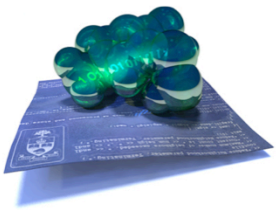
A
Bioinformatics
Course

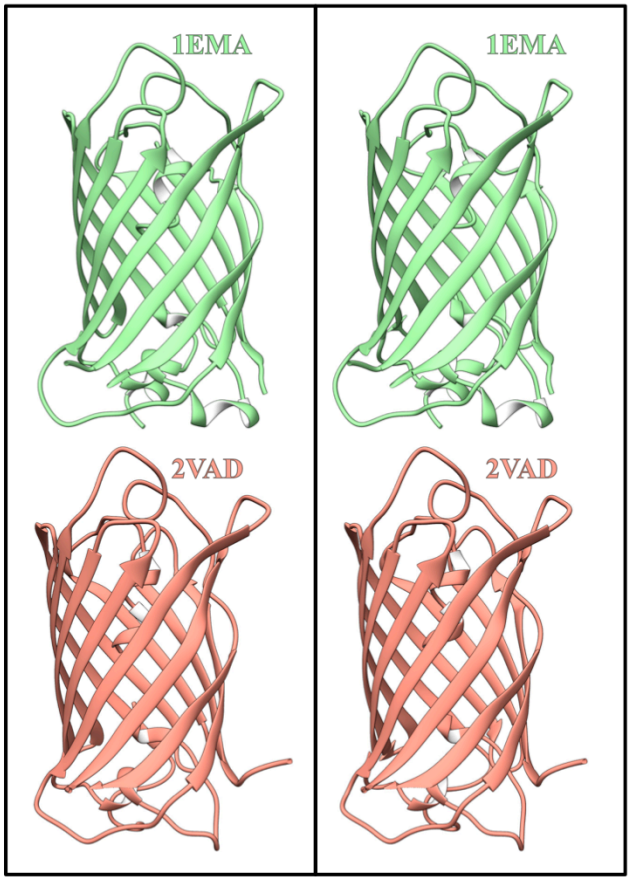
Structure Superposition



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

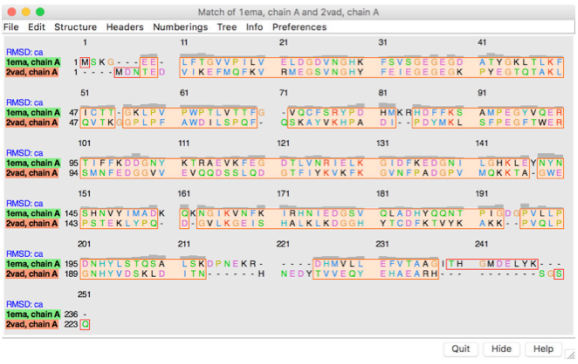
Superposition

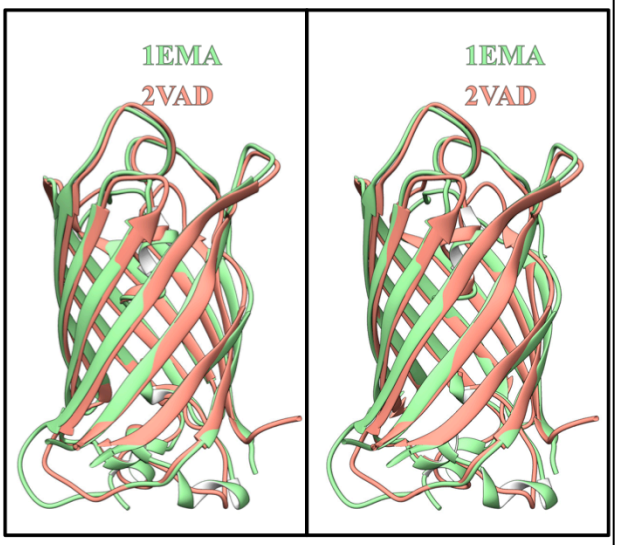


Homologous proteins have similar structures and structural superposition means to rotate and translate the structures so that corresponding atoms are as close to each other as possible. Structural similarity is very apparent in these two proteins, the Green Fluorescent Protein of *Aequorea victoria*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

Superposition

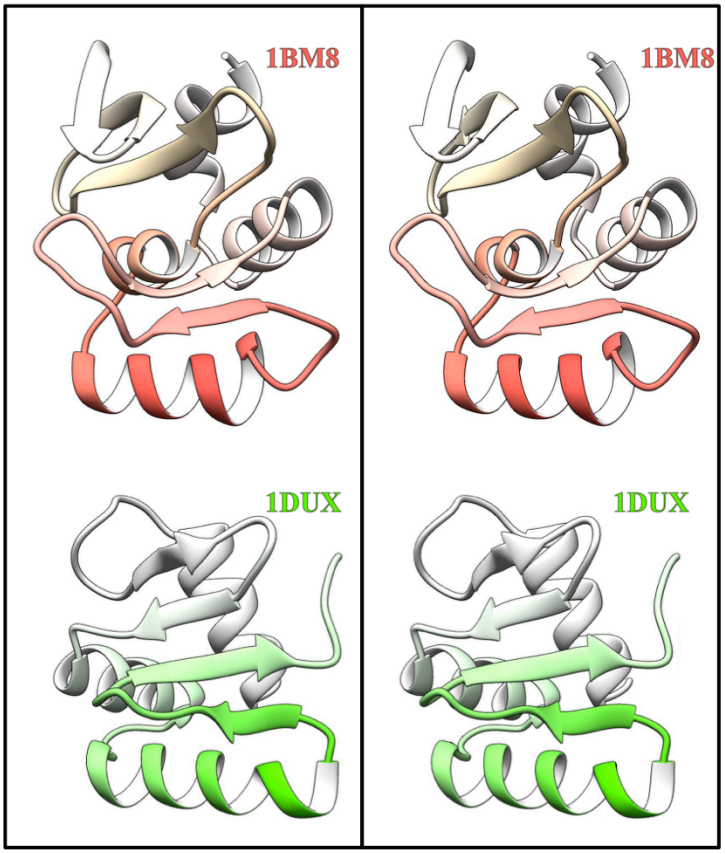
```
  1 MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTT     50
    |...|::....:...|.::|.||||.|.:.|||||....|..|.|...|.
  1 MDNTEDVIKEFMQFKVRMEGSVNGHYFEIEGEGEGKPYEGTQTAKLQVTK     50

 51 -GKLPVPWPTLVTTFTYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTIF     99
     |.||..|..|...|.||.:.:.::|..:.  |:.|.:.|||:..||::.
 51 GGPLPFAWDILSPQFQYGSKAYVKHPADIP--DYMKLSFPEGFTWERSMN     98

100 FKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLE-YNYNSH    148
    |:|.|..:.:.:...:..|.:.:::.||::|..||.::..|.. :..::.
