

Dali: an algorithm for optimal structure alignment using distance matrices

“Protein Structure Comparison by alignment of distance matrices” by Liisa Holm & Chris Sander

Presentation by: Athina Ropodi

Outline

- Introduction
- Dali algorithm
- Results
- Output
- DaliLite

Introduction

- Most newly determined protein sequences can be classified into families by sequence homology.
- However, protein families are known to retain the shape of the fold even when sequences share few similarities at the sequence level.
- These similarities can be detected by structural comparisons that merge protein families of known 3-D structure into structural classes.

Introduction

- A significant tool for the comparison of protein structures is the distance matrix.
- It is a 2D representation of the 3D structure, as it contains all pairwise distances between atoms – in this case C α atoms.
- They can be obtained by X-ray crystallography and Nuclear Magnetic Resonance (NMR).

Outline

- Introduction
- Dali algorithm
- Results
- Output
- DaliLite

Dali algorithm

- The “Protein Structure Comparison by alignment of distance matrices” by Liisa Holm & Chris Sander was first published in 1993.
- It was implemented in Fortran-77
- The name stands for Distance-matrix ALIGNment.

Dali algorithm – Methods

- Given proteins A and B, we define:

$$S = \sum_{i=1}^L \sum_{j=1}^L \phi(i, j)$$

- where i, j label pairs of matched residues, L is the number of these pairs, and ϕ the similarity measure based on the distances d_{ij}^A, d_{ij}^B

Dali algorithm – Similarity scores

- Objective: the search for the largest common substructure between 2 proteins

- Rigid similarity score:

$$\phi^R(i, j) = \theta^R - |d_{ij}^A - d_{ij}^B|,$$

where θ is the similarity threshold

- Elastic similarity score:

$$\phi^E(i, j) = \begin{cases} \left(\theta^E - \frac{|d_{ij}^A - d_{ij}^B|}{d_{ij}^*} \right) w(d_{ij}^*), & i \neq j \\ \theta^E, & i = j \end{cases}$$

where d_{ij}^* the average of the distances, and w is an envelope function.

Since pairs in the long distance are abundant their contribution is weighted down by the envelope function:
 $w(r) = \exp(-r^2 / a^2)$
where $a=20\text{\AA}$ calibrated on the size of a typical domain

Dali: a greedy algorithm

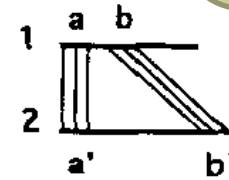
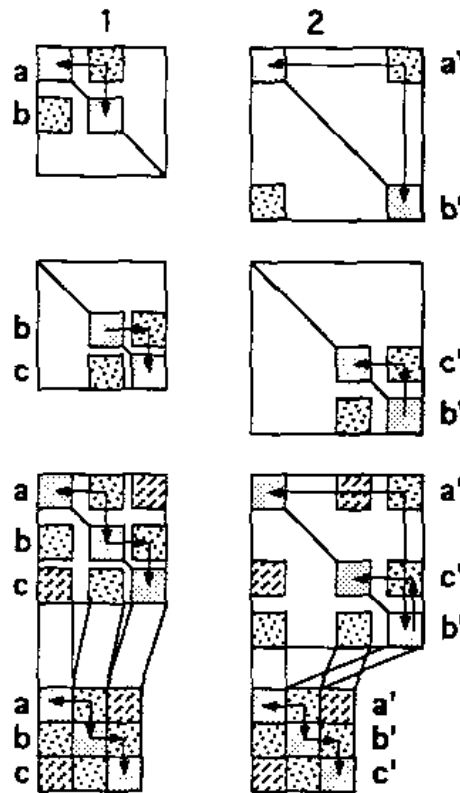
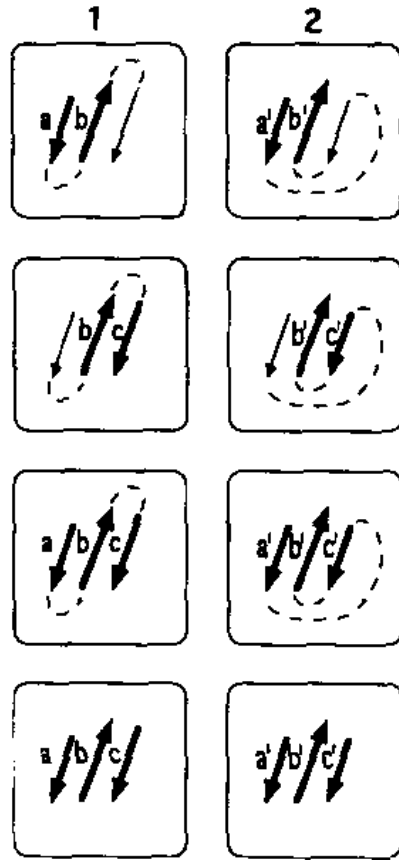
- It consists of 2 basic steps:
 1. Performs systematic comparisons of all elementary patterns (here hexapeptides). Similar patterns are stored in the “pair list”
 2. A Monte Carlo algorithm is used to deal with the combinatorial complexity of assembling patterns into larger consistent sets of pairs.

