Protein Structural Alignment

Structural alignment attempts to establish homology between two or more polymer structures based on their shape and three-dimensional conformation. This process is usually applied to protein tertiary structures but can also be used for large RNA molecules. In contrast to simple structural superposition, where at least some equivalent residues of the two structures are known, structural alignment requires no *a priori* knowledge of equivalent positions.

Structural alignment of thioredoxins from humans and the fly Drosophila melanogaster. The proteins are shown as ribbons, with the human protein in red, and the fly protein in yellow. Generated from PDB 3TRX and 1XWC.

The minimum information produced from a successful structural alignment is a set of superposed three-dimensional coordinates for each input structure. (Note that one input element may be fixed as a reference and therefore its superposed coordinates do not change.) The fitted structures can be used to calculate mutual RMSD values, as well as other more sophisticated measures of structural similarity .

The structural alignment also implies a corresponding one-dimensional sequence alignment from which a sequence identity, or the percentage of residues that are identical between the input structures, can be calculated as a measure of how closely the two sequences are related.

Structure Alignment Results

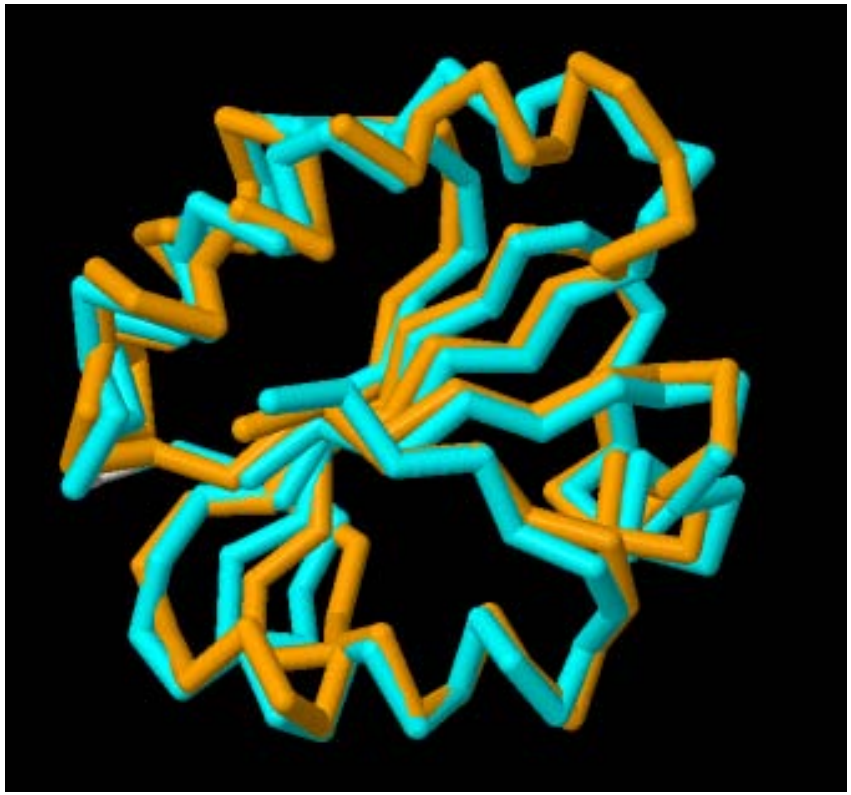
Alignment Details: **Query:** (orange/dark grey) ***THIOREDOXIN***

Z-score: 6.23
Score: 173.83
RMSD: 1.51
%Id: 48.1%

PDB ID: 3TRX
Chain ID: A
Length: 105
Similarity: 100%
EC number:

Subject: (cyan/light grey) ***thioredoxin***

PDB ID: 1XWC
Chain ID: A
Length: 106
Similarity: 99%
EC number:



The standard score of a raw score x [1] is

$$z = \frac{x - \mu}{\sigma}$$

where:

μ is the mean of the population;

σ is the standard deviation of the population.