 99 FEDGGVVEVQQDSSLQDGTFIYKVKFKGVNFPADGPVMQKKTAGWEPSTE    148

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

194 LLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK    238
    .||.|||:.::..::    |...|:.|:.::..|.....|..
188 QLPGNHYVDSKLDIT---NHNEDYTVVEQYEHAEARHSGSQ----    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).

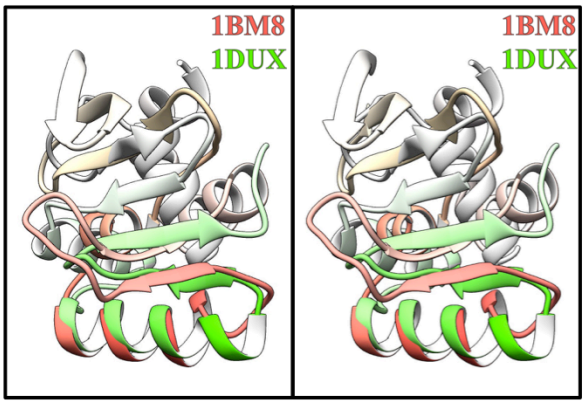
Superposition

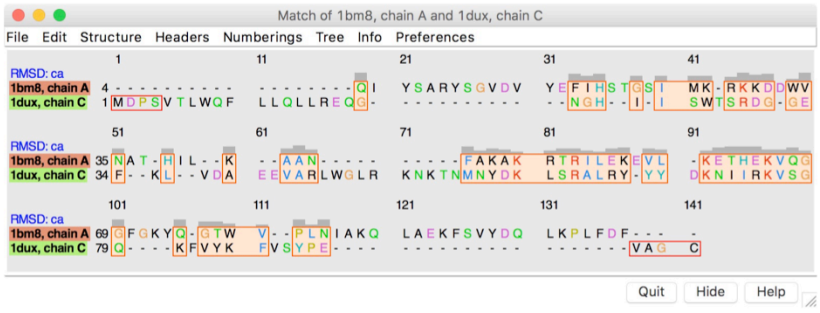


With more distant homologues, superposition may be more challenging. In this example, we compare the APSES domain of yeast Mbp1 (1BM8) and the ETS domain of the human Elk-1 transcription fator (1DUX). Both domains are members of the "winged helix" superfamily of DNA binding modules. But there are significant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

Superposition

```
  1 QIYSARYSGVDVYEFIHSTGSIMKRKKDDWVNATHILKAANFAKAKRTRI     50
  1 -------------------------------------------------       0
 51 LEKEVLKETHEKVQGGFGKYQGTWVPLNIAKQLAEKFSVYDQLKPLFDF-     99
                                          .|....|:.|
  1 ----------------------------------------MDPSVTLWQFL     11
100 -------------------------------------------------      99