Dali: 1st step

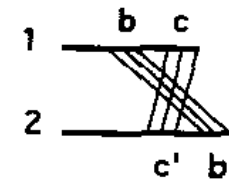
- In the first step of the algorithm, similar submatrices of size six in two proteins are found by comparing their distance matrices.
- These comparisons result in alignments of size six between two proteins. Then, compatible alignments are merged to obtain larger alignments called *seeds*.

Example of merging of co-alignments

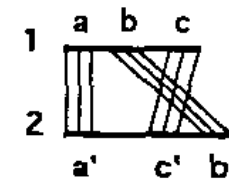
Upon comparison of distance matrices of proteins *A* and *B*, matrix component (a, b) is aligned to (a', b')



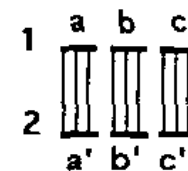
one pair



an overlapping pair

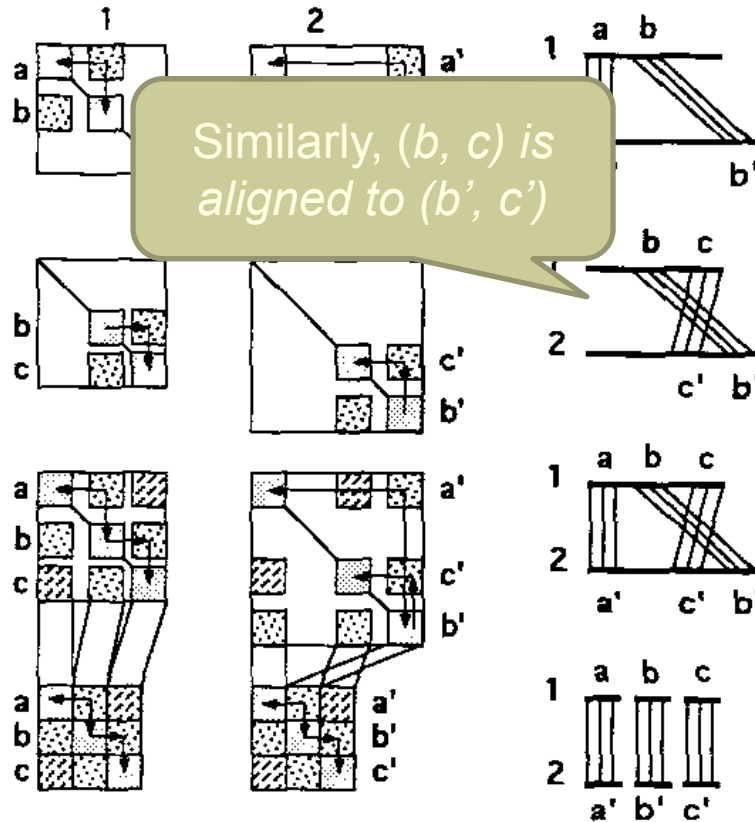
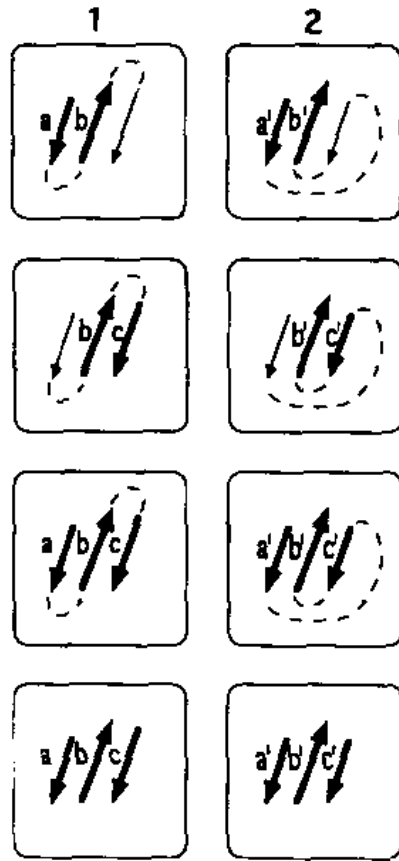


