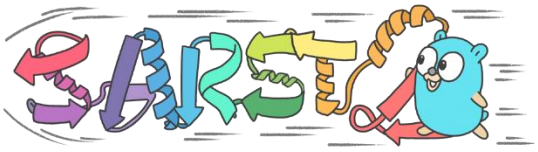
# SARST2 ユーザーマニュアル

## 日本語版



## 目次

# 1. ソフトウェアについて

SARST2 (Structural similarity search Aided by Ramachandran Sequential Transformation, version 2) は、高性能なタンパク質構造アラインメントアルゴリズムです。特定のクエリタンパク質（query protein）に対するデータベース構造比較検索、および２つのタンパク質間のペアワイズアラインメント（pairwise alignment）も実行できます。

本ソフトウェアは、下記の学術論文と共に発表されており、論文に記載の URL にて頻繁に更新版が提供されます。

| | |
|---|---|
| タイトル | SARST2, a high-throughput protein structure alignment algorithm for searching massive databases |
| 著者 | Wei-Cheng Lo*, Arieh Warshel, Chia-Hua Lo, Chia Yee Choke, Yan-Jie Li, Shih-Chung Yen, Jyun-Yi Yang and Shih-Wen Weng |
| 研究機関 | 中華民国、台湾、新竹、国立陽明交通大学<br>生物情報学とシステム生物学研究所 |

*責任著者（WadeLo@nycu.edu.tw)

ダウンロード URL：
https://github.com/NYCU-10lab/sarst
https://10lab.ceb.nycu.edu.tw/sarst2

# 2. ダウンロード、解凍、インストール

最新版の SARST2 プログラム、およびフォーマット済みのターゲットデータベースは上記の URL から入手できます。圧縮アーカイブのダウンロード後、ダウンロードしたアーカイブの形式に応じて、tar、gzip、または zip ユーティリティで解凍してください。解凍された SARST2 プログラムはコンパイル済みの実行ファイルであり、インストール不要でそのまま実行できます。

例えば、SARST2-v2.0.26-Linux.x86_64.tar.gz（Linux 環境の場合）をダウンロードした場合、以下のコマンドで解凍してください。

```
tar xfp SARST2-v.2.0.26-Linux.x86_64.tar.gz
```

解凍後、以下のコマンドでプログラムを実行できます。

```
cd SARST2-v.2.0.26-Linux.x86_64/bin
chmod +x sarst2
./sarst2 -h
```

## 3. ソフトウェアフォルダーの内容

| | |
|---|---|
| bin | 64-bit Linux、Windows、および macOS 用の実行ファイル。 |
| dbs | PDB-2022 と SCOP-2.07 のフォーマット済みデータベース。 |
| doc | 多言語版ユーザーマニュアル。 |

## 4. クイックガイド

ソフトウェアパッケージには３つの主要なプログラムがあります。

| | |
|---|---|
| sarst2 | SARST2 タンパク質構造アラインメントの実装プログラム。 |
| formatdb | ユーザーが SARST2 用フォーマット済みデータベースを自作できるデータベースジェネレーター。 |
| readdb | フォーマット済みデータベースから、各タンパク質のアミノ酸配列および構エンコード配列を抽出するためのツール。 |

引数なし、または -h を指定してプログラムを実行すると、クイックガイドが表示されます。

Linux, macOS

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

Windows (Cmd または PowerShell)

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

# 5. マニュアル：sarst2（構造比較主プログラム）

## 5.1 使い方

```
./sarst2    クエリ構造          検索対象          [オプション]
            ---------          -------
            > 1つの PDB/CIF ファイル   > できるのは
                                1．-db + 1つのフォーマット済みデータベース
                                2．複数の PDB/CIF ファイル
                                3．PDB/CIF ファイルのフォルダー
                                4．1つの PDB/CIF ファイル
                                   (ペアワイズアライメント)
```

