










## CATH 数据库介绍

CATH 是蛋白质结构分类数据库，其内的蛋白质结构均为从 PDB 数据库中下载而来。CATH 将下载而来的蛋白质按照结构进行分类，其分类方式为自动程序和人工处理混合。CATH 数据库由一个名为 Orengo Group 的小组进行维护，该小组由多名学者组成，目前已经经历过多个版本的变化，当前最新的版本为 CATH v4.0，其内包含 234848 个 CATH 结构域，2738 个 CATH 超家族，69058 个标注过的 PDB 结构。

CATH 数据库中，蛋白质按照结构分为多个层级，分别为：class, architecture, topology, homologous superfamily, sequence family, orthologous family, alike domain, identical domain, domain counter.具体细分标准如下图所示

Depth	Letter	Name	Clustering criteria
1		Class	Secondary structure content
2		Architecture	General spatial arrangement of secondary structures
3		Topology	Spatial arrangement and connectivity of secondary structures (fold)
4		Homologous Superfamily	Manual curation of evidence of evolutionary relationship (at least two criteria from sequence/structure/function must be observed)
5		Sequence Family (S35)	$\geq 35\%$ sequence similarity
6		Orthologous Family (S60) *	$\geq 60\%$ sequence similarity
7		âLikeâ domain (S95) *	$\geq 95\%$ sequence similarity
8		Identical domain (S100)	100% sequence similarity
9		Domain counter	Unique domains

CATH 作为数据库可为用户提供非常丰富的功能，现一一列举如下：

### 1、查询未知蛋白的功能

用户可以利用 CATH 提供的服务器 (CATHEDRAL) 将提交的未知蛋白与已有蛋白进行结构上的对比并分类。蛋白对比的结果会进行打分并排序, 而打分的依据则相当充足, 如标准化后的 RMSD, 重合度的百分比, 序列一致性等。而在结果出来以后, 用户还可以观察所提交蛋白质与已有蛋白质的三维结构的差异, 进一步确定蛋白质功能。

## 2、查看结构变化程度

CATH 能够让用户查询某一个超家族(superfamily)中的蛋白质结构的变化程度。以 HUP 超家族为例进行演示。用户可通过 CATH 编码搜索以及关键字搜索搜索到 hup 超家族的相关信息。

## 3、与 Gene3D 联合

CATH 编码在 Gene3D 数据库中也可以使用, 即用户可以在 Gene3D 数据库中使用 CATH 编码进行搜索, 从而查询到该超家族在 Gene3D 中的特有信息。

## 4、进行药物设计

CATH 数据库内置一个结构比较算法 SSAP, SSAP 能够对两个蛋白质逐个残基进行比较, 给出结构相似度的评分, 分数范围在 1 到 100。分数越高则相似度越高。用户可以将同一超家族的两个蛋白提交给 SSAP 进行比较, 从而能够全面的观测到两个蛋白质在结构方面的差异。

CATH 数据库的主页截图如下

**CATH / Gene3D**  
26 million protein domains classified into 2,738 superfamilies

[Browse »](#) [Search »](#) [Download »](#) [Take the Tour »](#)

### What is CATH?

**CATH is a classification of protein structures downloaded from the Protein Data Bank.** We group protein domains into superfamilies when there is sufficient evidence they have diverged from a common ancestor.

- Search CATH by text, ID or keyword
- Search CATH by protein sequence (FASTA)
- Search CATH by PDB structure
- Browse CATH Hierarchy
- CATH Release Notes
- CATH Tutorials

### Example pages

- PDB "2bop"
- Domain "1cukA01"
- Relatives of "1cukA01"
- Superfamily "HUPS"
- Functional Family
- FunFam Alignment
- Search for "enolase"
- Superfamily Comparison

### Latest Release Statistics

CATH v4.0 based on PDB dated March 26, 2013	
235,858	<a href="#">CATH Domains</a>
2,738	<a href="#">CATH Superfamilies</a>
69,058	<a href="#">Annotated PDBs</a>

Gene3D v12 released March 18, 2012	
6,131	<a href="#">Cellular Genomes</a>
21,662,155	<a href="#">Protein Sequences</a>
25,615,754	<a href="#">CATH Domain Predictions</a>