the two pairs combined



collapse

Example of merging of compatible alignments



Similarly, (b, c) is aligned to (b', c')

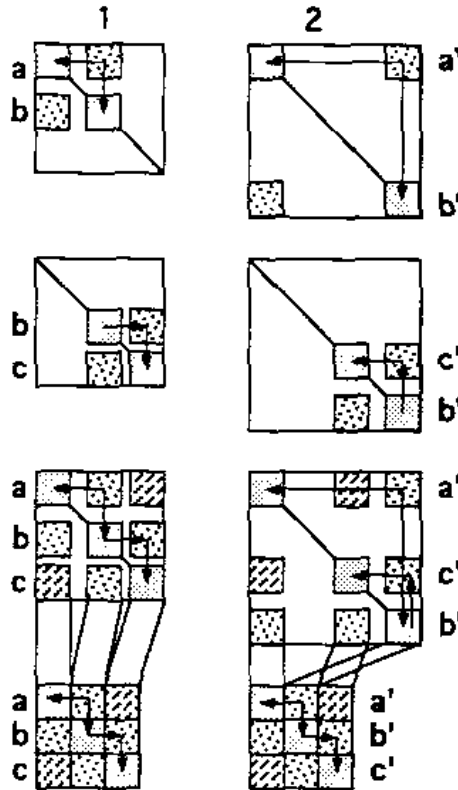
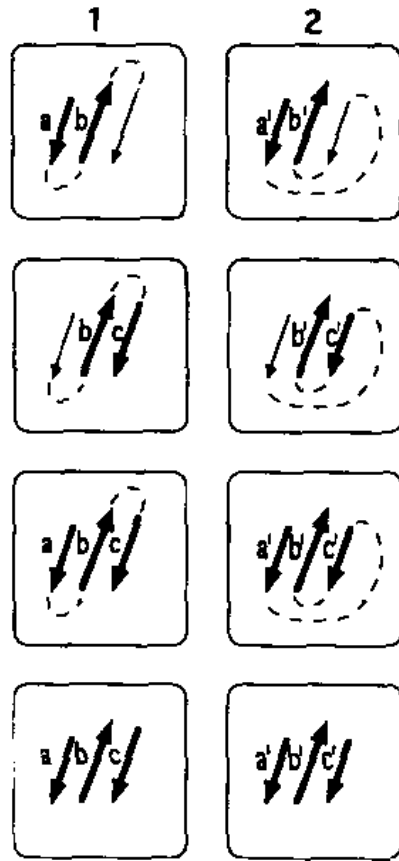
one pair

an overlapping pair

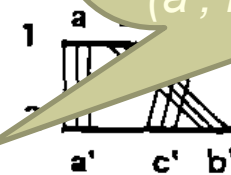
the two pairs combined

collapse

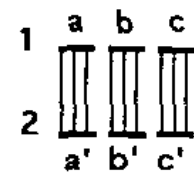
Example of merging of compatible alignments