## 5.2 プログラムのオプション

| | | |
|---|---|---|
| -db | [文字列] | 検索対象のターゲットデータベース。<br>複数の比較対象タンパク質構造が含まれています。<br>(既定値なし) |
| -brief | [整数] | 最大いくつかの高スコアタンパク質の構造類似性概要を出力します。<br>(既定値 500) |
| -detail | [整数] | 最大いくつかの高スコアタンパク質の詳細な構造比較結果を出力します。<br>(既定値 500) |
| -t | [整数] | 使用するスレッド数。<br>0 以上を指定してください。0 の場合、このコンピュータ全プロセッサコア数が自動的に設定されます。<br>(既定値 0、プロセッサコア数全て) |
| -w | [整数] | ワードサイズ (word size)。<br>(既定値 5) |
| -orderby | [整数] | 結果リストは以下の因子で並べ替えることができます。<br>1: Conf-score (--), 2: TM-score (--), 3: sequence identity (--), 4: RMSD (++)。<br>括弧内の「--」は降順、「++」は昇順を表します。<br>(既定値は 1、信頼度スコア)<br>(ペアワイズアライメントモードでは無効) |
| -mode | [整数] | 検索モード：<br>1: 精密、2: バランス、3: 高速、その他: 自動判別。<br>(既定値：プログラムによる自動判別) |
| -f | [整数] | 精密フィルタ：<br>0：無効、1：有効、その他：自動判定。<br>(既定値：プログラムによる自動判別)<br>(ペアワイズアライメントモードでは、0 に固定されます) |

| | | |
|---|---|---|
| -C | [実数] | 信頼度スコアの最小しきい値。<br>0 から 1 までの範囲で指定してください。0 の場合、しきい値は設定されません。<br>(既定値 0.5)<br>(ペアワイズアライメントモードでは無効) |
| -pC | [実数] | pC-value の上限、-log2(C) > 0 である必要があります。<br>0 の場合は、上限がないことを意味します。<br>(既定値 1.0、-C = 0.5)<br>(ペアワイズアライメントモードでは無効) |
| -e | [実数] | pC-value の上限は、各フィルタリングおよび精密計算ステップに適用されます。<br>-e と -pC の設定が同じ場合、-e の方がより多くの不相似な構造を除外します。<br>0 以上である必要があります。0 の場合、上限は設定されません。<br>(既定値 1.0)<br>(ペアワイズアライメントモードでは無効) |
| -tmcut | [実数] | TM-score 構造類似度スコアの上限です。<br>0 以上である必要があります。0 の場合は、上限は設定されません。SARST2 で計算された TM-score が 0.7 以上の場合、そのタンパク質は進化的に同族の類似体である可能性が高いです (SCOP の分類基準に基づく)。<br>(既定値 0.15)<br>(ペアワイズアライメントモードでは無効) |
| -mem | [T/F] | 計算プロセス中に、比較されるすべてのタンパク質の情報がメモリに格納されます。<br>(既定値 T、有効)<br>(ペアワイズアライメントモードでは無効) |
| -q | [T/F] | 簡易出力。<br>プログラムが自動的に解析しやすい、簡潔な形式で結果を出力します。<br>(既定値 F、無効) |
| -sa | [T/F] | 構造アライメントに基づく配列アライメント結果を表示。<br>(既定値 T、有効) |
| -mat | [T/F] | クエリタンパク質と対象タンパク質の構造を重ね合わせるための座標転置行列を表示します。<br>(既定値 F、無効) |
| -nmsbj | [T/F] | TM-score は、比較対象タンパク質のサイズで正規化されます。<br>(既定値 F、無効) |
| -nmavg | [T/F] | TM-score は、クエリタンパク質と対象タンパク質の平均サイズで正規化されます。<br>(既定値 F、無効) |
| -nmusr | [実数] | TM-score をユーザー定義のタンパク質サイズで正規化します。この値は、クエリタンパク質と対象タンパク質の、より小さい方のサイズ以上である必要があります。そうしないと、得られる TM-score が 1 より大きくなる可能性があります。 |

