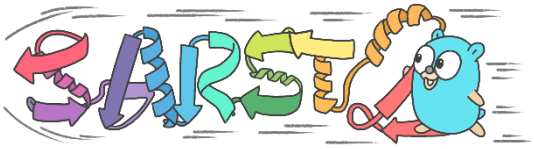
SARST2 使用手冊

繁體中文版



目錄

1. 關於此軟體

SARST2 (Structural similarity search Aided by Ramachandran Sequential Transformation, version 2) 是高效能蛋白質結構比對演算法，能對特定查詢蛋白 (query protein) 做資料庫結構比對搜索，也能執行雙蛋白比對 (pairwise alignment)。

此軟體隨下列期刊論文發表，且會頻繁於論文所附網址提供更新版本：

標題	SARST2, a high-throughput protein structure alignment algorithm for searching massive databases
作者	Wei-Cheng Lo (羅惟正)*, Arieh Warshel, Chia-Hua Lo (羅佳華), Chia Yee Choke (祝佳怡), Yan-Jie Li (李延杰), Shih-Chung Yen (顏士中), Jyun-Yi Yang (楊鈞詒) and Shih-Wen Weng (翁詩玟)
機構	中華民國、臺灣、新竹、國立陽明交通大學計算生物及系統生物研究所

*通訊作者

下載網址：
https://github.com/NYCU-10lab/sarst
https://10lab.ceb.nycu.edu.tw/sarst2

2. 下載、解壓縮與安裝

最新版 SARST2 程式及預先格式化好的目標資料庫 (target databases) 可由上列網址取得。下載壓縮檔後，須使用 tar, gzip, 或 zip 程式解壓縮，端看所下載的檔案格式。解壓縮後的 SARST2 程式是編譯好的執行檔，免安裝即可執行。

例如，所下載的壓縮檔是 SARST2-v2.0.26-Linux.x86_64.tar.gz, 在 Linux 環境下可以下列指令解壓縮：

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

解壓縮後，可用下列指令執行 sarst2 程式：

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

3. 各目錄之內容

bin	64 位元的 Linux, Windows, 和 macOS 執行檔。
dbs	預先格式化好的 PDB-2022 和 SCOP-2.07 目標資料庫。
doc	使用手冊，多種語言版本。

4. 快速指引

本軟體有三支主要程式，簡述如下：

sarst2	SARST2 蛋白質結構比對演算法之實作程式。
formatdb	資料庫格式化程式。讓使用者能自製目標資料庫，以利結構比對搜索。
readdb	序列讀取程式。能從格式化過的目標資料庫中擷取出各蛋白質的胺基酸序列及結構編碼序列。

執行上列程式，不下任何參數 (或下-h)，可獲得各程式之純文字簡要說明。

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 (結構比對主程式) 使用說明

5.1 指令格式

```
./sarst2    查詢蛋白結構          待比對蛋白結構          [參數]
            ----------            ------------
            > PDB 或 CIF 檔案     > 可以是
                                  1. -db 預先格式化的目標資料庫
                                  2. 數個 PDB/CIF 檔案
                                  3. 數個資料夾，內含 PDB/CIF 檔案
                                  4. 一個 PDB/CIF 結構檔案（雙蛋白比對）
```

5.2 程式參數

-db	[字串]	欲搜索之目標資料庫，內含多個待比對蛋白質結構。 (無預設值)
-brief	[整數]	輸出最多幾筆高分待比對蛋白之結構相似度摘要。 (預設值 500)
-detail	[整數]	輸出最多幾筆高分待比對蛋白之詳細結構比對結果。 (預設值 500)
-t	[整數]	欲使用多少執行緒。 必須 ≥0；若為 0，將自動設定為本電腦之處理器總核心數。 (預設值 0, 即處理器總核心數)
-w	[整數]	比對字節大小 (word size)。 (預設值 5)
-orderby	[整數]	依下列某因子來排列結果清單，1: Conf-score--, 2: TM-score--, 3: sequence identity--, or 4: RMSD++。其中 -- 和 ++ 代表遞減和遞增排序。 (預設值 1，即相似信心指數) (在雙蛋白比對模式中，此參數無作用)
-mode	[整數]	比對搜索模式，1: 精準, 2: 平衡, 3: 快速, 其他: 自動判斷。 (預設值 auto，即交由程式自動判斷)
-f	[整數]	細部過濾器，0: 停用, 1: 啟用, 其他: 自動判斷。 (預設值 auto，即交由程式自動判斷) (在雙蛋白比對模式中，此參數固定為 0)

