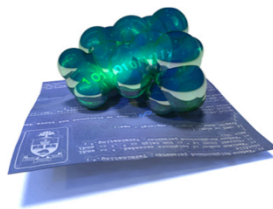


A
BIOINFORMATICS
COURSE

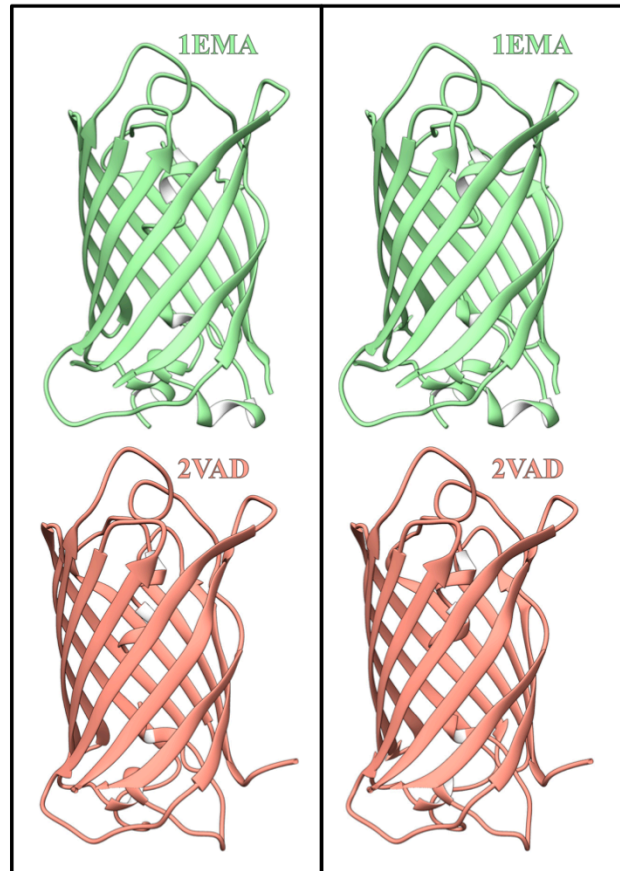
STRUCTURE SUPERPOSITION



BORIS STEIPE

*DEPARTMENT OF BIOCHEMISTRY – DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO*

SUPERPOSITION



Homologous proteins have similar structures and structural superposition means to rotate and translate the structures so that corresponding atoms are as close to each other as possible. Structural similarity is very apparent in these two proteins, the Green Fluorescent Protein of *Aequorea victoria* (1EMA) and the Red Fluorescent protein of *Discosoma striata*.