 12 LQLLREQGNGHIISWTSRDGGEFKLVDAEEVARLWGLRKNKTNMNYDKLS     61
100 ----------------------------------     99
 62 RALRYYYDKNIIRKVSGQKFVYKFVSYPEVAGC     94
```

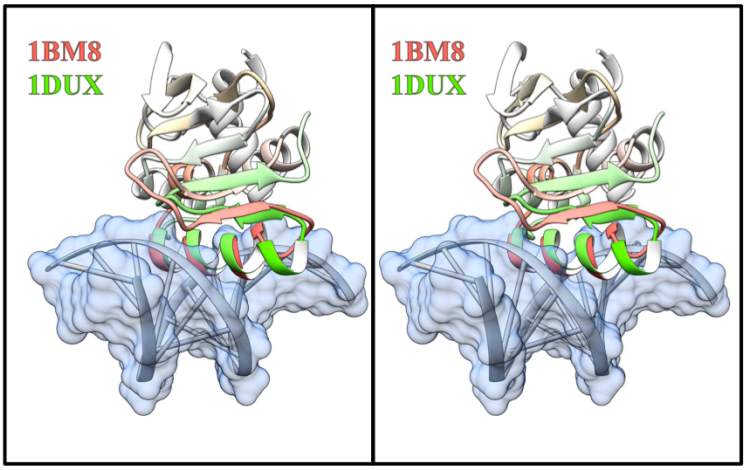




Indeed, sequence alignment fails completely to discover any reasonable region of similarity. However structural superposition matches the "helix and wing" motif quite well, and the superposition-derived alignment (bottom left) shows significant sequence conservation.

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.

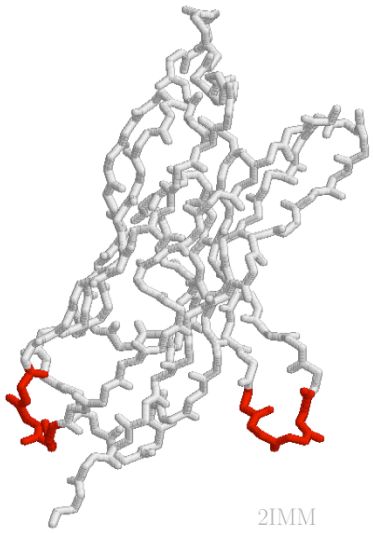
RMSD

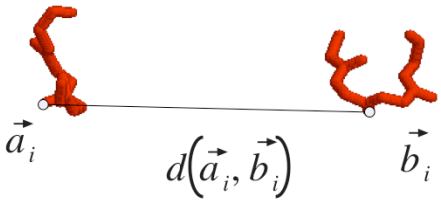


For example, after superposition of 1BM8 with 1DUX, which is a structure of a protein DNA complex, we can study the detailed interactions of the helix with the DNA major groove, and the apex of the "wing" with the DNA minor groove, **and evaluate whether these interactions may be conserved**.

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





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

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

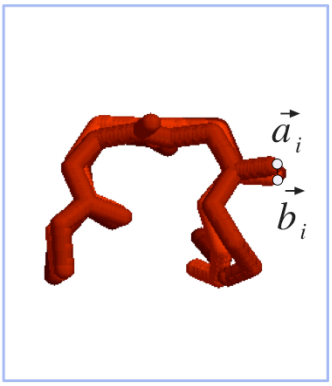
Optimal superposition aims to minimize the **RMSD** between two sets of matching atoms.