Root-mean-square deviation of atomic positions

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

where δ is the distance between N pairs of equivalent atoms (usually $C\alpha$ and sometimes $C, N, O, C\beta$).

Normally a rigid superposition which minimizes the RMSD is performed, and this minimum is returned. Given two sets of n points $\mathbf{v}$ and $\mathbf{w}$, the RMSD is defined as follows:

$$\mathrm{RMSD}(\mathbf{v},\mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix}-w_{ix})^2 + (v_{iy}-w_{iy})^2 + (v_{iz}-w_{iz})^2)}$$

An RMSD value is expressed in length units. The most commonly used unit in structural biology is the Ångström (Å) which is equal to 10^{-10}m.

Alignment Block(s) | Alignment with Sequence Conservation

```
Align 3TRX.A.pdb Length1: 105 with 1XWC.A.pdb Length2: 106
Z-score 6.23
Equ: 105
RMSD: 1.51
Score: 173.83
Align-len: 106
Gaps: 1 (0.94%)

Identity: 48.11%
Similarity: 64.15%

1:A       10:A      20:A      30:A      40:A      50:A      60:A
|         |    .    |    .    |    .    |    .    |    .    |    .
MVKQIESKTAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSEKYSN-VIFLEVDVDDCQDVASEC
||.|:..|......|..|..||||.||.||||||||||.|....||..:.. |:.|.||||:|:|:|.|.
MVYQVKDKADLDGQLTKASGKLVVLDFFATWCGPCKMISPKLVELSTQFADNVVVLKVDVDECEDIAMEY
|         |    .    |    .    |    .    |    .    |    .    |    .    |
1:A       10:A      20:A      30:A      40:A      50:A      60:A      70:A


 70:A      80:A      90:A      100:A
|    .    |    .    |    .    |    .
EVKCTPTFQFFKKGQKVGEFSGANKEKLEATINELV
.:...|||.|.|.|.||.||.|||..:||..|...:
NISSMPTFVFLKNGVKVEEFAGANAKRLEDVIKANI
    .    |    .    |    .    |    .
         80:A      90:A      100:A



     | ... Structurally equivalent and identical residues
     : ... Structurally equivalent and similar residues
     . ... Structurally equivalent, but not similar residues.

     To calculate the coordinates of chain 2 aligned on chain 1 apply the following transformation:

     X1 = (-0.185192)*Xorig + (-0.622374)*Yorig + (-0.760496)*Zorig + (   47.503000)
     Y1 = ( 0.386924)*Xorig + ( 0.665191)*Yorig + (-0.638600)*Zorig + (    9.933000)
     Z1 = ( 0.903324)*Xorig + (-0.412518)*Yorig + ( 0.117624)*Zorig + (   20.435000)
```

Sequence alignment derived from structural alignment

I = 48.11%

Calculate
- Two Chains
- Symmetry
- DB Search
- All vs All

Download
- Download jCE/jFatCat
- Documentation

History
- Release History
- CE History

Links
- Multiple Structure Alignment

Results
- Subdomains (PDF)

CruiseControl
- CruiseControl

Combinatorial Extension (CE)

A method for comparing and aligning protein structures

This page is intended as a pointer to get you to the most recent information on CE and to enable you to perform the calculations you need. CE is now an integral part of the RCSB Protein Data Bank (PDB) and continues to be developed in the Bourne laboratory as needed.

Key Pointers

- Access to CE from the RCSB PDB http://www.rcsb.org/pdb/workbench/workbench.do
- Standalone server http://source.rcsb.org/jfatcatserver/
- Access to the CE code in Java (jCE) and the original source http://source.rcsb.org/jfatcatserver/download.jsp
- Legacy CE - web site

What follows is a brief description of the history of CE and some additional references and pointers.