Since the two alignments are overlapping, *they are checked for compatibility.* If the nine matrix components for these two alignments are found similar to each other, alignments are merged to obtain a seed of $(a, b, c) - (a', b', c')$.



the two pairs combined



collapse

Dali: 1st step

A. Distance matrices

1lyz

No. of overlapping hexapeptides	124
Total no. of contact patterns	7626
No. of contact patterns in reduced distance matrix	5332

2lzm

No. of overlapping hexapeptides	159
Total no. of contact patterns	12,561
No. of contact patterns in reduced distance matrix	4709

B. Pair list

Total no. of pairs of contact patterns	96×10^6
Total no. of pairs of contact patterns after reduction	71×10^6
No. of checks by filters on row/column sums†	9×10^6
No. of residue-by-residue similarity score calculations	2×10^5
No. of kept pairs of contact patterns after ranking by score	4×10^4

Dali: Monte Carlo optimization

- The key idea is iteratively improving by a random exploration of the search space with occasional moves into non-optimal territory.
- A move is a randomly chosen change. The probability of accepting a move is:

$$p = \exp(\beta * (S' - S))$$

where S' , S the new and old scores, and β a parameter, inversely proportional to the temperature.

- Moves that improve the score are always accepted, but the higher the temperature the more probable are excursions downhill.

Dali: Monte Carlo optimization

- The basic moves are addition and deletion.
- A chain of configurations is called trajectory. For every trajectory the highest score is remembered.
- Optimization starts with a seed alignment.
- One expansion cycle corresponds to randomly testing all expansion candidates in the pair list.
- The addition of a new fragment may require the removal of inconsistent previous assignments.

Dali: selection protocol

- The range of alignments is narrowed onto the highest scoring ones in 3 stages.
- In stage 1, the pair list is screened for all triplets of non-overlapping hexapeptides.

A trimming cycle is performed after the first and every 5 consecutive expansion cycles

Dali: selection protocol (stage1)

- Singlets that overlap are merge into 1 seed.
- The maximum number of seeds is 100.
- Each seed initializes a trajectory and goes through one cycle.
- If the assignments of 2 trajectories are more than 50% identical, then the one with the lowest score is rejected.
- In practice, keeping the 10 highest scoring trajectories gives good results.

Dali: selection protocol (stage 2)

- Optimization is continued until the score has settled in an optimum, i.e. for 20 cycles.
- We reject trajectories with more than 80% identity with higher scoring ones.
- Trajectories with a score smaller than a fraction of the best score are also rejected.

Dali: selection protocol (stage 3)

- Refinement of the best scoring trajectory.
- The best alignment is used to initialize 10 parallel trajectories with 30% of aligned blocks removed.
- These are optimized as in stage 2, until the best score no longer improves.

Dali: Monte Carlo optimization

C. Monte Carlo optimization

Screening

No. of parallel trajectories	80
No. of expansion/trimming cycles‡	1
No. of kept alignments after ranking by score	10

Optimization of divergent alignments

No. of parallel trajectories	10
No. of expansion/trimming cycles‡	80
No. of kept alignments after ranking by score	1

Refinement of best alignment

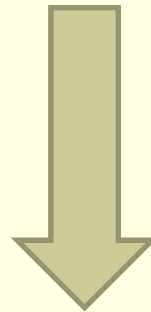
No. of parallel trajectories	10
No. of expansion/trimming cycles‡	40
No. of kept alignments after ranking by score	1

Outline

- Introduction
- Dali algorithm
- **Results**
- Output
- DaliLite

Results

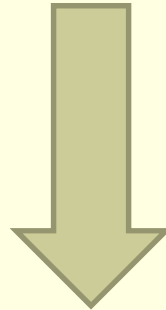
- In order to test whether the highest score found is the global optimum, comparisons with random number of seeds were performed.



