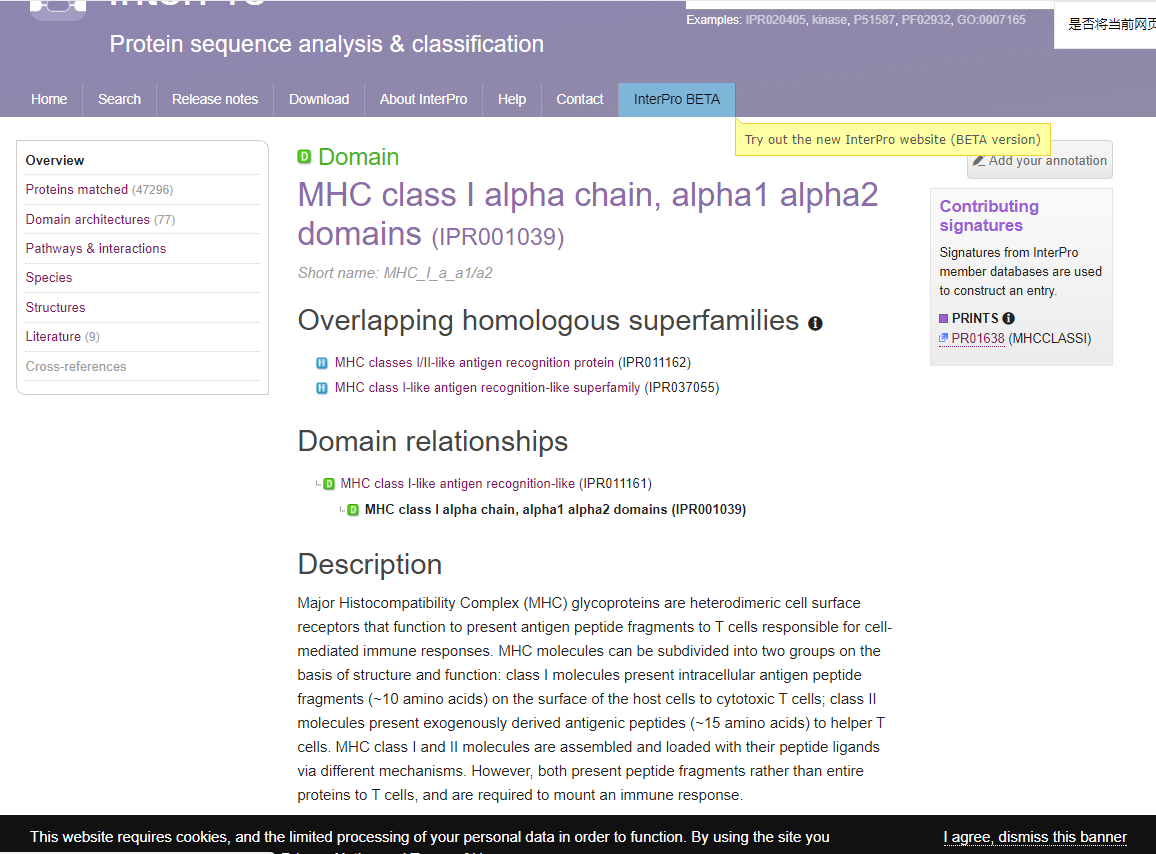
PDB中的。



接下来我们要查看一下这个域是做什么的

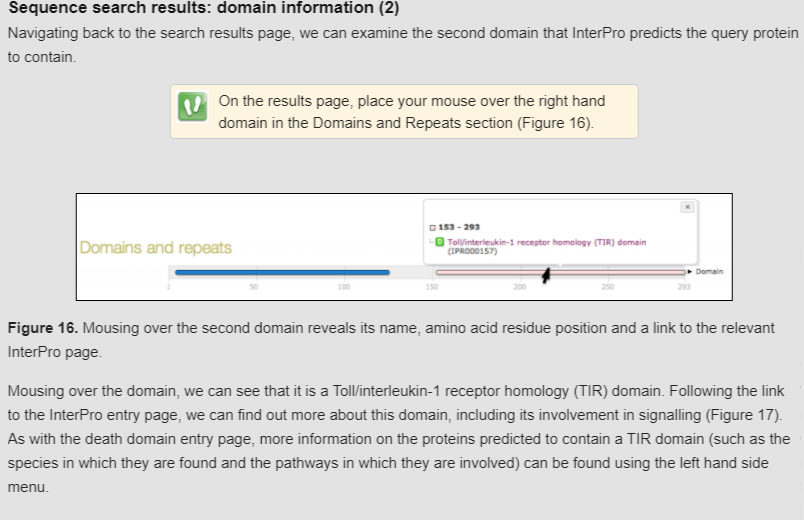
在我测试的条目中，有两个条目，我分别在新的页面中打开。（有新版网站和旧版网站的区分）