Chronology

- **1998** - CE method released and original paper published [1]
- **2000** - CE used to map existing protein fold space [2]
- **2001** - Pairwise alignment database made available [3]
- **2004** - A parallel version of CE was developed [4] (no longer relevant)
- **2004** - A multi-structure version of CE was released CE-MC [5]
- **2005** - A benchmark dataset of hand alignments was computed and run against CE [6]
- **2010** - Precalculated CE alignments and a pairwise alignment server made available from the RCSB PDB [7]

RCSB PDB PROTEIN DATA BANK

jCE/jFatCat Help
Explanation of frequently used terms.

Help

Calculate
- Two Chains
- Symmetry
- DB Search
- All vs All

Download
- Download jCE/jFatCat
- Documentation

History
- Release History
- CE History

Links
- Multiple Structure Alignment

Results
- Subdomains (PDF)

CruiseControl
- CruiseControl

Help

About

CE is a method for calculating **pairwise structure alignments**. CE aligns two polypeptide chains using characteristics of their local geometry as defined by vectors between C alpha positions. Matches are termed aligned fragment pairs (AFPs). Heuristics are used in defining a set of optimal paths joining AFPs with gaps as needed. The path with the best RMSD is subject to dynamic programming to achieve an optimal alignment. For specific families of proteins additional characteristics are used to weight the alignment. Complete details are described in the paper (PDF format).

jCE

jCE is a re-implementation of the original CE source code in the Java programming language. While the algorithm is principle exactly the same as in the original implementation, jCE provides several improvements over the original code:

- **New internal data model:** The internal representation of jCE is now based on the BioJava protein structure module. It supports both PDB file and mmCif files as input.
- **User Interface:** jCE provides a completely new user interface for easier set up of pairwise alignments and database searches.
- This 3D alignment program is based on BioJava and Jmol.

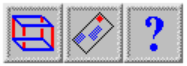
Structural alignment software

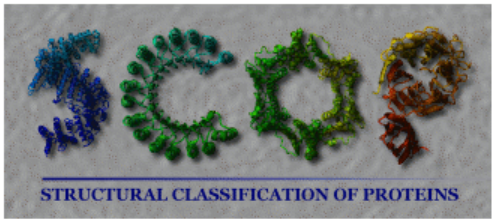
From Wikipedia, the free encyclopedia

This **list of structural comparison and alignment software** is a compilation of software tools and web portals used in pairwise or multiple structural comparison and structural alignment.

Structural comparison and alignment [edit]

