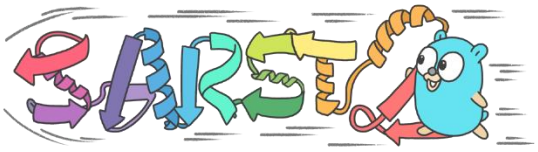
# SARST2 使用手册

## 简体中文版



## 目录

# 1. 关于此软件

SARST2 (Structural similarity search Aided by Ramachandran Sequential Transformation, version 2) 是高效能蛋白质结构比对算法，能对特定查询蛋白 (query protein) 作数据库结构比对搜索，也能执行双蛋白比对 (pairwise alignment)。

此软件随下列期刊论文发表，且会频繁于论文所附网址提供更新版本：

| | |
|---|---|
| 标题 | SARST2, a high-throughput protein structure alignment algorithm for searching massive databases |
| 作者 | Wei-Cheng Lo (罗惟正)*, Arieh Warshel, Chia-Hua Lo (罗佳华), Chia Yee Choke (祝佳怡), Yan-Jie Li (李延杰), Shih-Chung Yen (颜士中), Jyun-Yi Yang (杨钧诒) and Shih-Wen Weng (翁诗玟) |
| 机构 | 中华民国、台湾、新竹、国立阳明交通大学计算生物及系统生物研究所 |

*通讯作者 (WadeLo@nycu.edu.tw)

下载网址：
https://github.com/NYCU-10lab/sarst
https://10lab.ceb.nycu.edu.tw/sarst2

# 2. 下载、解压缩与安装

最新版 SARST2 程序及预先格式化好的目标数据库 (target databases) 可由上列网址取得。下载压缩文件后，须以 tar, gzip, 或 zip 程序解压缩，端看所下载的文件格式。解压缩后的 SARST2 程序是编译好的执行文件，免安装即可执行。

例如，所下载的压缩文件是 SARST2-v2.0.26-Linux.x86_64.tar.gz，在 Linux 环境可以下列指令解压缩：

```
tar xfp SARST2-v2.0.26-Linux.x86_64.tar.gz
```

解压缩后，可用下列指令执行 sarst2 程序：

```
cd SARST2-v2.0.26-Linux.x86_64/bin
chmod +x sarst2
./sarst2 -h
```

## 3. 各目录之内容

| | |
|---|---|
| bin | 64 位的 Linux, Windows, 和 macOS 执行文件。 |
| dbs | 预先格式化好的 PDB-2022 和 SCOP-2.07 目标数据库。 |
| doc | 使用手册，多种语言版本。 |

## 4. 快速指引

本软件有三支主要程序，简述如下：

| | |
|---|---|
| sarst2 | SARST2 蛋白质结构比对算法之实作程序。 |
| formatdb | 数据库格式化程序。让用户能自制目标数据库，以利结构比对搜索。 |
| readdb | 序列读取程序。能从格式化过的目标数据库中撷取出各蛋白质的胺基酸序列及结构编码序列。 |

执行上列程序，不下任何参数 (或下-h)，可获得各程序之纯文本简要说明。

Linux, macOS

```
./sarst2
./formatdb
./readdb
./sarst2 -h
./formatdb -h
./readdb -h
```

Windows (Cmd or PowerShell)

```
.\sarst2.exe
.\formatdb.exe
.\readdb.exe
.\sarst2.exe -h
.\formatdb.exe -h
.\readdb.exe -h
```

# 5. sarst2 (结构比对主程序) 使用说明

## 5.1 指令格式

```
./sarst2    查询蛋白结构          待比对蛋白结构          [参数]
            > PDB 或 CIF 文件     > 可以是
                                  1. -db 预先格式化的目标数据库
                                  2. 数个 PDB/CIF 文件
                                  3. 数个文件夹，内含 PDB/CIF 文件
                                  4. 一个 PDB/CIF 结构文件（双蛋白比对）
```