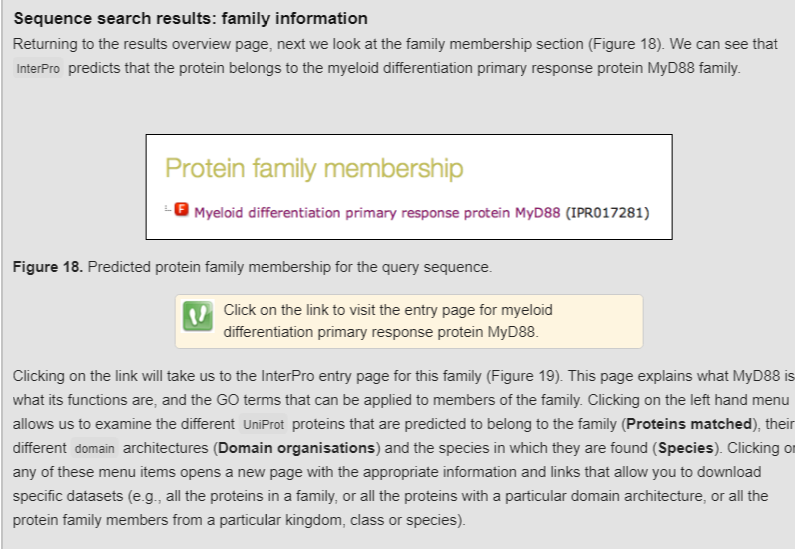


点击BETA按钮后变成蓝色，蓝色版本应该是新版。

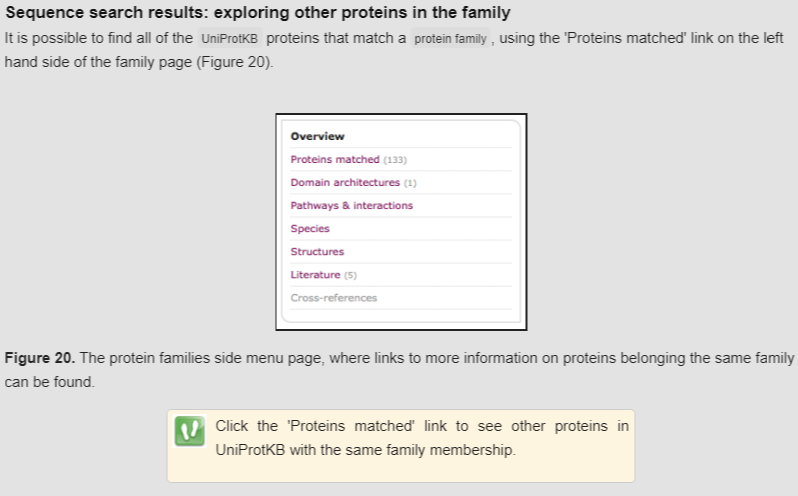
总之就是点击链接可以看到这个“域”的具体信息。



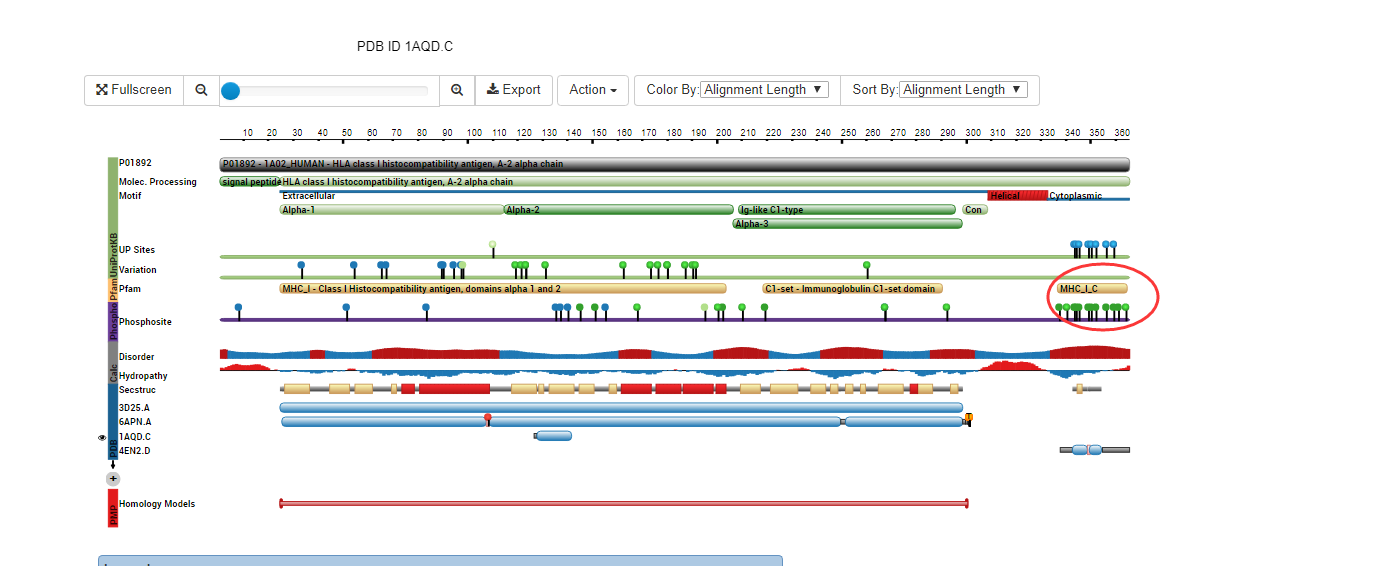
回到教程中，我们可以看到使用同样方法看有关域的信息。



我们还可以找其他蛋白质家族的信息。

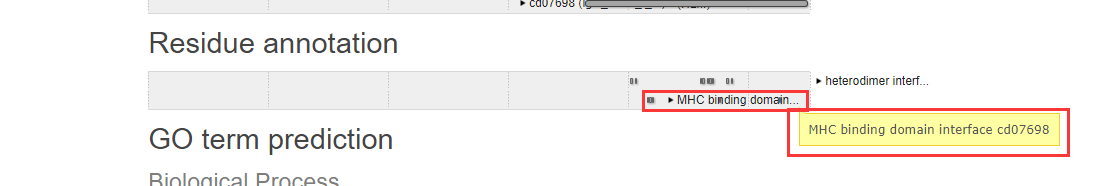


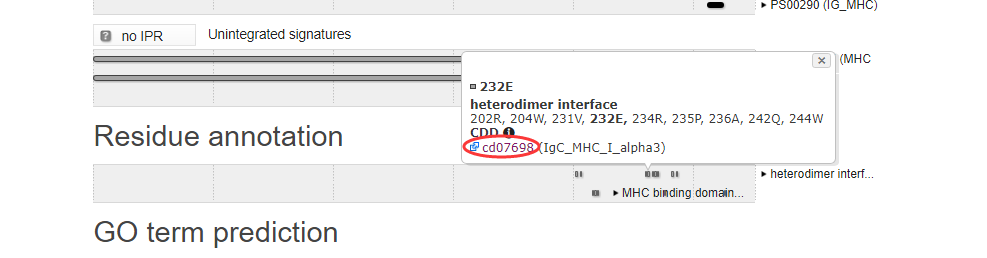
还可以找其家族中的其他蛋白质。



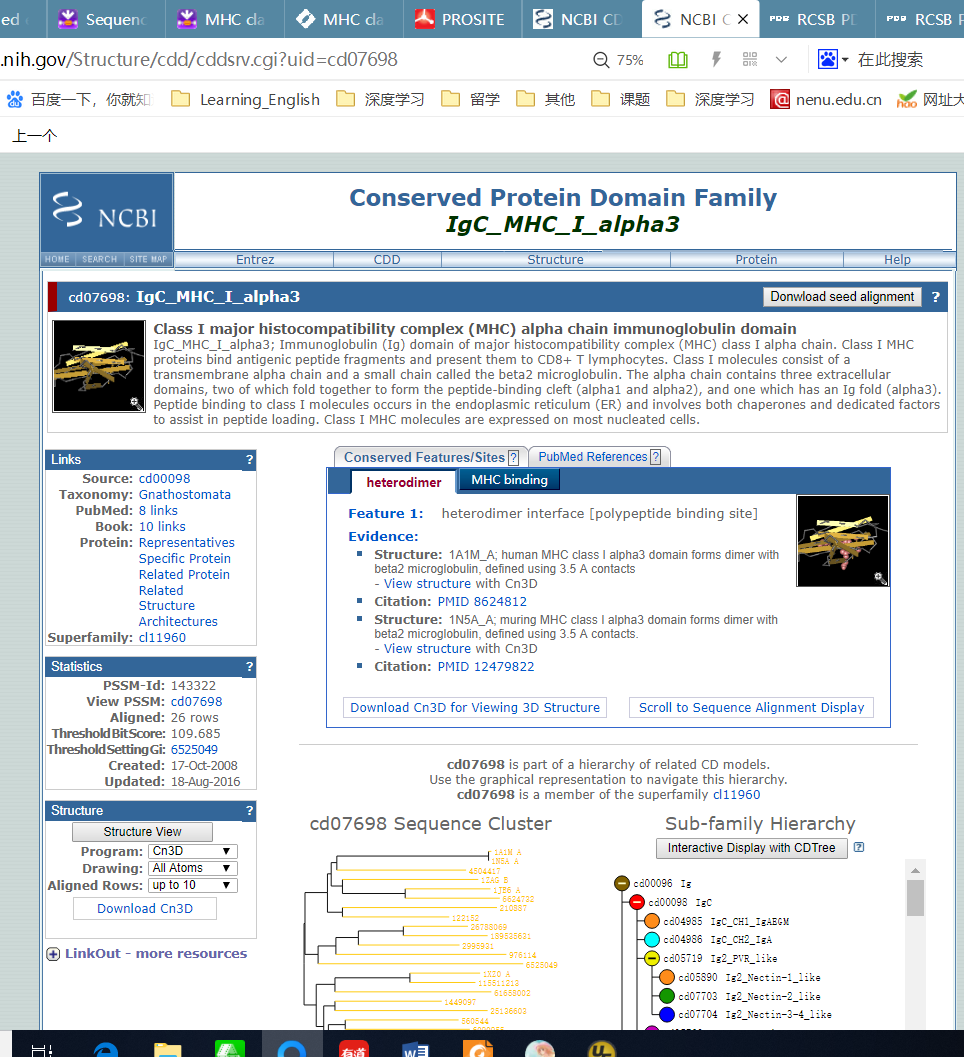