NAME	Description	Class	Type	Flexible	Link	Author	Year
MAMMOTH	**MA**tching **M**olecular **M**odels **O**btained from **T**heory	Cα	Pair	No	server download	CEM Strauss & AR Ortiz	2002
CE	**C**ombinatorial **E**xtension	Cα	Pair	No	server	I. Shindyalov	2000
CE-MC	**C**ombinatorial **E**xtension-**M**onte **C**arlo	Cα	Multi	No	server	C. Guda	2004
DaliLite	**D**istance Matrix **Ali**gnment	C-Map	Pair	No	server	L. Holm	1993
TM-align	**TM**-score based protein structure **align**ment	Cα	Pair	nil	server and download	Y. Zhang & J. Skolnick	2005
VAST	**V**ector **A**lignment **S**earch **T**ool	SSE	Pair	nil	server	S. Bryant	1996
PrISM	**P**rotein **I**nformatics **S**ystems for **M**odeling	SSE	Multi	nil	server	B. Honig	2000
SSAP	**S**equential **S**tructure **A**lignment **P**rogram	SSE	Multi	No	server	C. Orengo & W. Taylor	1989
SARF2	**S**patial **AR**rangements of Backbone **F**ragments	SSE	Pair	nil	server	N. Alexandrov	1996
KENOBI/K2	NA	SSE	Pair	nil	server	Z. Weng	2000
STAMP	**ST**ructural **A**lignment of **M**ultiple **P**roteins	Cα	Multi	No	site server	R. Russell & G. Barton	1992
MASS	**M**ultiple **A**lignment by **S**econdary **S**tructure	SSE	Multi	No	server	O. Dror & H. Wolfson	2003
SCALI	**S**tructural **C**ore **ALI**gnment of proteins	Seq/C-Map	Pair	nil	server download	X. Yuan & C. Bystroff	2004
DEJAVU	NA	SSE	Pair	nil	server	GJ. Kleywegt	1997
SSM	**S**econdary **S**tructure **M**atching	SSE	Multi	nil	server	E. Krissinel	2003
SHEBA	**S**tructural **H**omology by **E**nvironment-**B**ased **A**lignment	Seq	Pair	nil	server	B. Lee	2000
LGA	**L**ocal-**G**lobal **A**lignment	Cα	Pair	nil	server	A. Zemla	2003
POSA	**P**artial **O**rder **S**tructure **A**lignment	Cα	Multi	Yes	server	Y. Ye & A. Godzik	2005
PyMOL	"super" command does sequence-independent 3D alignment	Protein	Hybrid	No	site	W. L. DeLano	2007
FATCAT	**F**lexible **S**tructure **A**lignmen**T** by **C**haining **A**ligned **F**ragment Pairs **A**llowing **T**wists	Cα	Pair	Yes	server	Y. Ye & A. Godzik	2003
deconSTRUCT	Database search on substructural level and pairwise alignment.	SSE	Multi	No	server	ZH. Zhang et al.	2010
Matras	**MA**rkovian **TRA**nsition of protein **S**tructure	Cα & SSE	Pair	nil	server	K. Nishikawa	2000





Welcome to SCOP: Structural Classification of Proteins.
1.75 release (June 2009)

38221 PDB Entries. 1 Literature Reference. 110800 Domains. (excluding nucleic acids and theoretical models).
Folds, superfamilies, and families statistics here.
New folds superfamilies families.
List of obsolete entries and their replacements.

Authors. Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mrc-lmb.cam.ac.uk

Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]
Recent changes are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [PDF],
Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [PDF], and
Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acid Res.* advance access, doi:10.1093/nar/gkm993. [PDF].

Access methods

- **Enter SCOP at the top of the hierarchy**
- Keyword search of SCOP entries
- SCOP parseable files (MRC site)
- All SCOP releases and reclassified entry history (MRC site)
- **pre-SCOP - preview of the next release**
- SCOP domain sequences and pdb-style coordinate files (ASTRAL)
- Hidden Markov Model library for SCOP superfamilies (SUPERFAMILY)
- Structural alignments for proteins with non-trivial relationships (SISYPHUS)

- Online resources of potential interest to SCOP users

SCOP **mirrors** around the world may speed your access.

SCOP: Protein Classification

Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy, but the principal levels are **family, superfamily and fold.**

The exact position of boundaries between these levels are to some degree subjective. Our evolutionary classification is generally conservative: where any doubt about relatedness exists, we made new divisions at the family and superfamily levels.

Thus, some researchers may prefer to focus on the higher levels of the classification tree, where proteins with structural similarity are clustered.

The different major levels in the hierarchy are:

Family: ***Clear evolutionarily relationship***

Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. However, in some cases similar functions and structures provide definitive evidence of common descent in the absense of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%.

Superfamily: ***Probable common evolutionary origin***

Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies. For example, actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily.

Fold: ***Major structural similarity***

Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

SCOP: Structural Classification of Proteins. **1.75** release
38221 PDB Entries (23 Feb 2009). 110800 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	284	507	871
All beta proteins	174	354	742
Alpha and beta proteins (a/b)	147	244	803
Alpha and beta proteins (a+b)	376	552	1055
Multi-domain proteins	66	66	89
Membrane and cell surface proteins	58	110	123
Small proteins	90	129	219
Total	1195	1962	3902