## 5.2 程序参数

| | | |
|---|---|---|
| -db | [字符串] | 欲搜索之目标数据库，内含多个待比对蛋白质结构。<br>(无默认值) |
| -brief | [整数] | 输出最多几笔高分待比对蛋白之结构相似度摘要。<br>(默认值 500) |
| -detail | [整数] | 输出最多几笔高分待比对蛋白之详细结构比对结果。<br>(默认值 500) |
| -t | [整数] | 欲使用多少线程。<br>必须 ≥0；若为 0，将自动设定为本计算机之处理器总核心数。<br>(默认值 0，即处理器总核心数) |
| -w | [整数] | 比对字节大小 (word size)。<br>(默认值 5) |
| -orderby | [整数] | 依下列某因子来排列结果清单，1: Conf-score--, 2: TM-score--, 3: sequence identity--, or 4: RMSD++。其中 -- 和 ++ 代表递减和递增排序。<br>(默认值 1，即相似信心指数)<br>(在双蛋白比对模式中，此参数无作用) |
| -mode | [整数] | 比对搜索模式，1: 精准, 2: 平衡, 3: 快速，其他: 自动判断。<br>(默认值 auto，即交由程序自动判断) |
| -f | [整数] | 细部过滤器，0: 停用, 1: 启用，其他: 自动判断。<br>(默认值 auto，即交由程序自动判断)<br>(在双蛋白比对模式中，此参数固定为 0) |

| | | |
|---|---|---|
| -C | [实数] | 相似信心指数 (confidence score) 之最低门坎。<br>必须介于 0 到 1；若为 0，则表示不设门坎。<br>(默认值 0.5)<br>(在双蛋白比对模式中，此参数无作用) |
| -pC | [实数] | 精算步骤所得 pC-value，即 –log2(C)，之上限。<br>必须 ≥ 0；若为 0，则表示不设上限。<br>(默认值 1.0，相当于 -C = 0.5)<br>(在双蛋白比对模式中，此参数无作用) |
| -e | [实数] | pC-value 之上限，应用到各过滤与精算步骤。<br>当-e 和-pC 的设定值相同时，-e 会过滤掉较多不相似的结构。<br>必须 ≥ 0；若为 0，则表示不设上限。<br>(默认值 1.0)<br>(在双蛋白比对模式中，此参数无作用) |
| -tmcut | [实数] | TM-score 结构相似度分数之上限。<br>必须 ≥ 0；若为 0，则表示不设上限。<br>SARST2 计算所得之 TM-score ≥ 0.7 代表该等蛋白质在演化上可能是同族类似物 (基于 SCOP 的分类方式)。<br>(默认值 0.15)<br>(在双蛋白比对模式中，此参数无作用) |
| -mem | [T/F] | 计算过程中，将所有待比对蛋白之信息存放于内存中。<br>(默认值 T，是)<br>(在双蛋白比对模式中，此参数无作用) |
| -q | [T/F] | 简易输出。以简洁且较易以程序自动解析的格式来输出结果。<br>(默认值 F，否) |
| -sa | [T/F] | 显示以结构迭合为基础的序列比对结果。<br>(默认值 T，是) |
| -mat | [T/F] | 显示将待比对与查询蛋白之结构迭合所需的坐标转置矩阵。<br>(默认值 F，否) |
| -nmsbj | [T/F] | 以待比对蛋白之大小来标准化 TM-score。<br>(默认值 F，否) |
| -nmavg | [T/F] | 以查询蛋白与待比对蛋白之平均大小来标准化 TM-score。<br>(默认值 F，否) |
| -nmusr | [实数] | 以使用者自定义的蛋白质大小来标准化 TM-score。<br>必须 ≥ 查询和待比对蛋白之大小的较小值；否则所求得之 TM-score 可能会大于 1。 |

| | | |
|---|---|---|
| -d | [实数] | 设定 TM-score 计分公式中的数值缩放参数 d0，如 5.0 Angstroms (Å)。 |
| -ml | [T/F] | 启用机器学习。<br>(默认值 T，是)<br>(在双蛋白比对模式中，此参数无作用) |
| -fdp | [字符串] | 设定过滤步骤中，结构编码字符串比对所用的动态规划算法。<br>支持选项：NW (Needleman-Wunsch), SW (Smith-Waterman)<br>(默认值 NW) |
| -rdp | [字符串] | 设定精算步骤中，结构编码字符串比对所用的动态规划算法。<br>支持选项：NW (Needleman-Wunsch), SW (Smith-Waterman)<br>(默认值 NW) |
| -swp | [字符串] | 启用暂存盘，且设定其路径。<br>使用暂存盘可减少内存之用量。<br>(预设不启用) |
| -Sout | [字符串] | 设定输出结构迭合档案的文件夹并启用此输出。<br>若该文件夹不存在，将尝试建立之。<br>结构迭合档案的数量受-detail 参数之限制。<br>(预设不启用) |
| -html | [字符串] | 设定输出 HTML 结果展示文件之文件夹并启用此输出。<br>程序会自动在该 HTML 文件夹中储存结构迭合档案。<br>(预设不启用) |
| -jsmol | [字符串] | 设定 JSmol JavaScript 套件之路径，以利于 HTML 输出中展示迭合结构。<br>可以是本机磁盘文件夹或 HTTP(S) URL。<br>(无默认值)<br>(trial URL: "https://10lab.ceb.nycu.edu.tw/ext/jsmol") |
| -pssm_out | [字符串] | 设定用于输出本算法中所使用的胺基酸序列和结构编码的 PSSM 的档案，并启用此输出。<br>(无默认值) |
| -pssm_pC | [实数] | 建置 PSSM 时，用于筛选比对结果的 pC-value 上限。<br>(默认值 0.05) |
| -h | | 显示说明讯息。 |