- The algorithm converged to within 2% of the global optimum score with 96% fidelity.

Results

- To test the radius of convergence, alignments were generated from all seeds (even from the ones that would have been rejected).
- The vast majority corresponded to incorrect optima.



- Thus, the optimization procedure is not overly sensitive to the choice of initial alignment.

Outline

- Introduction
- Dali algorithm
- Results
- Output
- DaliLite

Output- structural alignment

- Obviously, an elastic score gives more common core residues than a rigid one.

```

=====
....._HHHHHHHHH_.._EEEEEE_TTS_EEEE.TTEEEES
.....MNIFEMLRIDEG..LRLKIYKDTTEGYTTIG.....SPLNAAKSELDKAIGRNCNGVITKDEAEKLFNQDVAAVRGI LRNAKLPVYD 2LZM
.....210353454344..21513451.0145454.345542.....5230.....-252.....1566..553455224431443.....
KVFGRaELAAAI.....QATNRNTDGGSTDYGILOINSRWwCNDGRTPGSRNL...dNIPcSAL...LSSD..ITASVNDAKKIVSDG..... 1LYZ
_B_HHHHHHHH.....S_EEE_TTS_EEETTTTEETTTS_B_SS_TT__T...T_SBGGGG...GSS..._HHHHHHHHHTTSS.....
=====
HS_HHHHHHHH.....HHHHHSSHHHHHSHHHHHHHHHHHHHSSHHHHH.....
SLDAVRRCALINMVFQMGETGVAGFTNSLRMLQQRWDEAAVNLAKSRYWYNQTPNRAKRVITTFRTGTWDAYKNL..... 2LZM
.....0235433.....01222.....32342.....
.....NGMNAWVAWRN.....RbKGT.....DVQAWIRGaRL 1LYZ
.....SGGGGSHHHHH.....HTTTS....._GCGGSTT_
=====

```

Common core of residues

The method is flexible in that gaps may be of any length.