查看PDB的大图看到右侧有结合位点信息。

在InterPro中也发现了结合位点信息！！！！





我点击了cd07698这个链接，得到：



MHC这种东西的结合位点！！！！但是这个MHC看起来不是我们提到的drug中的三位大写英文字母的编号。同时，这个蛋白质条目没有家族分类。（说明可能我们提供的蛋白质不是完整的蛋白质条目）

总结：

1，该数据库除了蛋白质序列信息中域的标注，还有蛋白质家族分类的功能。以及找同家族蛋白质的功能。

2，我们可以通过该数据库查找到某条序列中的各种“域”但域的种类？域的功能？需要我们学习了解后才可以应用。

3，目前测试的两条蛋白质序列，都没有准确找到对应的域：一条蛋白质中域的总类是很多的，同一域的功能也是多种多样的。我们的位点是否在“域”中，是个需要仔细考量的任务。

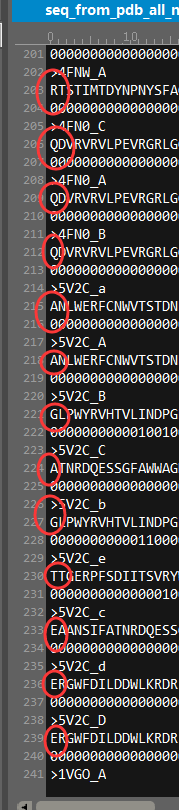
4，该数据库还可以找到某些结合位点。但结合位点的信息好像不是我们提到过的3位英文的药物名字。

5，使用该软件需要9G以上内存的Linux系统。（这部分我没有看呢。）

6，第二条目中我们的位点在序列的前侧，预测出来的位点在蛋白的后侧。

7，若没有相同蛋白，则进行预测。如果有相同蛋白，则不是预测，而是根据数据库中的数据进行显示。

其他：发现我们的数据中，蛋白质开头字母比较随机：



而上次讨论班中有同学提到，他所遇到的蛋白70%开头氨基酸都是”M”开头。（同时宫迎莉也表示遇到了这个情况）。我们当时想出来的解释是这样的：起始密码子的种类有限，所以开头的氨基酸种类也应该是固定的，或者就那么几种。出现M居多应该是可理解的。但是我的数据中M开头的较少，这是什么原因呢？