-C	[實數]	相似信心指數 (confidence score) 之最低門檻。 必須介於 0 到 1；若為 0，則表示不設門檻。 (預設值 0.5) (在雙蛋白比對模式中，此參數無作用)
-pC	[實數]	精算步驟所得 pC-value，即 –log2(C)，之上限。 必須 ≥0；若為 0，則表示不設上限。 (預設值 1.0, 相當於 -C = 0.5) (在雙蛋白比對模式中，此參數無作用)
-e	[實數]	pC-value 之上限，應用到各過濾與精算步驟。 當-e 和-pC 的設定值相同時，-e 會過濾掉較多不相似的結構。 必須 ≥0；若為 0，則表示不設上限。 (預設值 1.0) (在雙蛋白比對模式中，此參數無作用)
-tmcut	[實數]	TM-score 結構相似度分數之上限。 必須 ≥0；若為 0，則表示不設上限。 SARST2 計算所得之 TM-score ≥ 0.7 代表該等蛋白質在演化上可能是同族類似物 (基於 SCOP 的分類方式)。 (預設值 0.15) (在雙蛋白比對模式中，此參數無作用)
-mem	[T/F]	計算過程中，將所有待比對蛋白之資訊存放於記憶體中。 (預設值 T，是) (在雙蛋白比對模式中，此參數無作用)
-q	[T/F]	簡易輸出。以簡潔且較易以程式自動解析的格式來輸出結果。 (預設值 F，否)
-sa	[T/F]	顯示以結構疊合為基礎的序列比對結果。 (預設值 T，是)
-mat	[T/F]	顯示將待比對與查詢蛋白之結構疊合所需的座標轉置矩陣。 (預設值 F，否)
-nmsbj	[T/F]	以待比對蛋白之大小來標準化 TM-score。 (預設值 F，否)
-nmavg	[T/F]	以查詢蛋白與待比對蛋白之平均大小來標準化 TM-score。 (預設值 F，否)
-nmusr	[實數]	以使用者自訂的蛋白質大小來標準化 TM-score。 必須 ≥ 查詢和待比對蛋白之大小的較小值；否則所求得之 TM-score 可能會大於 1。

-d	[實數]	設定 TM-score 計分公式中的數值縮放參數 d0，如 5.0 Angstroms (Å)。
-ml	[T/F]	啟用機器學習。 (預設值 T，是) (在雙蛋白比對模式中，此參數無作用)
-fdp	[字串]	設定過濾步驟中，結構編碼字串比對所用的動態規劃演算法。 支援選項：NW (Needleman-Wunsch), SW (Smith-Waterman) (預設值 NW)
-rdp	[字串]	設定精算步驟中，結構編碼字串比對所用的動態規劃演算法。 支援選項：NW (Needleman-Wunsch), SW (Smith-Waterman) (預設值 NW)
-swp	[字串]	啟用暫存檔，且設定其路徑。 使用暫存檔可減少記憶體之用量。 (預設不啟用)
-Sout	[字串]	設定輸出結構疊合檔案的資料夾並啟用此輸出。 若該資料夾不存在，將嘗試建立之。 結構疊合檔案的數量受-detail 參數之限制。 (預設不啟用)
-html	[字串]	設定輸出 HTML 結果展示文件之資料夾並啟用此輸出。 程式會自動在該 HTML 資料夾中儲存結構疊合檔案。 (預設不啟用)
-jsmol	[字串]	設定 JSmol JavaScript 套件之路徑，以利於 HTML 輸出中展示疊合結構。 可以是本機磁碟資料夾或 HTTP(S) URL。 (無預設值) (trial URL: "https://10lab.ceb.nycu.edu.tw/ext/jsmol")
-pssm_out	[字串]	設定用於輸出本演算法中所使用的胺基酸序列和結構編碼的 PSSM 的檔案，並啟用此輸出。 (無預設值)
-pssm_pC	[實數]	建置 PSSM 時，用於篩選比對結果的 pC-value 上限。 (預設值 0.05)
-h		顯示說明訊息。

5.3 指令範例：對資料庫做結構相似度搜索

於目標資料庫中搜索查詢蛋白之結構類似物

```
./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
```

在此範例中，目標資料庫儲存於資料夾 "my_db" 中，而 "my_proteins.db" 是目標資料庫檔案的主檔名，詳見「**formatdb 程式使用說明**」。

5.4 指令範例：一對多蛋白質結構比對 (one-against-all alignments)