## 5.3 指令范例：对数据库作结构相似度搜索

于目标数据库中搜索查询蛋白之结构类似物

```
./sarst2 Qry.pdb -db my_db/my_proteins.db -brief 10 -w 7 -e 0.1
./sarst2 Qry.pdb -db my_db/my_proteins.db -brief 10 -d 5.0 -sa F
```

在此范例中，目标数据库储存于文件夹 "my_db" 中，而 "my_proteins.db" 是目标数据库档案的主文件名，详见「**formatdb (数据库制作程序) 使用说明**」。

## 5.4 指令范例：一对多蛋白质结构比对 (one-against-all alignments)

于使用者列举的待比对蛋白中搜索查询蛋白之结构类似物

```
./sarst2 Qry.pdb Sbj1.pdb Sbj2.cif Sbj3.pdb -mat T
```

依使用者以通配符列举的待比对蛋白档案路径搜索查询蛋白之结构类似物

```
./sarst2 Qry.pdb "set1/*.pdb" "set2/1a???.cif" -nmavg T
```

在此范例中，"set1/*.pdb" 和 "set2/1a???.cif" 用引号包起来并带有通配符。sarst2 程序会自行取得与通配符相符的文件名。若不加引号，则操作系统会自动列出符合通配符的所有文件名；当文件数量太多时，命令行参数将变得过长而无法运作。建议使用引号（即由 sarst2 程序自行列出文件），而非仰赖操作系统的默认文件列出行为。

于使用者列举的待比对蛋白文件夹中搜索查询蛋白之结构类似物

```
./sarst2 Qry.pdb set1 set2 -nmavg T
```

此范例中，set1 和 set2 是可能包含蛋白质结构档案的文件夹。两文件夹中所有文件清单 (即 set1/* 和 set2/*) 将由 sarst2 程序自行取得；若发现 PDB/CIF 格式的结构文件，该等结构将与查询蛋白之结构进行比对。

## 5.5 指令范例：双蛋白质结构比对 (pairwise structure alignment)

将查询蛋白与另一个蛋白作结构比对

```
./sarst2 Qry.pdb Sbj.pdb -sa F
./sarst2 Qry.pdb Sbj.cif -mat T
```

## 5.6 输出蛋白质结构迭合档案

输出查寻蛋白与各待比对蛋白之结构迭合文件

```
./sarst2 Qry.pdb -db prot/myDb -detail 100 -Sout output_folder
./sarst2 Qry.pdb "set1/*.cif" -detail 100 -Sout output_folder
```

使用 "-Sout output_folder" 参数，可将蛋白质结构迭合文件 (PDB 格式) 输出到使用者指定的文件夹。输出的数量由 -detail 参数定义。在该文件夹中，各结构迭合文件被命名为 Qry-SbjSN.pdb，SN 是结果清单中该待比对蛋白之序号。在结构迭合文件中，查询和待比对蛋白结构的多肽链代号分别为 Q 和 S，两链以 TER 记录分隔，如下所示：

```
ATOM    150  CA  LEU Q 150      29.000  -8.400   0.800                       C
ATOM    151  CA  GLY Q 151      26.000  -9.600   2.600                       C
ATOM    152  CA  TYR Q 152      25.400  -6.800   5.000                       C
ATOM    153  CA  GLN Q 153      23.600  -3.800   3.600                       C
ATOM    154  CA  GLY Q 154      22.800  -2.800   7.200                       C
TER
ATOM      1  CA  MET S   1      24.400   9.800 -10.000                       C
ATOM      2  CA  VAL S   2      27.200  11.800 -11.400                       C
ATOM      3  CA  LEU S   3      28.800  15.200 -10.400                       C
ATOM      4  CA  SER S   4      29.800  17.800 -13.000                       C
ATOM      5  CA  GLU S   5      33.400  19.200 -13.000                       C
ATOM      6  CA  GLY S   6      32.000  22.400 -11.600                       C
```