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 quantitatiy we are interested in: their intrinsic structural similarity.

See also: Structural Alignment (Wikipedia)

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

$RMSD_{OPT}$



$$RMSD_{opt} = \min(RMSD_{coord})$$

$$RMSD_{opt} = RMSD_{coord}(\mathbf{A}, (\mathbf{B}\text{-}T_s) \times R_s)$$

The translation vector T_s and the rotation matrix M_s define a *superposition* of the vector set **B** on **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!

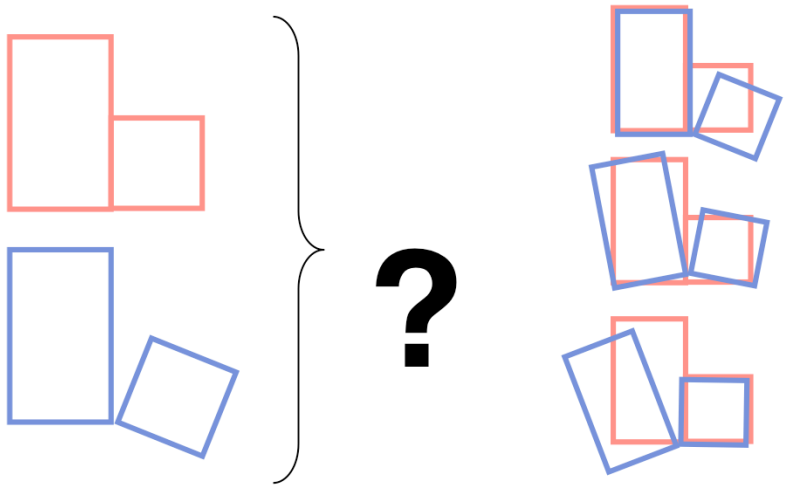
A meaningful comparison of structural segments requires that the coordinate sets at first be **optimally** "superimposed": this means find a translation and rotation that minimizes the residual RMSD.

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 straigthforward 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 gloabal 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.

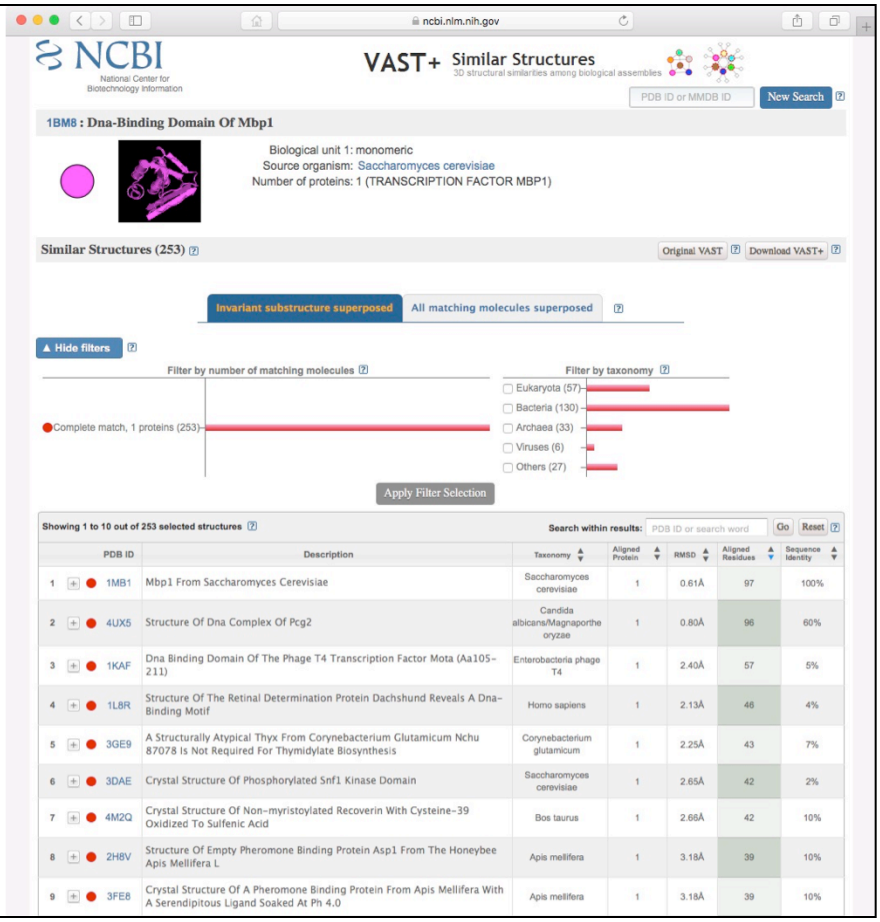
cf. Yuzhen Ye, Y. and Godzik, A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res. 32: W582–W585. (PMID 15215455)

Superposition Servers

A number of Web servers offer collections of superimposed domains.