| | | |
|---|---|---|
| -d | [実数] | TM-score のスコア計算式における数値スケーリングパラメーター d0 を設定します。例えば、5.0 オングストローム(Å)などです。 |
| -ml | [T/F] | 機械学習を有効にします。<br>(既定値 T、有効)<br>(ペアワイズアライメントモードでは無効) |
| -fdp | [文字列] | フィルタリングステップにおける、構造エンコード文字列の比較に用いる動的計画法アルゴリズムを設定します。<br>サポートオプション：<br>NW (Needleman-Wunsch)、SW (Smith-Waterman)<br>(既定値 NW) |
| -rdp | [文字列] | 精密計算するステップにおける、構造エンコード文字列の比較に用いる動的計画法アルゴリズムを設定します。<br>サポートオプション：<br>NW (Needleman-Wunsch)、SW (Smith-Waterman)<br>(既定値 NW) |
| -swp | [文字列] | スワップファイルを有効にし、そのパスを設定します。<br>スワップファイルを使用することで、メモリ使用量を削減できます。<br>(既定では無効) |
| -Sout | [文字列] | 構造重ね合わせファイルを出力するフォルダーです。<br>フォルダーが存在しない場合は作成されます。<br>出力される重ね合わせファイルの数は、-detail オプションによって制限されます。<br>(既定では無効) |
| -html | [文字列] | HTML 出力フォルダーを作成します。<br>HTML フォルダー内には、構造重ね合わせファイルも生成されます。<br>(既定では無効) |
| -jsmol | [文字列] | HTML 出力に重ね合わせ構造を表示するための、JSmol JavaScript パッケージのパスを設定します。<br>これは、ローカルディスクフォルダー、または HTTP(S) URL のいずれかです。<br>(既定値なし)<br>(試用 URL: "https://10lab.ceb.nycu.edu.tw/ext/jsmol") |
| -pssm_out | [文字列] | このアルゴリズムで適用される構造コードとアミノ酸配列コードの PSSM を格納するファイルです。<br>(既定値なし) |
| -pssm_pC | [実数] | PSSM 構築における pC-value のカットオフ値。<br>(既定値 0.05) |
| -h | | クイックガイドを表示されます。 |

## 5.3コマンド例：データベースの構造類似性検索

対象データベース内で、クエリタンパク質の構造類似体検索

```
./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
```

この例では、対象データベースは "my_db" フォルダーに保存されており、"my_proteins.db" は対象データベースファイルの本体ファイル名です (詳細は「**マニュアル：formatdb**」を参照)。

## 5.4コマンド例：1 対多のタンパク質構造比較

ユーザーが列挙した比較対象タンパク質の中から、クエリタンパク質の構造類似体を検索します

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

ユーザーがワイルドカードで指定した比較対象タンパク質ファイルのパスに基づいて、クエリタンパク質の構造類似体を検索します

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

この例では、"set1/*.pdb" と "set2/1a???.cif" は引用符で囲まれ、ワイルドカードが含まれています。sarst2 プログラムは、ワイルドカード文字に一致するファイル名を自動的に取得します。引用符がない場合、オペレーティングシステムは、ワイルドカード文字に一致するすべてのファイル名を自動的にリストします。ファイルが多すぎると、コマンドラインパラメーターが長すぎて機能しなくなる可能性があります。オペレーティングシステムのデフォルトのファイルリスト動作に依存するのではなく、引用符を使用すること (つまり、sarst2 プログラム自体にファイルをリストさせること) をお勧めします。

ユーザーが指定した比較対象タンパク質フォルダーの中から、クエリタンパク質の構造類似体を検索します

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

この例では、set1 と set2 はタンパク質構造ファイルを含む可能性のあるフォルダーです。両方のフォルダー (set1/* と set2/*) 内のすべてのファイルのリストは、sarst2 プログラム自体によって取得されます。PDB/CIF 形式の構造ファイルが見つかった場合、これらの構造はクエリの構造と比較されます。

## 5.5コマンド例：構造のペアワイズアライメント

クエリタンパク質の構造を別のタンパク質とアライメントします

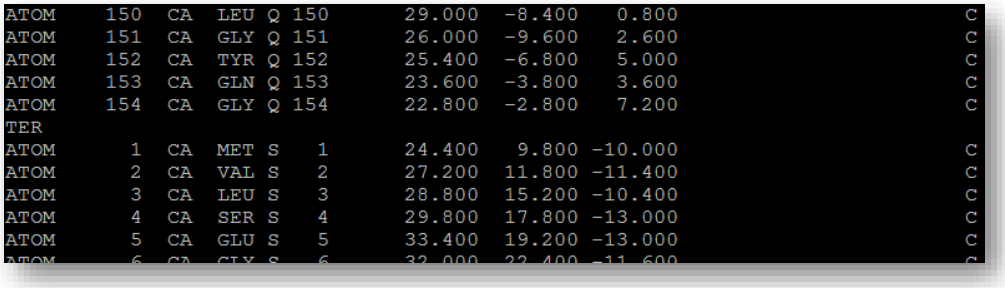
```
./sarst2 Qry.pdb Sbj.pdb -sa F
./sarst2 Qry.pdb Sbj.cif -mat T
```