SUPERPOSITION

```

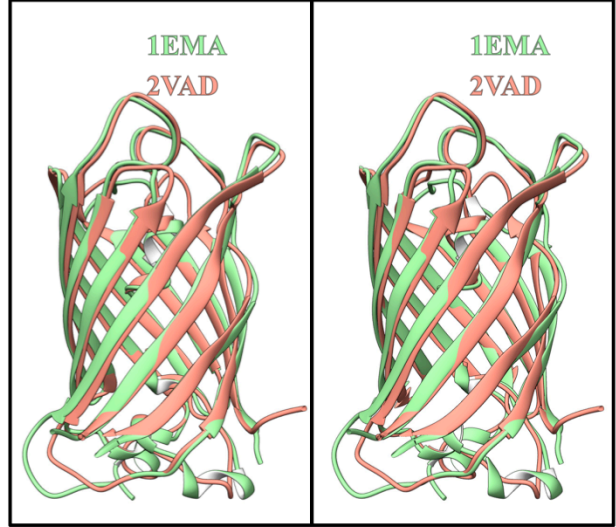
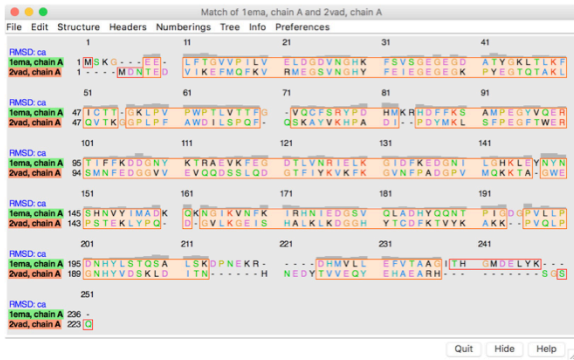
1 MSKGEELFTGVVPILVELDGDVNGHKFVSVSGEGEDATYKLTTLKFICTT      50
|...|:.....|:..|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:
1 MDNTEDEVIFEPMQFKVRMEGCVNGHYFEIEGEGEKPYEGTQAKLQVTK      50

51 -GKLPVPWPTLVTTFTYGVQCFSRYPDHMKRHDFKSAPEGYVQERTIF      99
|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:
51 GGPLPFAWDILSPQFYGSKAYVKHPADIP--DYMKLSFPEGFTWERSMN      98

100 FKDDGNYKTRAEVKFEGDVLVNRIELKGIDFKEDGNILGHKLE-YNYNSH    148
|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:
99 FEDGGVVEVQQDSSLQDGTFTFYKVKFGVNFADGPMQKKTAGWEPSTE      148

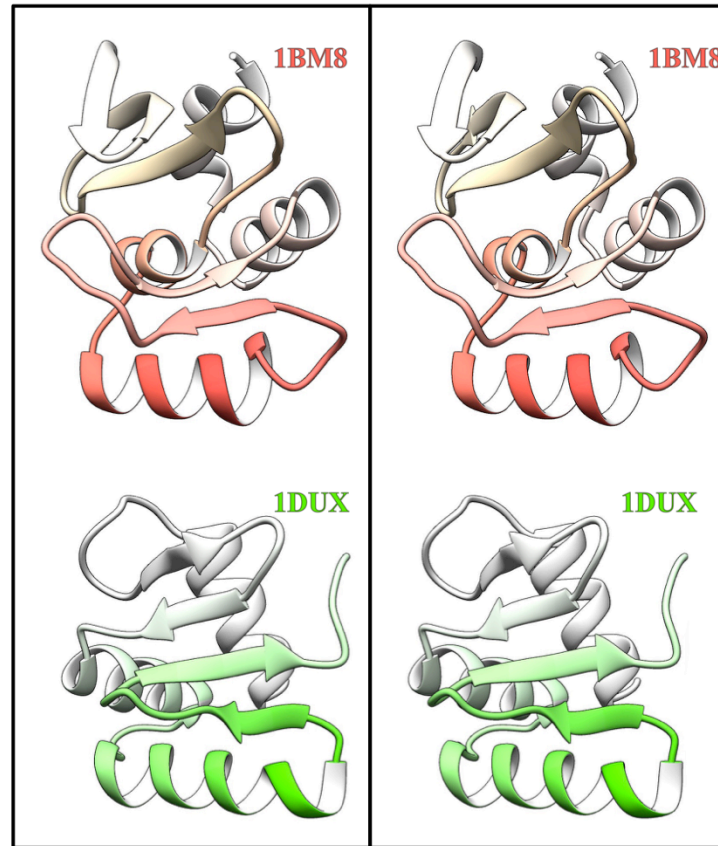
149 NVYIMADKQKNGIKVNFKIRH--NIEDGSQLADHYQ---QNTPIGDGPV      193
.:| :.:|: :.:|: :.:|: :.:|: :.:|: :.:|: :.:|: :.:|: :.:|:
149 KLY-----PQDGV-LKGEISHALKLKDGG-----HYTCDFKTVYKAKKPV      187

194 LLPDNHYLSTQSALS KDPNEKRDMVLLFVTAAGITHGMDELYK        238
|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:
188 QLPGNHVDSKLDIT---NHNEDYTVVQEYEAHARHSGSQ----        225
    
```



After superposition, the structure of these two proteins virtually overlap. Sequence similarity is recognizable over the whole length of the domains (top left), although slightly less than 25% identity. Also, the sequence alignment corresponds closely to the alignment derived from spatially close matching residues, computed by Chimera (bottom left).