Active sites

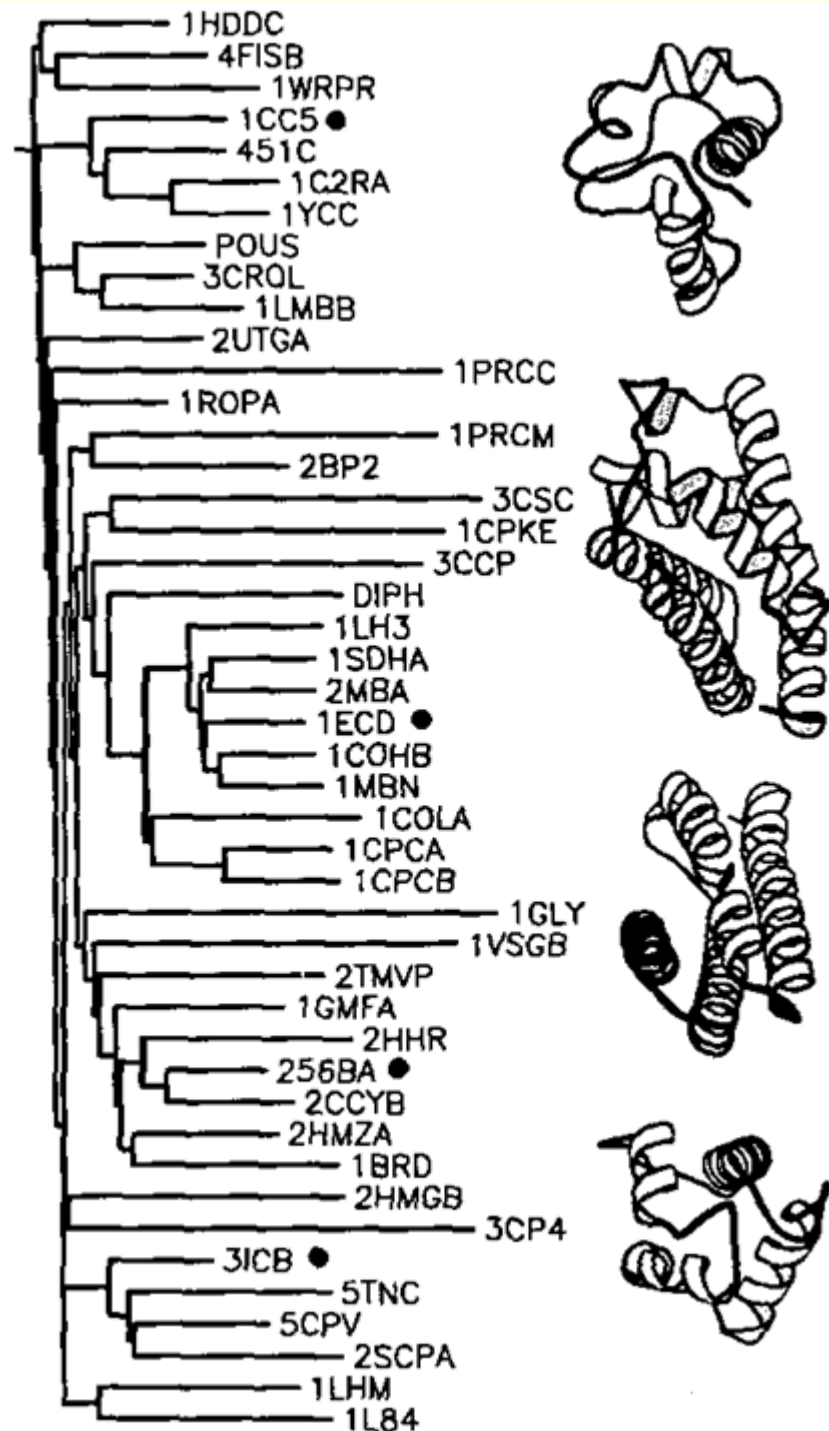
Gaps, non-aligned segments & trailing ends

H, a-helix;
 G, 3-helix;
 E & B, β -sheet;
 T, turn;
 S, bend;
 -, non-hydrogen bonded structure.

Protein structure family trees by average linkage clustering

α-helical proteins:

1HDDC	<i>engrailed</i> homeodomain	1COHB	hemoglobin
4FISB	factor for inversion stimulation	1MBN	myoglobin
1WRPR	Trp repressor	1COLA	colicin A
1CC5	cytochrome c5	1CPCA	C-phycoyanin
451C	cytochrome c551	1CPCB	C-phycoyanin
1C2RA	cytochrome c2	1GLY	glucoamylase
1YCC	cytochrome c	1VSGB	variant surface glycoprotein
POUS	POU-specific domain	2TMVP	tobacco mosaic virus
3CROL	434cro	1GMFA	growth factor
1LMBB	lambda repressor	2HHRA	human growth hormone
2UTGA	uteroglobin	256BA	cytochrome b562
1PRCC	photosynthetic reaction centre	2CCYB	cytochrome c'
1ROPA	ROP (repressor of primer) protein	2HMZA	hemerythrin
1PRCM	photosynthetic reaction centre	1BRD	bacteriorhodopsin
2BP2	phospholipase	2HMGB	hemagglutinin
3CSC	citrate synthase	3CP4	cytochrome P450 CAM
1CPKE	cAMP-dependent protein kinase	3ICB	intestinal calcium-binding protein
3CCP	cytochrome c peroxidase	5TNC	troponin C
DIPH	diphtheria toxin	5CPV	parvalbumin B
1LH3	leghemoglobin	2SCPA	sarcoplasmic calcium-binding protein
1SDHA	hemoglobin	2LHM	human lysozyme
2MBA	myoglobin	1L84	T4 lysozyme
1ECD	erythrocrucorin		