於使用者列舉的待比對蛋白中搜索查詢蛋白之結構類似物

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

依使用者以萬用字元列舉的待比對蛋白檔案路徑搜索查詢蛋白之結構類似物

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

在此範例中，"set1/*.pdb" 和 "set2/1a???.cif" 用引號包起來並帶有萬用字元。sarst2 程式會自行取得與萬用字元相符的檔名。若不加引號，則作業系統會自動列出符合萬用字元的所有檔名；當檔案數量太多時，命令列參數將變得過長而無法運作。建議使用引號（即由 sarst2 程式自行列出檔案），而非仰賴作業系統的預設檔案列出行為。

於使用者列舉的待比對蛋白資料夾中搜索查詢蛋白之結構類似物

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

此範例中，set1 和 set2 是可能包含蛋白質結構檔案的資料夾。兩資料夾中所有檔案清單 (即 set1/* 和 set2/*) 將由 sarst2 程式自行取得；若發現 PDB/CIF 格式的結構檔案，該等結構將與查詢蛋白之結構進行比對。

5.5 指令範例：雙蛋白質結構比對 (pairwise structure alignment)

將查詢蛋白與另一個蛋白做結構比對

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

5.6 輸出蛋白質結構疊合檔案

輸出查尋蛋白與各待比對蛋白之結構疊合檔案

```
./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" 參數，可將蛋白質結構疊合檔案 (PDB 格式) 輸出到使用者指定的資料夾。輸出的數量由 -detail 參數定義。在該資料夾中，各結構疊合檔案被命名為 Qry-SbjSN.pdb，SN 是結果清單中該待比對蛋白之序號。在結構疊合檔案中，查詢和待比對蛋白結構的多肽鏈代號分別為 Q 和 S，兩鏈以 TER 記錄分隔，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
```

上圖還展示著，只有 α 碳 (Cα) 原子會出現在疊合結構檔案中。因為 SARST2 在計算過程中僅使用 Cα 座標。查詢蛋白之結構於空間中的擺放方式，在所有疊合結構檔案中皆為固定。各待測蛋白根據其與查詢蛋白之比對結果，經座標旋轉、平移後與查詢蛋白之結構相疊合。

想用視覺化方式查看疊合結構，我們推薦 RasMol (http://www.openrasmol.org/) 和 RasWin。由於疊合檔案中只有 Cα 原子，RasMol 之顯示模式應設定為 "backbone"。

5.7 輸出 HTML 互動式結果展示網頁

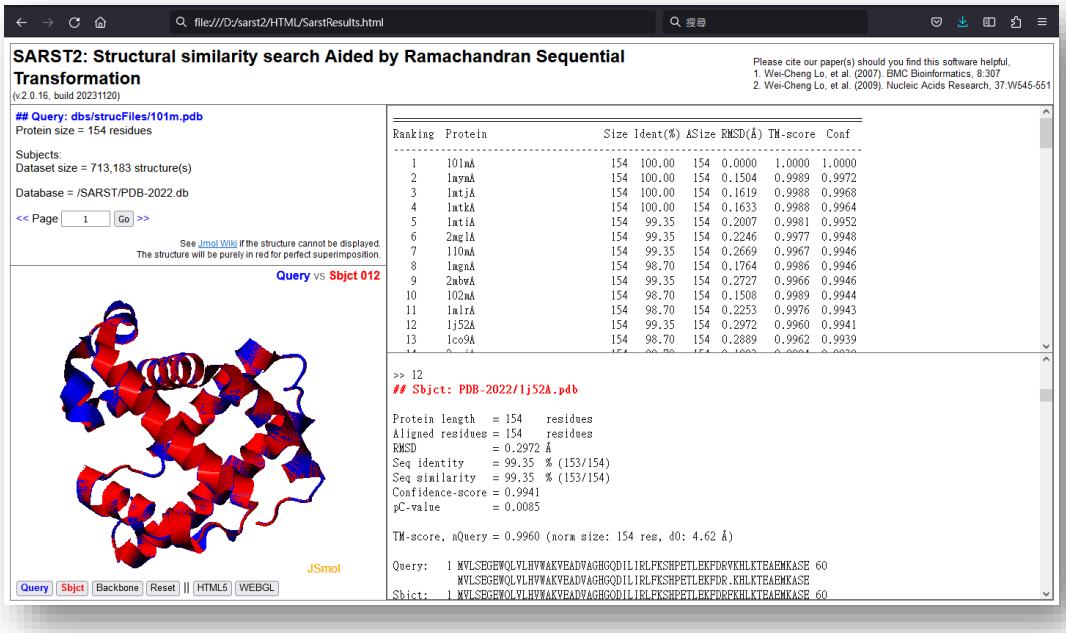
輸出附加線上 JSmol 動態結構展示功能的 HTML 結果網頁

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

輸出附加本機磁碟 JSmol 動態結構展示功能的 HTML 結果網頁 (Windows)