SUPERPOSITION

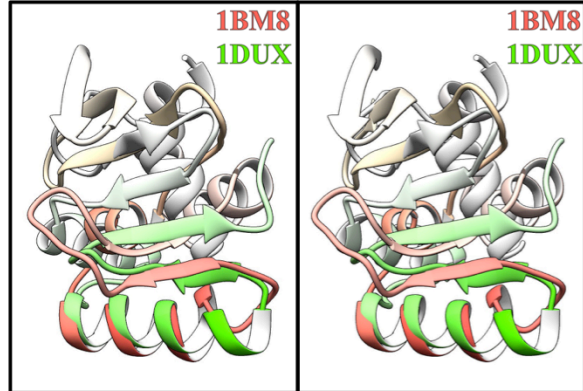


With more distant homologues, superposition may be more challenging. In this example, we compare the APSES domain of yeast Mbp1 (1BM8) and the ETS domain of the human Elk-1 transcription factor (1DUX). Both domains are members of the “winged helix” superfamily of DNA binding modules. But there are significant topological rearrangements which make it challenging to match corresponding residues.

SUPERPOSITION

```

1 QIYSARYSGVDVYEFIHSTGSIIMKRKKDDWVNATHILKAANFAKAKRTRI 50
1 ----- 0
51 LEKEVLKETHKVEKVGQGGFGKYOGTWVPLNIAKQLAEKFSVYDQLKPLDF- 99
   .|...|:|
1 -----MDPSVTLWQFL 11
100 ----- 99
12 LQLLREQGNHIIISWTSRDGGEFKLVDAEEVARLWGLRKNKTNMNYDKLS 61
100 ----- 99
62 RALRYYYDKNIIRKVSQKQFVYKFVSYPEVAGC 94
  
```



Match of 1bm8, chain A and 1dux, chain C

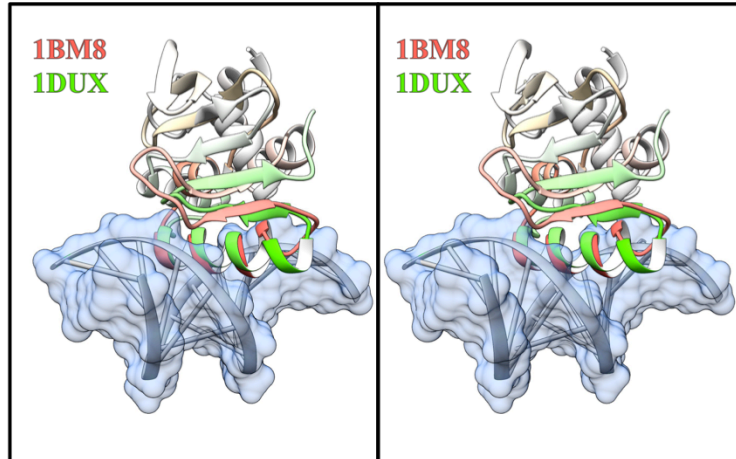
File	Edit	Structure	Headers	Numberings	Tree	Info	Preferences	
RMSD: ca				11	21	31	41	
1bm8, chain A	4	-----	-----	Q I	Y S A R Y S G V D V	Y E F I H S T G S I	M K - R K K D D W V	
1dux, chain C	1	M D P S V T L W Q F	L L Q L L R E Q C	-----	-----	N G H - - I - I	S W T S R D G - G E	
RMSD: ca				51	61	71	81	91
1bm8, chain A	35	N A T - H I L - - K	- - A A N - - - -	- - - - -	F A K A K	R T R I L E K E V L	- K E T H E K V Q Q	
1dux, chain C	34	F - - K L - - V A	E E V A R L W G L R	K N K T N M N Y D K	L S R A L R Y - Y Y	D K N I I R K V S G		
RMSD: ca				101	111	121	131	141
1bm8, chain A	68	F G K Y Q - G T W	V - - P L N I A K Q	L A E K F S V Y D Q	L K P L F D F - - -			
1dux, chain C	79	Q - - - K F V Y K	F V S Y P E - - - -	-----	-----	- - - - -	V A G C	

Quit Hide Help

Indeed, sequence alignment fails completely to discover any reasonable region of similarity. However structural superposition matches the “helix and wing” motif quite well, and the superposition-derived alignment (bottom left) shows significant sequence conservation.

Superposition is valuable for the analysis of distant family relationships and conservation patterns, but it has other important uses too, for the analysis of interaction sites.