Outline

- Introduction
- Dali algorithm
- Results
- Output
- DaliLite


DaliLite workbench



- The Dali server is routinely used
 1. to compare newly solved structures against those in PDB,
 2. to compare predicted structures to the real ones and
 3. to maintain the FSSP database of structural networks.
- It is available on the internet
- There is now a stand-alone distribution which contains programs written in Perl and Fortran77

DaliLite workbench-Characteristics

- Two alignment options –pairwise comparison and database search.
- The input is one or two sets of atomic coordinates of proteins in PDB format.
- The output is a FSSP file
- A visualization script is included to convert FSSP alignments to graphical output.

http://www.ebi.ac.uk/DaliLite/

EMBL-EBI  All Databases

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index  

- Help Index
- General Help
- Formats
- References
- DaliLite Help

- DaliLite Programmatic Access

EBI > Tools > Structural Analysis


DaliLite Pairwise comparison of protein structures

DaliLite is a program for pairwise structure comparison. Compare your structure(first structure) to a reference structure(second structure).

	<u>RESULTS</u>	<u>SEARCH TITLE</u>	<u>YOUR EMAIL</u>
	<input type="text" value="interactive"/>	<input type="text" value="Sequence"/>	<input type="text"/>
First Structure	<u>PDB entry code:</u> <input type="text"/>	Chain ID: <input type="text"/>	or upload a file in PDB format (.pdb,.ent,.dat,.brk) <input type="text"/> <input type="button" value="Αναζήτηση..."/>
Second Structure	<u>PDB entry code:</u> <input type="text"/>	Chain ID: <input type="text"/>	or upload a file in PDB format (.pdb,.ent,.dat,.brk) <input type="text"/> <input type="button" value="Αναζήτηση..."/>

http://ekhidna.biocenter.helsinki.fi/ dali_server/

Dali server



SERVICES & TOOLS	GROUP MEMBERS	NEWS & VACANCIES	RESEARCH	PUBLICATIONS
------------------	---------------	------------------	----------	--------------

Protein Structure Database Searching by DaliLite v. 3

The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank (PDB). You receive an email notification when the search has finished. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

If you want to know the structural neighbours of a protein already in the Protein Data Bank (PDB), you can find them in the [Dali Database](#).

If you want to compare two particular structures, you can do it in the [pairwise DaliLite](#) server.

Upload a structure:

Or enter PDB identifier: chain: (optional)

Enter email address for notification:

Repeat email address:

PDB search results for epidermal growth factor legf

DaliLite: Structural Neighbours

Query: mol1A MOLECULE: EPIDERMAL GROWTH FACTOR;

Matches are sorted by Z-score. Similarities with a Z-score lower than 2 are spurious.

Summary

No:	Chain	Z	rmsd	lali	nres	%id	Description
1:	legf	12.7	0.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES)
2:	3egf	10.6	1.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES AFTEF
3:	1mox-D	4.4	3.0	46	48	28	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
4:	1ivo-C	4.3	2.7	44	47	61	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
5:	1mox-C	4.2	3.1	47	49	30	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
6:	1ivo-D	4.2	2.7	44	47	61	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
7:	1j19-A	4.1	2.2	41	42	71	MOLECULE: EPIDERMAL GROWTH FACTOR;
8:	1ydt-B	3.9	2.0	40	41	33	MOLECULE: DIPHTHERIA TOXIN.

PDB search results for epidermal growth factor legf

Data root-mean-square deviation of C- α atoms in the least-squares superimposition of the structurally equivalent C- α atoms. The program does not optimise rmsd.

Neighbours

Query EPIDERMAL GROWTH FACTOR;

Match more lower than 2 are spurious.