上图还展示着，只有 α 碳 (Cα) 原子会出现在迭合结构文件中。因为 SARST2 在计算过程中仅使用 Cα 坐标。查询蛋白之结构于空间中的摆放方式，在所有迭合结构文件中皆为固定。各待测蛋白根据其与查询蛋白之比对结果，经坐标旋转、平移后与查询蛋白之结构相迭合。

想用可视化方式查看迭合结构，我们推荐 RasMol (http://www.openrasmol.org/) 和 RasWin。由于迭合档案中只有 Cα 原子，RasMol 之显示模式应设定为 "backbone"。

## 5.7 输出 HTML 交互式结果展示网页

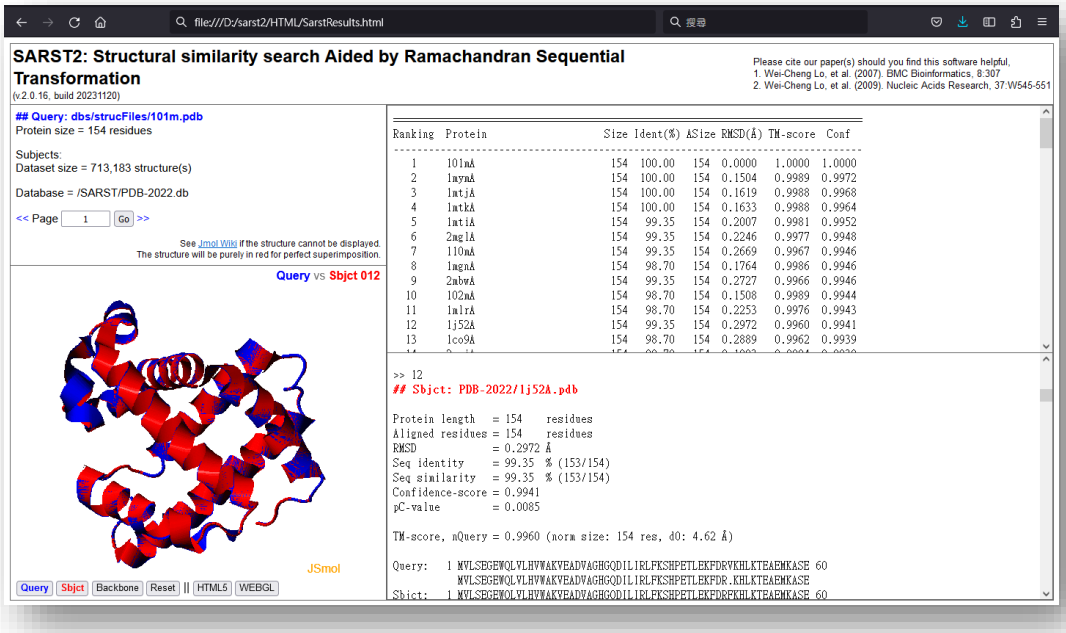
输出附加在线 JSmol 动态结构展示功能的 HTML 结果网页

```
./sarst2 Qry.pdb -db my_db/my_proteins.db -html output_folder
-jsmol "https://10lab.ceb.nycu.edu.tw/ext/jsmol"
```

输出附加本机磁盘 JSmol 动态结构展示功能的 HTML 结果网页 (Windows)

```
./sarst2 Qry.pdb "set1/*.cif" -detail 100 -html output_folder
-jsmol "file:///D:/bioinfo/software/jsmol"
```

使用 "-html output_folder" 参数，HTML 结果展示文件和蛋白质结构迭合文件将输出到该指定文件夹。此-html 选项必须与-jsmol 一起使用，后者指定 JSmol 套件 (2013 版) 的 URL。JSmol 是一款于网页浏览器运作的交互式立体分子结构查看器，支持多数主流网页浏览器。




在 HTML 输出文件夹中，主控文件是 "SarstResults.html"，请以网页浏览器开启；其他 HTML 档案乃以内框架 (inner frame) 形式整合到网页画面中。有个名为 "sup" 的子文件夹，储存着查询蛋白以及与之结构类似的各待比对蛋白的结构迭合文件。