## 5.6 タンパク質構造重ね合わせファイルの出力

クエリタンパク質と各比較対象タンパク質の構造重ね合わせファイルを出力します

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

"-Sout Output_folder" オプションを使用すると、タンパク質構造重ね合わせファイル (PDB 形式) をユーザーが指定したフォルダーに出力できます。出力されるファイルの数は、-detail パラメーターによって定義されます。このフォルダー内では、各構造重ね合わせファイルは Qry-SbjSN.pdb という形式で命名されます。SN は結果リストにおける比較対象タンパク質の配列番号です。構造重ね合わせファイルでは、クエリと比較対象タンパク質構造のポリペプチド鎖コードはそれぞれ Q と S となり、以下の例に示すように、2 つの鎖は TER レコードによって区切られます。



```
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
```

上の図は、重ね合わせ構造ファイルにはアルファ炭素（Cα）原子のみが表示されることも示しています。これは、SARST2 が計算プロセスで Cα 座標のみを使用するためです。クエリタンパク質の空間内での配置方法は、すべての重ね合わせ構造ファイルで固定されています。各検査対象タンパク質は、クエリタンパク質との比較結果に基づいて、座標の回転・平行移動を行った後にクエリタンパク質の構造と重ね合わされます。

重ね合わせ構造を視覚化するには、RasMol および RasWin (http://www.openrasmol.org/) を推奨します。重ね合わせファイルには Cα 原子しかないため、RasMol の表示モードを「backbone」に設定する必要があります。

## 5.7 HTML インタラクティブ結果表示ウェブページの出力

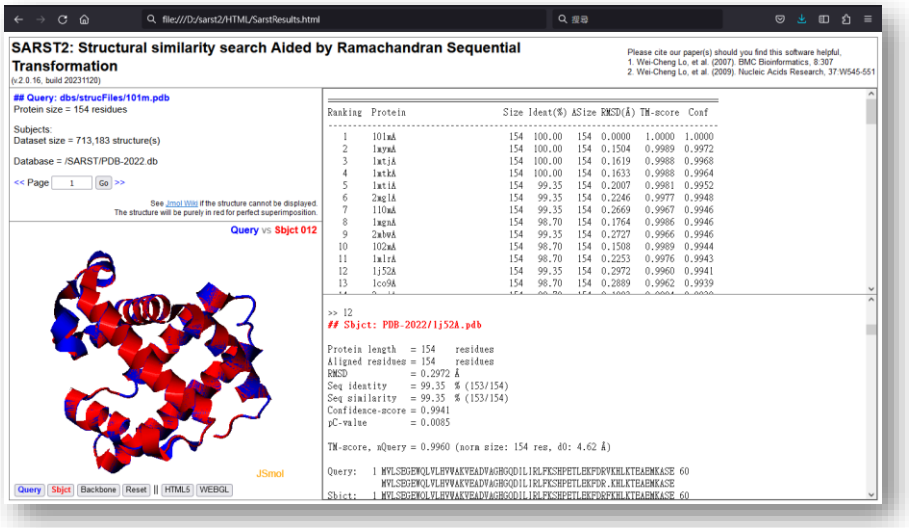
オンライン JSmol 動的構造表示機能付き HTML 結果ページの出力

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

ローカルディスク JSmol 動的構造表示機能付き HTML 結果ページの出力 (Windows)

```
./sarst2 Qry.pdb "set1/*.cif" -brief 10 -detail 100 -html
output_folder -jsmol "file:///D:/bioinfo/software/jsmol"
```

"-html Output_folder" パラメーターを使用すると、HTML 結果表示ドキュメントとタンパク質構造重ね合わせファイルが指定されたフォルダーに出力されます。この -html オプションは、JSmol パッケージ（2013 年版）の URL を指定する -jsmol オプションと併用する必要があります。JSmol は、ウェブブラウザで動作するインタラクティブな立体分子構造ビューアで、主要なウェブブラウザのほとんどをサポートしています。




HTML 出力フォルダー内のメインファイルは "SarstResults.html" です。これをウェブブラウザで開いてください。他の HTML ファイルは、インナーフレーム形式でウェブページに統合されています。また、"sup" という名前のサブフォルダーがあり、そこにはクエリタンパク質と、それに構造が類似する各比較対象タンパク質の構造重ね合わせファイルが保存されています。

JSmol の立体構造表示オブジェクトが正常に動作するためには、ウェブブラウザが JavaScript を実行し、"sup" ディレクトリ内の蛋白質構造重ね合わせファイルを読み込めるよう、オペレーティングシステム、ブラウザ、またはアンチウイルスソフトウェアのセキュリティ設定を調整する必要がある場合があります。