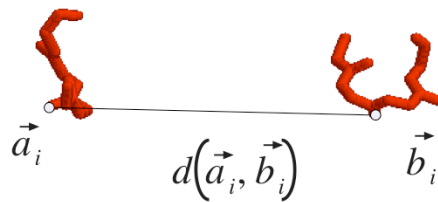
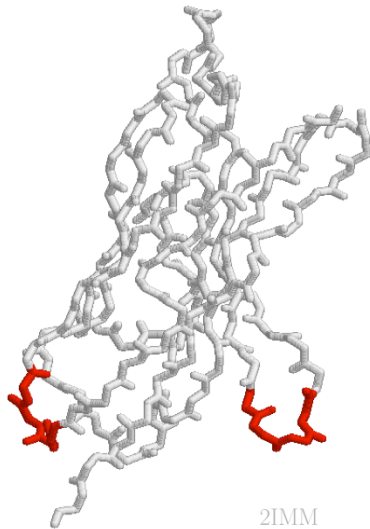
RMSD



For example, after superposition of 1BM8 with 1DUX, which is a structure of a protein DNA complex, we can study the detailed interactions of the helix with the DNA major groove, and the apex of the “wing” with the DNA minor groove, **and evaluate whether these interactions may be conserved.**

This analysis may allow conclusions about the DNA binding mode of 1BM8, for which no structure of a protein-DNA complex has been determined so far.

RMSD



$$\text{RMSD}_{\text{coord}}(\mathbf{A}, \mathbf{B}) = \sqrt{\frac{1}{n} \sum_{i=1}^n d(\vec{a}_i, \vec{b}_i)^2}$$

To calculate a RMSD, a *pairwise correspondence* of points has to be defined.

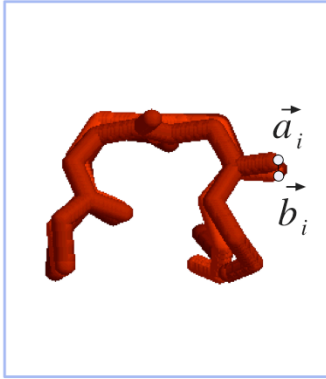
Optimal superposition aims to minimize the **RMSD** between two sets of matching atoms.

RMSD or *root mean square deviation* is simply the square root of the average sum of squared coordinate distances. However, this is just a measure of the relationship between two sets of points in space - it depends on the distance between the point sets, their rotation and the quantitativ we are interested in: their intrinsic structural similarity.

See also: Structural Alignment (Wikipedia)

(http://en.wikipedia.org/wiki/Protein_structural_alignment)

RMSD_{OPT}



$$\text{RMSD}_{\text{opt}} = \min(\text{RMSD}_{\text{coord}})$$

$$\text{RMSD}_{\text{opt}} = \text{RMSD}_{\text{coord}}(\mathbf{A}, (\mathbf{B} - \mathbf{T}_s) \times \mathbf{R}_s)$$

The translation vector \mathbf{T}_s and the rotation matrix \mathbf{M}_s define a *superposition* of the vector set \mathbf{B} on \mathbf{A} .

An analytic solution of the superposition problem is available, but not straightforward (involves an eigenvalue problem). But fortunately, the measure is a true metric!

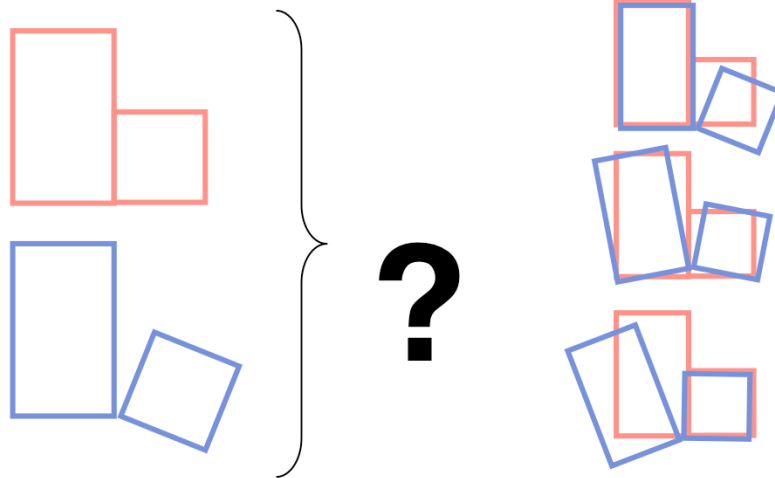
A meaningful comparison of structural segments requires that the coordinate sets at first be **optimally** “superimposed”: this means find a translation and rotation that minimizes the residual RMSD.