VAST

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



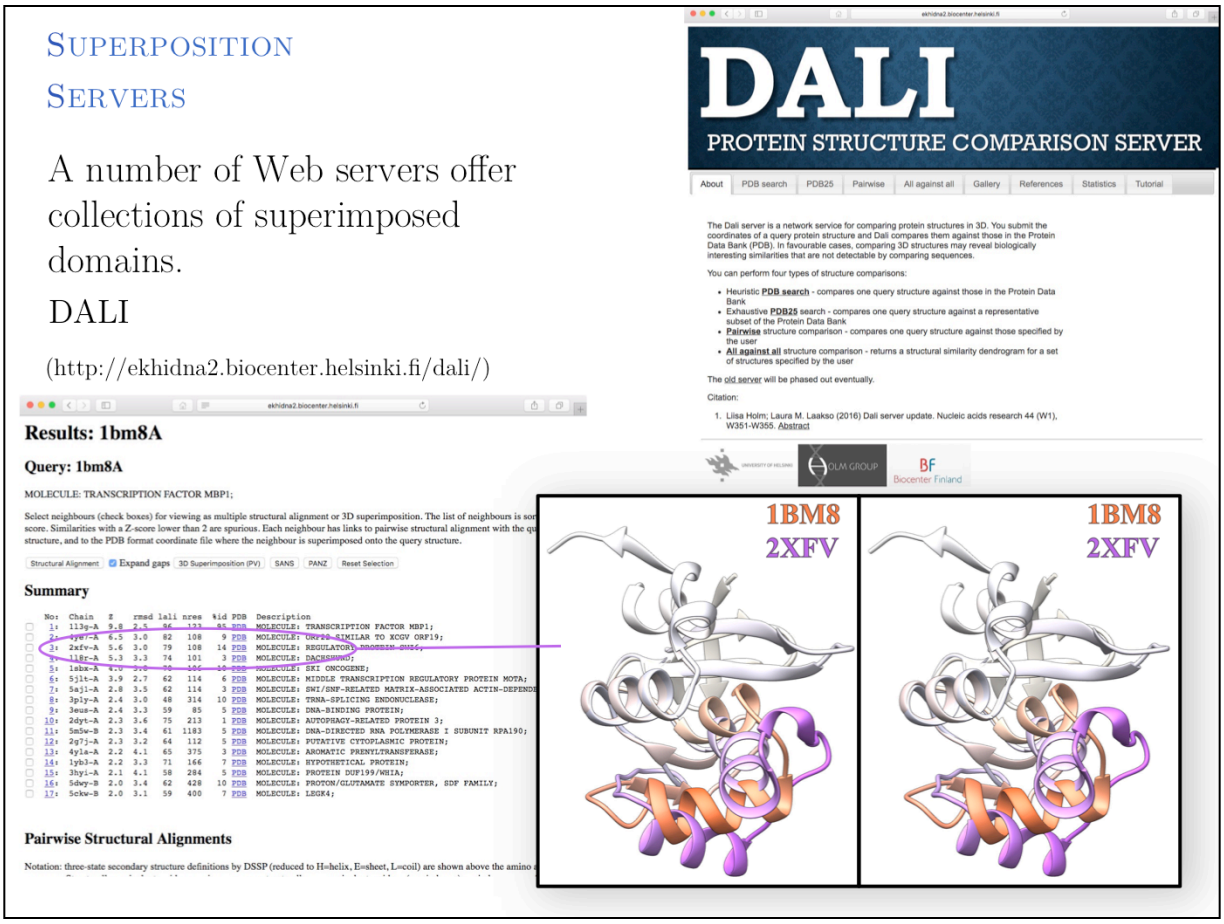
VAST (Vector Alignment Search Tool) finds similar structures by searching for similarly orianted and arranged elements of secondary structure.

Superposition Servers

A number of Web servers offer collections of superimposed domains.

DALI

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



The 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