用户可能需要调整操作系统、浏览器、或防病毒软件之安全性设定，允许浏览器执行 JavaScript 及读取 "sup" 目录中的蛋白质结构迭合档案，JSmol 立体结构展示对象才能正常运作。

# 6. formatdb (数据库制作程序) 使用说明

## 6.1 指令格式

```
./formatdb      待比对蛋白结构     -db database    [参数]
                ------------
                > 可以是
                1. 数个 PDB/CIF 文件
                2. 数个文件夹，内含 PDB/CIF 文件
                3. 纯文本文件，内为 PDB/CIF 结构列表
```

## 6.2 5.2 程序参数

| | | |
|---|---|---|
| -db | [字符串] | 欲建立之目标数据库，内含多个待比对蛋白质结构。<br>(无默认值) |
| -flist | [字符串] | 纯文本文件，内容为蛋白质结构档案列表。本参数可与其他档案路径参数合并使用。<br>(无默认值) |
| -t | [整数] | 欲使用多少线程。<br>必须 ≥0；若为 0，将自动设定为本计算机之处理器总核心数。<br>(默认值 0，即处理器总核心数) |
| -split | [整数] | 将数据库拆分为子集合，每个子集合最多包含此选项所指定的待比对蛋白数量。此拆分有助防止数据库档案超过磁盘的档案大小限制。<br>(无默认值) |
| -save_disk | [T/F] | 将原子坐标从小数点后三位四舍五入到小数点后一位，以节省磁盘空间。<br>(默认值 F) |
| -keep_order | [T/F] | 使数据库中储存的待比对结构之顺序与格式化指令中的蛋白质输入顺序相同。设定成 T 会降低数据库之格式化速度。<br>(默认值 F) |
| -h | | 显示说明讯息。 |

## 6.3 指令范例

将用户列举的待比对蛋白结构文件制成目标数据库

```
./formatdb Sbj1.pdb Sbj2.cif Sbj3.pdb -db myDb -keep_order T
```

制作完成后，磁盘中将出现几个文件名以 myDb 开头的数据库文件。启用-keep_order 可使目标数据库中待比对蛋白的存放顺序跟它们在命令行参数中被列出的顺序一样。

找出使用者以通配符列举的待比对蛋白结构文件，制成目标数据库

```
./formatdb "set1/*.pdb" "set2/*.cif" Sbj1.pdb Sbj2.cif -db myDb
```

列出待比对蛋白质之结构文件时，可混用含通配符和不含通配符的参数。建议含通配符的参数以引号包住。

依档案列表指定的蛋白结构，制成目标数据库

```
./formatdb -flist protlist.txt Sbj1.pdb Sbj2.cif -db myDb
```

文件清单 protlist.txt 的格式为一列一个文件路径。

在使用者列举的待比对蛋白文件夹中找出所有结构文件，制成目标数据库

```
./formatdb folder1 folder2 -db myDb -save_disk T -split 50000
```

启用-save_disk 会将 Cα 坐标四舍五入到小数点后一位以节省储存空间。而设定 "-split 50000" 将产生多个数据库子集合，每个子集合最多包含 50,000 个结构。若格式化后的数据库文件大小有可能超过某些操作系统或磁盘格式的上限，此分割功能将特别有用。

# 7. readdb (数据库序列读取程序) 使用说明

## 7.1 指令格式

```
./readdb  目标数据库          输出文件          [-seq 序列类型]
          ---------           -------           ------------
          > 必须是 SARST2     > 输出格式为      > 可以是
            数据库格式          FASTA           1. AA
                                                2. AAT
                                                3. SARST
                                                4. SSE
```

## 7.2 程序参数

| | | |
|---|---|---|
| -seq | [字符串] | 指定输出序列之类型。<br>AA　　胺基酸序列<br>AAT　　五分类的胺基酸类型序列<br>SARST　SARST 结构编码序列<br>SSE　　四分类的二级结构元素 (SSE) 序列<br>(默认值 AA) |
| -h | | 显示说明讯息。 |

## 7.3 指令范例

自 SARST2 目标数据库中读取出所需类型之序列

```
./readdb my_db/my_proteins.db seqs.fasta
./readdb my_db/my_proteins.db seqs.fasta -seq SARST
./readdb my_db/my_proteins.db seqs.fasta -seq AAT
```

未指定-seq 时，预设输出胺基酸序列。输出之序列文件 (seqs.fasta) 格式为 FASTA。若 readdb 执行前输出文件已存在，则其将被覆写。