Note that only this analytic part is a solved problem. The **choice** of coordinate pairs to superimpose is difficult. Just as with sequence alignment, this choice is only straightforward if the number of coordinates (residues) in both proteins is the same. But if there are indels, that number changes, and disordered sections of loops or termini should not be included in the superposition anyway. Moreover, RMSD values are lowest for a small, structurally conserved set of residues which may not be representative of global structural distortions.

Thus the major computational challenge is to find which pairs of atoms should be matched between two structures. This problem has no clear algorithmic solution, and successful algorithms apply heuristics that may include global and local similarities, coarse grained approximations of secondary structure elements, and iterated improvements.

CHOICE OF COORDINATE SETS

Rigid body movement of domains or subdomains:



Relative domain motion (and sub-domain motion to a degree) can often be approximated as independent rigid bodies, joined by a flexible hinge. Global superposition may not give satisfactory results. Local superposition requires to define the domain boundaries and does not preserve interface geometries. Superposition applies the same rotation and translation to all atoms, a smoothly varying deformation may be more appropriate to model “real” molecular relationships.

The Godzik lab’s FATCAT server addresses this problem, the algorithm is available for structure comparisons at the PDB.

cf. Yuzhen Ye, Y. and Godzik, A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.* 32: W582–W585. (PMID 15215455)

SUPERPOSITION SERVERS

A number of
Web servers offer
collections of
superimposed
domains.

VAST

([https://
www.ncbi.nlm.nih.gov/
Structure/VAST/
vast.shtml](https://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml))

1BM8 : Dna-Binding Domain Of Mbp1

Biological unit 1: monomeric
Source organism: *Saccharomyces cerevisiae*
Number of proteins: 1 (TRANSCRIPTION FACTOR MBP1)

Similar Structures (253)

Filter by number of matching molecules: Complete match, 1 proteins (253)

Filter by taxonomy:

- Eukaryota (57)
- Bacteria (130)
- Archaea (33)
- Viruses (6)
- Others (27)

PDB ID	Description	Taxonomy	Aligned Protein	RMSD	Aligned Residues	Sequence Identity
1	1MB1 Mbp1 From <i>Saccharomyces Cerevisiae</i>	<i>Saccharomyces cerevisiae</i>	1	0.61Å	97	100%
2	4JX5 Structure Of Dna Complex Of Pcg2	<i>Candida albicans/Magnaporthe oryzae</i>	1	0.80Å	96	60%
3	1KAF Dna Binding Domain Of The Phage T4 Transcription Factor Mota (Aa105-211)	<i>Enterobacteria phage T4</i>	1	2.40Å	57	5%
4	1L8R Structure Of The Retinal Determination Protein Dachshund Reveals A Dna-Binding Motif	<i>Homo sapiens</i>	1	2.13Å	46	4%
5	3GE9 A Structurally Atypical Thyx From <i>Corynebacterium Glutamicum</i> Nchu 87078 Is Not Required For Thymidylate Biosynthesis	<i>Corynebacterium glutamicum</i>	1	2.25Å	43	7%
6	3DAE Crystal Structure Of Phosphorylated Snf1 Kinase Domain	<i>Saccharomyces cerevisiae</i>	1	2.65Å	42	2%
7	4M2Q Crystal Structure Of Non-myristoylated Recoverin With Cysteine-39 Oxidized To Sulfenic Acid	<i>Bos taurus</i>	1	2.66Å	42	10%
8	2HBV Structure Of Empty Pheromone Binding Protein Asp1 From The Honeybee <i>Apis Mellifera</i> L.	<i>Apis mellifera</i>	1	3.18Å	39	10%
9	3FEB Crystal Structure Of A Pheromone Binding Protein From <i>Apis Mellifera</i> With A Serendipitous Ligand Soaked At Ph 4.0	<i>Apis mellifera</i>	1	3.18Å	39	10%

VAST (Vector Alignment Search Tool) finds similar structures by searching for similarly oriented and arranged elements of secondary structure.