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

使用 "-html output_folder" 參數，HTML 結果展示文件和蛋白質結構疊合檔案將輸出到該指定資料夾。此-html 選項必須與-jsmol 一起使用，後者指定 JSmol 套件 (2013 版) 的 URL。JSmol 是一款於網頁瀏覽器運作的互動式立體分子結構檢視器，支援多數主流網頁瀏覽器。



在 HTML 輸出資料夾中，主檔案是 "SarstResults.html", 請以網頁瀏覽器開啟；其他 HTML 檔案乃以內框架 (inner frame) 形式整合到網頁畫面中。有個名為 "sup" 的子資料夾，儲存著查詢蛋白以及與之結構類似的各待比對蛋白的結構疊合檔案。

使用者可能需要調整作業系統、瀏覽器、或防毒軟體之安全性設定，允許瀏覽器執行 JavaScript 及讀取 "sup" 目錄中的蛋白質結構疊合檔案，JSmol 立體結構展示物件才能正常運作。

6. formatdb (資料庫製做程式) 使用說明

6.1 指令格式

```
./formatdb      待比對蛋白結構     -db database    [參數]
                ------------
                > 可以是
                1. 數個 PDB/CIF 檔案
                2. 數個資料夾，內含 PDB/CIF 檔案
                3. 純文字檔案，內為 PDB/CIF 結構清單
```

6.2 5.2 程式參數

-db	[字串]	欲建立之目標資料庫，內含多個待比對蛋白質結構。 (無預設值)
-flist	[字串]	純文字檔，內容為蛋白質結構檔案清單。本參數可與其他檔案路徑參數合併使用。 (無預設值)
-t	[整數]	欲使用多少執行緒。 必須 ≥0；若為 0，將自動設定為本電腦之處理器總核心數。 (預設值 0, 即處理器總核心數)
-split	[整數]	將資料庫拆分為子集合，每個子集合最多包含此選項所指定的待比對蛋白數量。此拆分有助防止資料庫檔案超過磁碟的檔案大小限制。 (無預設值)
-save_disk	[T/F]	將原子座標從小數點後三位四捨五入到小數點後一位，以節省磁碟空間。 (預設值 F)
-keep_order	[T/F]	使資料庫中儲存的待比對結構之順序與格式化指令中的蛋白質輸入順序相同。設定成 T 會降低資料庫之格式化速度。 (預設值 F)
-h		顯示說明訊息。

6.3 指令範例

將使用者列舉的待比對蛋白結構檔案製成目標資料庫

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

製作完成後，磁碟中將出現幾個檔名以 myDb 開頭的資料庫檔案。啟用-keep_order 可使目標資料庫中待比對蛋白的存放順序跟它們在命令列參數中被列出的順序一樣。

找出使用者以萬用字元列舉的待比對蛋白結構檔案，製成目標資料庫

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

列出待比對蛋白質之結構檔案時，可混用含萬用字元和不含萬用字元的參數。建議含萬用字元的參數以引號包住。

依檔案清單指定的蛋白結構，製成目標資料庫

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

檔案清單 protlist.txt 的格式為一列一個檔案。

在使用者列舉的待比對蛋白資料夾中找出所有結構檔案，製成目標資料庫

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

啟用-save_disk 會將 Cα 座標四捨五入到小數點後一位以節省儲存空間。而設定 "-split 50000" 將產生多個資料庫子集合，每個子集合最多包含 50,000 個結構。若格式化後的資料庫檔案大小有可能超過某些作業系統或磁碟格式的上限，此分割功能將特別有用。

7. readdb (資料庫序列讀取程式) 使用說明

7.1 指令格式

```
./readdb  目標資料庫          輸出檔案          [-seq 序列類型]
          ---------           -------           ------------
          > 必須是 SARST2     > 輸出格式為      > 可以是
            資料庫格式          FASTA           1. 胺基酸序列
                                                2. 胺基酸類型序列
                                                3. SARST 結構編碼序列
                                                4. 四分類的二級結構元素序列
```

7.2 程式參數

-seq	[字串]	指定輸出序列之類型。 AA　　胺基酸序列 AAT　　五分類的胺基酸類型序列 SARST　SARST 結構編碼序列 SSE　　四分類的二級結構元素 (SSE) 序列 (預設值 AA)
-h		顯示說明訊息。

7.3 指令範例

自 SARST2 目標資料庫中讀取出所需類型之序列

```
./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 時，預設輸出胺基酸序列。輸出之序列檔案 (seqs.fasta) 格式為 FASTA。若 readdb 執行前輸出檔案已存在，則其將被覆寫。