Summary

No:	Chain	Z	rmsd	lali	nres	%id	Description
1:	legf	12.7	0.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES)
2:	3egf	10.6	1.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES AFTER
3:	1mox-D	4.4	3.0	46	48	28	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
4:	1ivo-C	4.3	2.7	44	47	61	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
5:	1mox-C	4.2	3.1	47	49	30	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
6:	1ivo-D	4.2	2.7	44	47	61	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
7:	1j19-A	4.1	2.2	41	42	71	MOLECULE: EPIDERMAL GROWTH FACTOR;
8:	1ydt-B	3.9	2.0	40	41	33	MOLECULE: DIPHTHERIA TOXIN.

PDB search results for epidermal growth factor legf

DaliLite: Structural Neighbours

Query: mol1A MOLECULE: EPIDERMAL GROWTH FACTOR;

Matches are sorted by Z-score. Similarities with a Z-score lower than 2 are spurious.

Summary

No:	Chain	Z	rmsd	lali	nres	%id	Description
1:	legf	12.7	0.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES)
2:	3egf	10.6	1.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES AFTER
3:	1mox-D	4.4	3.0	46	48	28	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
4:	1ivo-C	4.3	2.7	44		61	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
5:	1mox-C	4.2	3.1	47			MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
6:	1ivo-D	4.2	2.7	44			MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
7:	1j19-A	4.1	2.2	41			MOLECULE: EPIDERMAL GROWTH FACTOR;
8:	1ydt-B	3.9	2.0	40			MOLECULE: DIPHTHERIA TOXIN.

number of structurally equivalent residues

PDB search results for epidermal growth factor legf

DaliLite: Structural Neighbours

Query: mol1A MOLECULE: EPIDERMAL GROWTH FACTOR;

Matches are sorted by Z-score. Similarities are spurious.

number of amino acids in the protein

Summary

No:	Chain	Z	rmsd	lali	nres	%id	Description
1:	legf	12.7	0.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES)
2:	3egf	10.6	1.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES AFTER
3:	1mox-D	4.4	3.0	46	48	28	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
4:	1ivo-C	4.3	2.7	44	47	61	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
5:	1mox-C	4.2	3.1	47	49	30	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
6:	1ivo-D	4.2	2.7	44	47	61	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
7:	1j19-A	4.1	2.2	41	42	71	MOLECULE: EPIDERMAL GROWTH FACTOR;
8:	1ydt-B	3.9	2.0	40	41	33	MOLECULE: DIPHTHERIA TOXIN.

PDB search results for epidermal growth factor legf

DaliLite: Structural Neighbours

Query: mol1A MOLECULE: EPIDERMAL GROWTH FACTOR;

Matches are sorted by Z-score. Similarities with a Z-score lower than 2 are spurious.

Summary

No:	Chain	Z	rmsd	lali	nres	%id	Description
1:	legf	12.7	0.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES)
2:	3egf	10.6	1.0	53	53	100	EPIDERMAL GROWTH FACTOR (EGF) (NMR, 16 STRUCTURES AFTER
3:	1mox-D	4.4	3.0	46	48	28	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
4:	1ivo-C	4.3	2.7	44	47	61	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
5:	1mox-C	4.2	3.1	47	49	30	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
6:	1ivo-D	4.2	2.7	44	47	61	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
7:	1j19-A	4.1	2.2	41	42	71	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
8:	1ydt-B	3.9	2.0	40	41	33	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;

percentage of identical amino acids over all structurally equivalent residues

Bibliography

- “Protein Structure Comparison by alignment of distance matrices”, 1993, Liisa Holm & Chris Sander
- “Dali: a network tool for protein structure comparison”, 1995, Liisa Holm & Chris Sander
- “DaliLite workbench for protein structure comparison”, 2000, Liisa Holm & John Park
- “Computational Methods for Protein Structure Prediction and Modeling, Volume 1: Basic Characterization”, Ying Xu, Dong Xu, and Jie Liang, p.151-155
- “Introduction to Bioinformatics”, Arthur M. Lesk, 2002, p.221-222
- “BIOINFORMATICS-A Practical Guide to the Analysis of Genes and Proteins”, SECOND EDITION, Andreas D. Baxevanis, B. F. Francis Ouellette, p.100,275

The end...