# 6. マニュアル：formatdb (データベース作成プログラム)

## 6.1使い方

```
./formatdb      検索対象      -db データベース  [オプション]
                --------
                > できるのは
                1．複数の PDB/CIF ファイル
                2．PDB/CIF ファイルのフォルダー
                3．PDB/CIF ファイルパスを記述したテキストファイル
```

## 6.2プログラムのオプション

| | | |
|---|---|---|
| -db | [文字列] | 作成する対象データベース。複数の比較対象タンパク質構造が含まれています。<br>(既定値なし) |
| -flist | [文字列] | 蛋白質構造ファイルのリストが記述された純粋なテキストファイルです。このパラメーターは、他のファイルパスパラメーターと組み合わせて使用できます。<br>(既定値なし) |
| -t | [整数] | 使用するスレッドの数。<br>0 以上を指定してください。0 の場合、このコンピュータ全プロセッサコア数が自動的に設定されます。<br>(既定値 0、プロセッサコア数全て) |
| -split | [整数] | データベースを、このオプションで指定された比較対象タンパク質の最大数を含むサブセットに分割します。その分割により、データベースファイルがディスクのファイルサイズ制限を超えるのを防ぐことができます。<br>(既定値なし) |
| -save_disk | [T/F] | ディスク領域を節約するため、原子座標を小数点以下 3 桁から小数点以下 1 桁に四捨五入します。<br>(既定値 F) |
| -keep_order | [T/F] | データベースに保存される比較対象構造の順序を、入力された順序と同じになるように維持します。これを T に設定すると、データベースの作成が遅くなります。<br>(既定値 F) |
| -h | | クイックガイドを表示されます。 |

## 6.3コマンド例

ユーザーが列挙した比較対象タンパク質構造ファイルからターゲットデータベースを作成します

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

作成が完了すると、myDb で始まるファイル名を持つ複数のデータベースファイルがディスク上に生成されます。-keep_order を有効にすると、ターゲットデータベースに格納される比較対象タンパク質の順序が、コマンドライン引数で指定された順序と同じになります。

ユーザーがワイルドカードで指定した比較対象タンパク質構造ファイルを検索し、ターゲットデータベースを作成します

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

比較対象タンパク質の構造ファイルをリストする際、ワイルドカードを含むパラメーターと含まないパラメーターを混在させることができます。ワイルドカードを含むパラメーターは引用符で囲むことを推奨します。

ファイルリストで指定されたタンパク質構造から、ターゲットデータベースを作成します

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

ファイルリスト protlist.txt の形式は、1 行につき 1 つのファイルです。

ユーザーが列挙した比較対象タンパク質フォルダー内からすべての構造ファイルを検索して、ターゲットデータベースを作成します

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

-save_disk を有効にすると、ストレージ容量を節約するため、Cα 座標が小数点第 1 位に四捨五入されます。また、"-split 50000" を設定すると、複数のデータベースサブセットが生成され、各サブセットには最大 50,000 個の構造が含まれます。この分割機能は、フォーマットされたデータベースファイルのサイズが、特定のオペレーティングシステムやディスクフォーマットの上限を超える可能性がある場合に特に役立ちます。

# 7. マニュアル：readdb (データベース配列読み込みプログラム)

## 7.1 使い方

```
./readdb  ターゲットデータベース      出力ファイル      [-seq 配列タイプ]
          ------------------------    ----------      ----------------
          > フォーマット済み          > FASTA 形式    > できるのは
            SARST2 形式                               1. AA
                                                      2. AAT
                                                      3. SARST
                                                      4. SSE
```

## 7.2 プログラムのオプション

| | | |
|---|---|---|
| -seq | [文字列] | 出力配列のタイプ。<br>AA　　アミノ酸配列<br>AAT　　５つのシンボル構成するアミノ酸タイプ配列<br>SARST　SARST 構造配列<br>SSE　　４つのシンボル構成する二次構造配列<br>(既定値 AA) |
| -h | | クイックガイドを表示されます。 |

## 7.3 コマンド例

SARST2 ターゲットデータベースから、必要なタイプの配列を読み出します

```
./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
```

-seq オプションを指定しない場合、既定値でアミノ酸配列が出力されます。出力される配列ファイル (seqs.fasta) は FASTA 形式です。readdb の実行前に出力ファイルが既に存在する場合、そのファイルは上書きされます。