SUPERPOSITION SERVERS

A number of Web servers offer collections of superimposed domains.

DALI

(<http://ekhidna2.biocenter.helsinki.fi/dali/>)

The screenshot displays the DALI Protein Structure Comparison Server interface. At the top, the DALI logo and "PROTEIN STRUCTURE COMPARISON SERVER" are visible. Below the navigation menu, a brief description of the server's function is provided. The main content area shows the search results for the query "1bm8A".

Results: 1bm8A

Query: 1bm8A

MOLECULE: TRANSCRIPTION FACTOR MBP1;

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Structural Alignment Expand gaps 3D Superimposition (3V) SANS PANZ Reset Selection

Summary

No:	Chain	Q	rmsd	lali	nres	qid	PDB	Description
1:	113a	A	2.8	2.8	86	100	1BM8	MOLECULE: TRANSCRIPTION FACTOR MBP1;
2:	2xy7	A	6.5	3.0	82	108	9 PDB	MOLECULE: UNP00-SIMILAR TO XCOV ORP19;
3:	2z4v	A	5.4	3.0	79	108	14 PDB	MOLECULE: REGULATORY DOMAIN OF SWI6;
4:	1l8c	A	5.3	3.0	74	101	3 PDB	MOLECULE: SMOGSHW07;
5:	1aba	A	4.70	2.90	66	106	4 PDB	MOLECULE: SFI ONCOGENE;
6:	5j1a	A	3.9	2.7	62	114	6 PDB	MOLECULE: MIDDLE TRANSCRIPTION REGULATORY PROTEIN MTRF1;
7:	8qj1	A	2.8	3.0	62	114	3 PDB	MOLECULE: SMI/SNF-RELATED MATRIX-ASSOCIATED ACTIN-DEPENDENT CELL CYCLES 1;
8:	3p1y	A	2.4	3.0	48	314	10 PDB	MOLECULE: TRNA-SPlicing ENDO nucleic acid; RNA-SPlicing ENDO nucleic acid;
9:	3eas	A	2.4	3.0	59	85	5 PDB	MOLECULE: DNA-BINDING PROTEIN;
10:	2dy1	A	2.3	3.4	75	213	1 PDB	MOLECULE: AUTOPHAGY-RELATED PROTEIN 3;
11:	5a5v	B	2.3	3.4	61	1183	5 PDB	MOLECULE: DNA-DIRECTED RNA POLYMERASE I SUBUNIT RPA190;
12:	2q7r	A	2.3	3.2	64	112	5 PDB	MOLECULE: NUTRIENT CYTOSOLIC PROTEIN;
13:	4y1a	A	2.2	4.1	65	375	3 PDB	MOLECULE: AROMATIC PRENYLTRANSFERASE;
14:	1yb3	A	2.2	3.3	71	166	7 PDB	MOLECULE: HYPOTHETICAL PROTEIN;
15:	3hy1	A	2.1	4.1	58	284	5 PDB	MOLECULE: PROTEIN SDF19/PH1A;
16:	5dwy	B	2.0	3.4	62	428	10 PDB	MOLECULE: PROTON/GLUTAMATE SYMPORTER, SDF FAMILY;
17:	5ekv	B	2.0	3.1	59	400	7 PDB	MOLECULE: LEGK4;

Pairwise Structural Alignments

Notation: three-state secondary structure definitions by DSSP (reduced to H=helix, E=sheet, L=coil) are shown above the amino acid sequence.

The interface also features two 3D ribbon diagrams comparing the query structure (1BM8) and a hit (2XFV). The query is shown in white, and the hit is shown in orange and purple. The DALI logo and "PROTEIN STRUCTURE COMPARISON SERVER" are visible at the top of the interface.

The list of matches returned by DALI for a search with 1BM8 has a number of very interesting hits, more, and more relevant than the hits VAST had discovered.

2XFV – the N-terminal domain of yeast Swi6 – does **not** bind DNA, and the structural superposition rationalizes this well. The two sequences have only about 10% pairwise identity after alignment: homology can not be inferred from sequence similarity in this case.

<http://steipe.biochemistry.utoronto.ca/abc>

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA