Fast Substructure Searching Using the Conformation Likeness Method

Misha P. Ponomarenko
Institute of Cytology and Genetics
The Russian Academy of Sciences, Siberian Branch, Novosibirsk, 630090 Russia
pon@cgi.nsk.su

Ilya N. Shindyalov and Philip E. Bourne
San Diego Supercomputer Center
P.O. Box 85608, San Diego, CA 92186-5608
shindyal@sdsc.edu bourne@sdsc.edu http://www.sdsc.edu/pb/Group.html

Abstract

Conformational Likeness is a new method to search for similar 3-D structure fragments within the complete PDB, or perform a detailed structure comparison of one structure against another. The method is distinguished by the speed of searching and the large choice of structural features to use in the comparison. Searches can be performed through the World Wide Web at the URL http://xtal1.sdsc.edu/misha/misha.html.

This paper does not detail the theory of the method, which is being published elsewhere, but concentrates on how to search effectively using specific examples and interpreting the results of those searches. An on-line version of this tutorial can be found at http://xtal1.sdsc.edu/misha/tut_cl.htm

1 Introduction

3-D structure similarity - finding structures with common folding motifs or similar spatial arrangements of key structural features - is important to both structural biologists and crystallographers. Beyond the obvious questions relating to biological function, are the questions that crystallographers frequently ask of themselves, “ this piece of poypeptide chain I have just fitted looks unusual; how unusual is it? Could I have made a mistake in the chain tracing?” This paper describes a methodology to addresses these questions.

The literature is rich in methods for determining 3-D structure similarity, but relatively poor in available software or Web sites for 3-D similarity searching. Our goal was to come up with a practical methodology for substructure searching that could be run through the Web, that is, was close to real-time and yet could keep pace with searching against an exponential growing body of data.

In simple terms, methods for determining 3-D similarity consist of 3 basic steps.

1. Represent macromolecular structures in a way that facilitates comparison with a single structure or substructure.
2. Apply a method of comparison.
3. Determine that the results are meaningful.

It is not the purpose of this paper to provide a review of the various methods used at each step. However, here are some examples to provide the context for introducing conformational likeness.

Common representations used in step 1 are:

- Cα contact maps (Holm and Sander, 1995).
- Property Profiles (Zhang and Eisenberg, 1994).
- Side chain contacts (Godzik, Skolnick, and Kolinski, 1993).
- Geometric hashing (Nussinov and Wolfson, 1991).
- Spacial arrangement of secondary structure elements (Mizuguchi and Go, 1995, Alexandrov, 1996).

and for step 2:

- Monte Carlo (Holm and Sander, 1995).
- Dynamic Programming (Orengo and Taylor, 1993).
- Fuzzy Logic (this method).

There are a variety of statistical significance tests applied in step 3.

The types of problems that plague these methods are:

- The best structural alignment requires human intervention.
- Limited to finding similarity in very similar 3-D structures.
- Find too many non-significant homologies.
- Require a high degree (>30%) sequence similarity.
- Sensitive to insertions and deletions.
- Time consuming to compute.

Notwithstanding, these methods have been used successfully to produce databases organized according to structural similarity, for example FSSP (Holm and Sander, 1996) which uses DALI (Holm and Sander, 1995) to define the alignment; CATH (Orengo *et al.,* 1993), which uses SSAP (Taylor, Flores, and Orengo, 1994); and Entrez which uses VAST (see Bryant and Hogue this volume).

The conformational likeness method is seen to complement these existing approaches. A potential advantage is the variety of parameters that can be used in the comparison. A disadvantage is in results interpretation and in interpretation of the significance of the results obtained - there has been to equivalent to a Z score determined at this time.

2 Qualitative Description of the Method

From a user's perspective the method of conformational likeness can be characterized by the following steps..

1. Calculate a comprehensive set (currently 495) of conformational and physiochemical features on each protein in the PDB. This set is referred to as a "conformational likeness profile" on each structure.
2. Store the conformational likeness profile in a way that facilitates searching for a biological meaningful subset (e.g., local geometry, overall topology, secondary structure) of profiles. This is based on new data models which are beyond the scope of this paper (Shindyalov and Bourne, 1996, Shindyalov and Bourne 1997). Proceed to steps 3 or 5.
3. **For a search against the complete database** take a starting structure (polypeptide chain or fragment) and search for a like subset by determining whether there is conformational likeness between the starting structure and each structure in the database for a small profile. Comparison to a random sample is used to determine the significance of any possible matches. If the starting structure is not already in the database, step 1 must first be performed on it
4. Present the results as a list of possible hits with the degree of likeness indicated.
5. **For a detailed structure alignment** Perform a more detailed alignment between two polypeptide chains or fragments. The alignment is based on dynamic programming and uses a larger number of conformational features than the database scan.
6. Present the results as: (i) a conformational likeness matrix with the alignment highlighted; (ii) a sequence comparison based on the structure alignment; (iii) a stereo plot of the superimposed structures using a least squares minimization of CA distances (Hendrickson, 1979), where the superposition is color-coded to highlight the agreement in the likeness profiles; (iv) be able to invoke Rasmol as a helper application so that the color-coded superposition can be analyzed in more detail; and (v) be able to download the superimposed coordinates in PDB format and render them locally with the software of choice.

2.1 Likeness Profiles

All conformational features are based upon pentapeptides as the base unit. Pentapeptides were chosen since they encode the local properties of a region of polypeptide chain yet are not computationally time-consuming to calculate.

The following nomenclature is used to describe the geometrical features of pentapeptides, starting from the N terminus:
A is the CA-atom of the 1st residue
B is the CA-atom of the 2nd residue
C is the CA-atom of the 3rd residue
D is the CA-atom of the 4th residue
E is the CA-atom of the 5th residue

K is the center of mass for A, B, and C
I is the center of mass for B, C, and D
J is the center of mass for C, D, and E

M is the center of mass-center for A, B, C, D, and E
P is the center of mass for the complete protein based on all CA positions
L is the center of mass of the decamer (again based on CA positions) preceding the given pentapeptide
R is the center of mass of the decamer following the given pentapeptide
G is the center of mass of all CA's that reside within a 25Å shell around the given pentapeptide.

A total of 495 conformational likeness features have been defined, in part, based of this set of geometrical points {A, B, C, D, E, K, I, J, M, P, L, R, G}, for each residue of each protein in the PDB and stored. These features are defined in Table 1. When measuring protein similarity these features can be combined together in various ways described subsequently.

Table 1. Components of a Conformational Likeness Profile

Geometrical (Sets 1-5) and physical and chemical (Sets 6-12) features used to characterize structural similarity, grouped by feature type. The nomenclature used is that described in the text.

Set	Feature Description	Feature Parameters
1	Absolute (Euclid's) distance, XY, between points *X* and *Y*,	*XY={ AC, AD, AE, BE, CE, AB, BC, BD, CD, DE, MA, MB, MC, MD, ME, PA, PB, PC, PD, PE, GA, GB, GC, GD, GE, LA, LB, LC, LD, LE, RA, RB, RC, RD, RE, LM, LJ, LR, KR, MR, LK, KM, KJ, MJ, JR, GL, GK, GM, GJ, GR}*
2	Relative distance, *XYdZ*, that is : $XYdZ = \frac{XY - dZ}{XY + dZ}$ where *dZ* is an average distance from a point Z to all CA-atoms of the protein. Clearly, *XYdZ* has the following simple behavior: $XYdZ=0$ when $XY=dZ$ $0<XYdZ<1$ when $XY>dZ$ $-1<XYdZ<0$ when $XY<dZ$	*XYdZ={PAdA, PBdB, PCdC, PDdD, PEdE, PAdP, PBdP, PCdP, PDdP, PEdP, LadA, LBdB, LCdC, LDdD, LedE, LAdP, LBdP, LCdP, LDdP, LEdP, RAdA, RBdB, RCdC, RDdD, RedE, RAdP, RBdP, RCdP, RDdP, REdP, GAdA, GBdB, GCdC, GDdD, GEdE, GAdP, GBdP, GCdP, GDdP, GEdP}*
3	Angle between vectors *X*→ *Y* and *Y*→ *Z*	*XYZ={ABE, ACD, ACE, ADE, BCE, AMC, AMD, AME, BME, CME, ABC, ABD, BCD, BDE, CDE, AMB, BMC, BMD, CMD, DME, APC, APD, APE, BPE, CPE, APB, BPC, BPD, CPD, DPE, AGC, AGD, AGE, BGE, CGE, AGB, BGC, BGD, CGD, DGE, ALC, ALD, ALE, CRD, DRE, LKR, LMJ, LMR, LJR, KMR, LGM, LGJ, LGR, KGR, MGR, LKM, LKJ, KMJ, KJR, MJR, LGK, KGM, KGJ, MGJ, JGR, MLG, MIR, MIG, MRG, LIG, MPI, MPR, MPG, LPG, IPG, MLI, MLR, LIR, LRG, IRG, MPL, LPI, LPR, IPR, RPG, PMA, PMB, PMC, PMD, PME, LMA, LMB, LMC, LMD, LME, RMA, RMB, RMC, RMD, RME, GMA, GMB, GMC, GMD, GME}*
4	Angle XY-ZU between vectors X→ *Y* and *Z*→ *U*	*XY-ZU={PM-AC, PM-AD, PM-AE, PM-BE, PM-CE, PM-AB, PM-BC, PM-BD, PM-CD, PM-DE, LM-AC, LM-AD, LM-AE, LM-BE, LM-CE, LM-AB, LM-BC, LM-BD, LM-CD, LM-DE, RM-AC, RM-AD, RM-AE, RM-BE, RM-CE, RM-AB, RM-BC, RM-BD, RM-CD, RM-DE, GM-AC, GM-AD, GM-AE, GM-BE, GM-CE, GM-AB, GM-BC, GM-BD, GM-CD, GM-DE }*

Table 1 cont. Components of a Conformational Likeness Profile

5	Dihedral angle built for points *X*, *Y*, *Z*, *U* (between planes *XYZ* and *YZU*)	*XYZU={ABCD, ABCE, ABDE, ACDE, BCDE, MBDA, MBDC, MBDE, AMCE, BMCD, PABE, PACD, PACE, PADE, PBCE, PABC, PABD, PBCD, PBDE, PCDE, APMC, APMD, APME, BPME, CPME, APMB, BPMC, BPMD, CPMD, DPME, PACM, PADM, PAEM, PBEM, PCEM, PABM, PBCM, PBDM, PCDM, PDEM, LABE, LACD, LACE, LADE, LBCE, LABC, LABD, LBCD, LBDE, LCDE, ALMC, ALMD, ALME, BLME, CLME, ALMB, BLMC, BLMD, CLMD, DLME, LACM, LADM, LAEM, LBEM, LCEM, LABM, LBCM, LBDM, LCDM, LDEM, RABE, RACD, RACE, RADE, RBCE, RABC, RABD, RBCD, RBDE, RCDE, ARMC, ARMD, ARME, BRME, CRME, ARMB, BRMC, BRMD, CRMD, DRME, RACM, RADM, RAEM, RBEM, RCEM, RABM, RBCM, RBDM, RCDM, RDEM, GABE, GACD, GACE, GADE, GBCE, GABC, GABD, GBCD, GBDE, GCDE, AGMC, AGMD, AGME, BGME, CGME, AGMB, BGMC, BGMD, CGMD, DGME, GACM, GADM, GAEM, GBEM, GCEM, GABM, GBCM, GBDM, GCDM, GDEM, LKMJ, LKMR, LKJR, LMJR, KMJR, GKJL, GKJM, GKJR, LGMR, KGMJ, MLIR, MLIG, MLRG, MIRG, LIRG, PLRM, PLRI, PLRG, MPIG, LPIR}*
6	Values of Exposure and Polarity for each side-chains in a pentapeptide calculated according to Lee & Richards (1971).	
7	Amino acid codes for all 5 residues in the pentapeptide (in 1-letter alphabet)	
8	Values of static physical and chemical properties of amino acids defined by sequence: exposure, polarity, hydrophobicity, isoelectric point, volume, number of chemical bonds, molecular weight, Chou-Fasman alpha-helix and beta-strand propensities;	
9	Amino acid frequencies observed in: (a) whole protein; (b) 25-peptide segment centered at the pentapeptide; and (c) set of residues which reside in a 25Å shell around the pentapeptide.	
10	Number of residues residing in a 25Å shell around the pentapeptide	
11	Codes for secondary structure defined for the pentapeptide according to Kabsch & Sander (1983) using in the following alphabet:{H, G, I, T, B, E, S}.	
12	Main chain dihedral angles phi and psi.	

Table 2. Grouping of Likeness Profiles used in Fast Searching.

Alignment and search modes (bold) when concatenated indicate the option presented to the user, for example, LocalDist when performing a complete database search.

Alignment	Search	Conformational Features
Local	**Dist**_ance_	*Euclid’s distances XY={AC, AD, AE, BE, CE, AB, BC, BD, CD, DE, MA, MB, MC, MD, ME}*
	Angle	*Plane angles XYZ={ABE, ACD, ACE, ADE, BCE, AMC, AMD, AME, BME, CME, ABC, ABD, BCD, BDE, CDE, AMB, BMC, BMD, CMD, DME }*
	Twist	*Dihedral angles XYZU={ABCD, ABCE, ABDE, ACDE, BCDE, MBDA, MBDC, MBDE, AMCE, BMCD }*
	Surface	Exposure and polarity calculated by Lee & Richards, (1971).
	PhiPsi	Phi-, psi-angle of main chain
	Sec_ondary_ **Str**_ucture_	Secondary structure defined by Kabsch-and Sander, (1983) and represented as the alphabet: H, G, I, T, B, E, and S.
Seq_uence_	**PAMatrix**	Amino acid codes represented as the single-letter alphabet.
Feat_ure_	**Surface**	Exposure and polarity defined by sequence
	Hydro_philicity_	Hydrophilicity and isoelectric point defined by sequence
	Shape	Volume, number of chemical bonds, and molecular weight.
	Freq_uency_	Amino acid frequencies in the protein, in a 25-peptide region of the sequence centered at the pentapeptide and in 25Å shell around the pentapeptide;
	Chou _&_ **Fasman**	Chou-Fasman alpha- and beta-structural coefficients;

Prot_ein_	**Dist**_ance_	*Euclid’s distances XY={ PA, PB, PC, PD, PE };*
		Relative distances XYdZ={ APdA, PBdB, PCdC, PDdD, PEdE, PAdP, PBdP, PCdP, PDdP, PEdP }

	Angle	*Plane angles XYZ={ PMA, PMB, PMC, PMD, PME, APC, APD, APE, BPE, CPE, APB, BPC, BPD, CPD, DPE }*
	Thang_ences_	*Plane angles XY-ZU={PM-AC, PM-AD, PM-AE, PM-BE, PM-CE, PM-AB, PM-BC, PM-BD, PM-CD, PM-DE }*
	Twist	*Dihedral angles XYZU={PACM, PADM, PAEM, PBEM, PCEM, PABM, PBCM, PBDM, PCDM, PDEM }*
	Ring	*Dihedral angles XYZU={ PABE, PACD, PACE, PADE, PBCE, PABC, PABD, PBCD, PBDE, PCDE }*
	Round	*Dihedral angles XYZU={ APMC, APMD, APME, BPME, CPME, APMB, BPMC, BPMD, CPMD, DPME }*
Envir_onment_	**Cont**_act_ **Num**_ber_	Number of residues in a 25Å shell around the pentapeptide.

	Dist_ance_	*Euclid’s distances XY={ GA, GB, GC, GD, GE };*
		Relative distances XYdZ={ GAdA, GBdB, GCdC, GDdD, GEdE, GAdP, GBdP, GCdP, GDdP, GEdP }

	Angle	*Plane angles XYZ={ GMA, GMB, GMC, GMD, GME, AGC, AGD, AGE, BGE, CGE, AGB, BGC, BGD, CGD, DGE }*
	Thang_ences_	*Plane angles XY-ZU={ GM-AC, GM-AD, GM-AE, GM-BE, GM-CE, GM-AB, GM-BC, GM-BD, GM-CD, GM-DE }*
	Twist	*Dihedral angles XYZU={ GACM, GADM, GAEM, GBEM, GCEM, GABM, GBCM, GBDM, GCDM, GDEM }*
	Ring	*Dihedral angles XYZU={ GABE, GACD, GACE, GADE, GBCE, GABC, GABD, GBCD, GBDE, GCDE }*
	Round	*Dihedral angles XYZU={ AGMC, AGMD, AGME, BGME, CGME, AGMB, BGMC, BGMD, CGMD, DGME }*

Table 2 cont Grouping of Likeness Profiles used in Fast Searching

Neigh*bors*	**Dist***ance*	*Euclid's distances XY={ LA, LB, LC, LD, LE, RA, RB, RC, RD, RE };*
	Relat*ions*	*Relative distances XYdZ={ ALdA, LBdB, LCdC, LDdD, LEdE, LAdP, LBdP, LCdP, LDdP, LEdP, RAdA, RBdB, RCdC, RDdD, REdE, RAdP, RBdP, RCdP, RDdP, RedP }*
	Angle	*Plane angles XYZ={ LMA, LMB, LMC, LMD, LME, RMA, RMB, RMC, RMD, RME }*
	Sect*ors*	*Plane angles XYZ={ ALC, ALD, ALE, BLE, CLE, ALB, BLC, BLD, CLD, DLE, ARC, ARD, ARE, BRE, CRE, ARB, BRC, BRD, CRD, DRE }*
	Thang*ences*	*Plane angles XY-ZU={LM-AC, LM-AD, LM-AE, LM-BE, LM-CE, LM-AB, LM-BC, LM-BD, LM-CD, LM-DE, RM-AC, RM-AD, RM-AE, RM-BE, RM-CE, RM-AB, RM-BC, RM-BD, RM-CD, RM-DE }*
	Twist	*Dihedral angles XYZU={LACM, LADM, LAEM, LBEM, LCEM, LABM, LBCM, LBDM, LCDM, LDEM, RACM, RADM, RAEM, RBEM, RCEM, RABM, RBCM, RBDM, RCDM, RDEM }*
	Ring	*Dihedral angles XYZU={ LABE, LACD, LACE, LADE, LBCE, LABC, LABD, LBCD, LBDE, LCDE, RABE, RACD, RACE, RADE, RBCE, RABC, RABD, RBCD, RBDE, RCDE }*
	Round	*Dihedral angles XYZU={ ALMC, ALMD, ALME, BLME, CLME, ALMB, BLMC, BLMD, CLMD, DLME, ARMC, ARMD, ARME, BRME, CRME, ARMB, BRMC, BRMD, CRMD, DRME }*
Topol*ogy*	**Dist***ance*	*Euclid's distances XY={ LM, LJ, LR, KR, MR, LK, KM, KJ, MJ, JR, GL, GK, GM, GJ, GR }*
	Angle	*Plane angles XYZ={ LKR, LMJ, LMR, LJR, KMR, LGM, LGJ, LGR, KGR, MGR, LKM, LKJ, KMJ, KJR, MJR, LGK, KGM, KGJ, MGJ, JGR }*
	Thang	*Plane angles XYZ={ MLG, MIR, MIG, MRG, LIG, MPI, MPR, MPG, LPG, IPG, MLI, MLR, LIR, LRG, IRG, MPL, LPI, LPR, IPR, RPG }*
	Twist	*Dihedral angles XYZU={ LKMJ, LKMR, LKJR, LMJR, KMJR, GKJL, GKJM, GKJR, LGMR, KGMJ, MLIR, MLIG, MLRG, MIRG, LIRG, PLRM, PLRI, PLRG, MPIG, LPIR }*

2.2 Store and Group Likeness Profiles

The features introduced in the previous section are grouped into seven categories according to their structural relationship (Table 2).

In a detailed one-on-one conformation search it is one of these seven groupings of feature that is, by default, used in the alignment. Likewise, a default subset of these features is used in the fast 3-D search as shown in column 2 of Table 2. Each of the groupings and their composite features are discussed, however, it should be noted that these defaults sets can be overridden in favor of any combination of the 495 parameters:

Local takes into account the local conformational features of a pentapeptide and is useful for detecting local similarities. The composite features are:

Dist the distances, {*AC, AD, AE, BE, CE, AB, BC, BD, CD* and *DE}* defined for the 5 CA-atoms, *ABCDE*, of a pentapeptide and the distances, {*MA*,*MB*, *MC*, *MD* and *ME*}, between the mass-center, *M*, of the pentapeptide and the CA-atoms of the pentapeptide;

Angle all possible plane angles, {*ABE*, *ACD*, *ACE*, *ADE*, *BCE*, *ABC*, *ABD*, *BCD*, *BDE*, *CDE*}, defined for the pentapeptide *ABCDE* (here: *XYZ* is for the angle between vectors *Y*⟶ *X* and *Y*⟶ *Z*) and all possible plane angles, {*AMC*, *AMD*, *AME*, *BME*, *CME*, *AMB*, *BMC*, *BMD*, *CMD*, *DME*} at the mass-center, *M*, of the pentapeptide *ABCDE*;

Twist all possible dihedral angles, {*ABCD*, *ABCE*, *ABDE*, *ACDE*, *BCDE*} defined for the pentapeptide *ABCDE* (here: *XYZU* is the angle between the 1st plane defined by points *XYZ* and the 2nd plane defined by points *YZU*); all possible dihedral angles, {*MBDA*, *MBDC*, *MBDE*}, that measure the deviation of a pentapeptide conformation from the plane defined by points *MBD*; and torsion angles, {*AMCE* and *BMCD*}, that measure a twist of the pentapeptide around its symmetry axis defined by points *M* and *C*;

PhiPsi phi and psi torsional angles for each residue in the pentapeptide;

Secstr secondary structure values for residues in the pentapeptide defined by Kabsch and Sander (1983).

Protein considers for each pentapeptide analyzed the conformational features that are defined by the center of mass, *P*, of a protein. These features describe how each pentapeptide is located relative to the center of mass of the protein. The following conformational features are used:

Dist the absolute distances, {*PA, PB, PC, PD* and *PE*}, between the protein center of mass, *P*, and each of the CA-atoms of the pentapeptide and the relative distances (set 2 in Table 1) *PXdP* and *PXdX*, that indicate how far a point *X* (where X is *A*, *B*, *C*, *D* or *E*) is located from *P*, relative to an average distance, *dP*, from P to all CA-atoms in the structure and in the comparison with the distance, *dX*, from *X* to all CA-atoms of this protein;

Angle all possible plane angles, {*APC, APD, APE, BPE, CPE, APB, BPC, BPD, CPD, DPE*}, between the protein center of mass, P, and the CA-atoms of the pentapeptide, *ABCD* and *E*;

Twist, Ring, and **Round** dihedral angles, {*PABE, PACD, PACE, PADE, PBCE, PABC, PABD, PBCD, PBDE, PCDE, APMC, APMD, APME, BPME, CPME, APMB, BPMC, BPMD, CPMD, DPME, PACM, PADM, PAEM, PBEM, PCEM, PABM, PBCM, PBDM, PCDM, PDEM*} defined by combinations of four points from the following: protein center of mass *P*, pentapeptide center of mass *M* and CA-atoms *A, B, C, D, E*

Thang plane angles defined by the vector $P \longrightarrow M$, the center of mass of the protein and the pentapeptide, respectively, and by each of the vectors from $\{A\longrightarrow C, A\longrightarrow D, A\longrightarrow E, B\longrightarrow E, C\longrightarrow E, A\longrightarrow B, B\longrightarrow C, B\longrightarrow D, C\longrightarrow D, D\longrightarrow E\}$.

Neighbors considers decamers both preceding and following the pentamer in the amino acid sequence. These features are used to characterize the orientation of the main chain of a pentapeptide relative to adjoining regions. Each of these decamers is described by a single point, the center of mass (*R* for the center of mass preceding the pentapeptide; *L* for the center of mass following the pentapeptide). Conformational likeness features defined by **Neighbors** are the same as those described for **Protein**, except that the protein center of mass *P* is substituted with *R* and *L,* the centers of mass of the decamers preceding and following the pentapeptide, respectively.

Environment considers conformational features that describe each pentapeptide with respect to the residues located in a 25Å-shell around the pentapeptide. That is, these features define a pentapeptide's location relative to its proximal environment in a globular protein. All residues in the shell are described by their centers of mass, *G*. All Environment conformational features are then described in the same way as those in **Protein,** but substituting the protein center of mass, *P,* with *G*.

Topology considers conformational features that describe the pentapeptide by the centers of mass, {*L, M, P, G, R*} and three additional centers of mass *{K, I, J}:* These features are defined to take into account the topological properties of the main-chain fragment centered at the pentapeptide. The three additional mass-centers used here are:

K defined by the 3 CA-atoms *A, B* and *C* of the pentapeptide;

I defined by the 3 CA-atoms *B, C* and *D* of the pentapeptide;

J defined by the 3 CA-atoms *C, D* and *E* of the pentapeptide.

Analogous to the above mentioned **Local** mode, where pentapeptide *ABCDE* and its center of mass *M* have been used, **Topology** mode uses 2 pseudo-pentapeptides and the centers of mass as follows:

- pseudo-pentapeptide *LKMJR* and center of mass *G*;
- pseudo-pentapeptide *MLIRG* and center of mass *P*.

All conformational features defined for the **Local** mode are also calculated for this **Topology** mode**.**

Sequence considers the standard PAM-matrix (Dayhoff, 1978) of "Accepted Point Mutations" that is used to compare aligned amino acid sequences. Residue types at all 5 positions of the pentapeptide are used as a conformational feature. These features describe a "substitution-ability" of side chains in the pentapeptide.

Features takes into account the physical, chemical and statistical features of amino acid residues for each pentapeptide analyzed. Thus **Feature** is defined fully by the amino acid sequence and describes an "optimal conformation" of both main and side chains of a pentapeptide using.

- exposure, polarity, hydrophilicity, isoelectric point, volume, number of chemical bonds, molecular weight, Chou-Fasman alpha- and beta-structural propensities
- observed frequencies of residues in the protein analyzed
- observed frequencies of residues in a 25Å shell centered at the pentapeptide.

2.3 Profile Comparison

This will be reported in detail elsewhere. In short, the method of comparison of conformational likeness values depends on the type of value. For example, PAM matrices are treated differently to local distances. The significance of the difference (or lack thereof) is determined by accumulative probabilities based on comparison to a random sample. The overall likeness is then presented for each polypeptide chain or fragment with each other in the database. Results are normalized so that the starting structure will find itself with a value of 1.0. For detailed comparisons the likeness at each amino acid position is calculated and presented as a color coded comparison. This is best illustrated by example.

3.0 Results

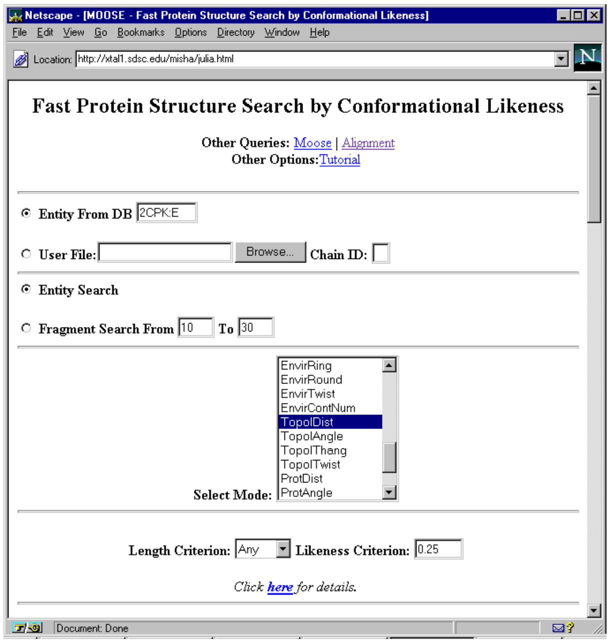
Questions are posed as example searches and the results discussed.

3.1 What structures are conformationally similar to cAMP dependent protein kinase (PDB code 2CPK)?

We have found that **TopolDist** (distances associated with various centers of mass of pentamers but excluding the protein center of mass) is a good parameter to use in finding structures with a similar overall topology. Other **Topol** parameters are not so good - angles are very insensitive; **TopolThang** (as **TopolDist** but for angles) orders the kinases with the highest values and also returns many other structures, (maybe desirable under some circumstances); **TopolTwist** (as **TopolDist** but for dihedral angles) is far too insensitive, finding structures without any apparent structural similarity. A database search with:

- Complete polypeptide chain of 2CPK, designated as E (350 residues)
- Any length of match
- A likeness measure of 0.25 or above (a simple filter to limit the number of hits).

was made with the input window shown in Fig. 1:


Netscape - [MOOSE - Fast Protein Structure Search by Conformational Likeness]
File Edit View Go Bookmarks Options Directory Window Help
Browse...
Document: Done

Location: http://xtal1.sdsc.edu/misha/julia.html

Fast Protein Structure Search by Conformational Likeness

Other Queries: Moose | Alignment
Other Options: Tutorial

● Entity From DB 2CPK:E

○ User File: Chain ID:

● Entity Search

○ Fragment Search From 10 To 30

Select Mode: EnvirRing, EnvirRound, EnvirTwist, EnvirContNum, TopolDist (selected), TopolAngle, TopolThang, TopolTwist, ProtDist, ProtAngle

Length Criterion: Any Likeness Criterion: 0.25

*Click **here** for details.*

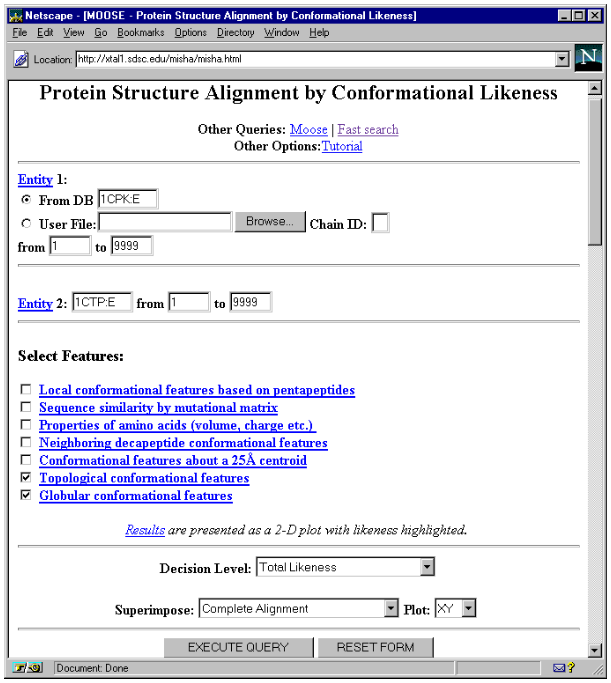
Figure 1 Input Web Form for a Substructure Search.

The search which took approximately one minute of CPU time on a 275Mhz DEC Alpha processor revealed the following 8 structures:

```
1) 0.760 # 1APME # $C-/AMP$-DEPENDENT PROTEIN
KINASE (E.C.
2) 0.863 # 1ATPE # $C-/AMP$-DEPENDENT PROTEIN
KINASE (E.C.
3) 0.717 # 1CDKA # MOL_ID: 1; MOLECULE: CAMP-
DEPENDENT PRO
4) 0.731 # 1CDKB # MOL_ID: 1; MOLECULE: CAMP-
DEPENDENT PRO
5) 0.390 # 1CMKE # CAMP-DEPENDENT PROTEIN
KINASE CATALYTIC
6) 1.000 # 2CPKE # $C-/AMP$-DEPENDENT PROTEIN
KINASE (E.C.
7) 0.362 # 1CSN_ # MOLECULE: CASEIN KINASE-1;
EC: 2.7.1.-;
8) 0.332 # 1CTPE # CAMP-DEPENDENT PROTEIN
KINASE (E.C.2.7.
```

These list of corresponds was the same as that found with DALI, using the same PDB database, with one exception. 1IRK - the tyrosine receptor kinase. It appears **Topology** is too stringent a similarity measure to detect the conformational similarity between these two proteins which shae a common catalytic core, but show distinct variations overall. Also missing are model structures (present in the PDB but purposely ignored here) and the calmodulin binding domain of calmodulin dependent protein kinase (1CDM) - also not detected by DALI, but with a similar catalytic core.

Notice the difference in the likeness values of say 1ATPE(0.863) and 1CTPE(0.332) when compared with 2CPKE (1.000). The difference is the shift between the open and closed conformation that occurs on substrate binding. Both 2CPKE(1.000) and 1ATPE (0.863) are closed conformations, whereas 1CTPE (0.332) is an open conformation. This difference is clearly seen in the detailed alignment by conformational likeness using **Topology** and **Globular** conformational features (Figure 2).


Netscape - [MOOSE - Protein Structure Alignment by Conformational Likeness]
File Edit View Go Bookmarks Options Directory Window Help

Location: http://xtal1.sdsc.edu/misha/misha.html

Protein Structure Alignment by Conformational Likeness

Other Queries: Moose | Fast search
Other Options: Tutorial

Entity 1:
● From DB 1CPK:E
○ User File: Browse... Chain ID:
from 1 to 9999

Entity 2: 1CTP:E from 1 to 9999

Select Features:

□ Local conformational features based on pentapeptides
□ Sequence similarity by mutational matrix
□ Properties of amino acids (volume, charge etc.)
□ Neighboring decapeptide conformational features
□ Conformational features about a 25Å centroid
☑ Topological conformational features
☑ Globular conformational features

Results are presented as a 2-D plot with likeness highlighted.

Decision Level: Total Likeness

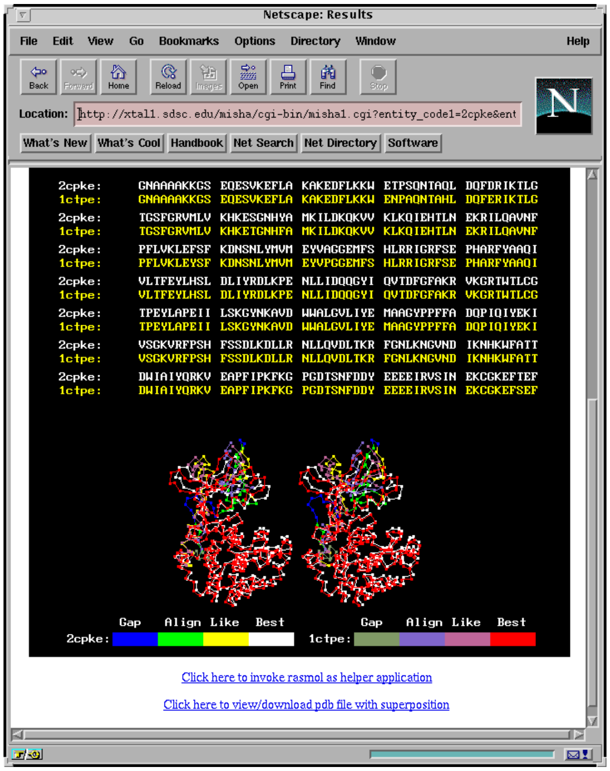
Superimpose: Complete Alignment Plot: XY

EXECUTE QUERY RESET FORM

Document: Done

Figure 2. Input Form for Detailed Structure Alignment.

Notice that the grouping of searchable parameters corresponds to that given in Table 2. Additional parameters which are defined are the decision level and the type of superposition. Decision level defines the cut-off in the probability distribution. Three values are possible: “Total Likeness”, “Total/pentapeptide likeness”, and “Pentapeptide likeness”. They represent, in decreasing order of stringency, the degree of likeness to impose. Possible values for superposition are: “Complete alignment”, “Fragments with positive likeness”, and “Best fragment” and define the parts of the two structures that will be aligned in the display. This is a display-only feature and does not define what fragments are used to make the structure alignment. Those fragments are defined relative to the polypeptide chain of each structure. Part of the results display of this alignment is shown in Fig. 3.


Netscape: Results
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
What's New What's Cool Handbook Net Search Net Directory Software

Location: http://xtal1.sdsc.edu/misha/cgi-bin/misha1.cgi?entity_code1=2cpke&ent

```
2cpke:    GNAAAAKKGS EQESVKEFLA KAKEDFLKKW ETPSQNTAQL DQFDRIKTLG
1ctpe:    GNAAAAKKGS EQESVKEFLA KAKEDFLKKW ENPAQNTAHL DQFERIKTLG
2cpke:    TGSFGRVMLV KHKESGNHYA MKILDKQKVV KLKQIEHTLN EKRILQAVNF
1ctpe:    TGSFGRVMLV KHKETGNHFA MKILDKQKVV KLKQIEHTLN EKRILQAVNF
2cpke:    PFLVKLEFSF KDNSNLYMVM EYVAGGEMFS HLRRIGRFSE PHARFYAAQI
1ctpe:    PFLVKLEYSF KDNSNLYMVM EYVPGGEMFS HLRRIGRFSE PHARFYAAQI
2cpke:    VLTFEYLHSL DLIYRDLKPE NLLIDQQGYI QVTDFGFAKR VKGRTWTLCG
1ctpe:    VLTFEYLHSL DLIYRDLKPE NLLIDQQGYI QVTDFGFAKR VKGRTWTLCG
2cpke:    TPEYLAPEII LSKGYNKAVD WWALGVLIYE MAAGYPPFFA DQPIQIYEKI
1ctpe:    TPEYLAPEII LSKGYNKAVD WWALGVLIYE MAAGYPPFFA DQPIQIYEKI
2cpke:    VSGKVRFPSH FSSDLKDLLR NLLQVDLTKR FGNLKNGVND IKNHKWFATT
1ctpe:    VSGKVRFPSH FSSDLKDLLR NLLQVDLTKR FGNLKNGVND IKNHKWFATT
2cpke:    DWIAIYQRKV EAPFIPKFKG PGDTSNFDDY EEEEIRVSIN EKCGKEFTEF
1ctpe:    DWIAIYQRKV EAPFIPKFKG PGDTSNFDDY EEEEIRVSIN EKCGKEFSEF
```

	Gap	Align	Like	Best
2cpke:	🟦	🟩	🟨	⬜
1ctpe:	▨	▨	▨	🟥

Click here to invoke rasmol as helper application

Click here to view/download pdb file with superposition

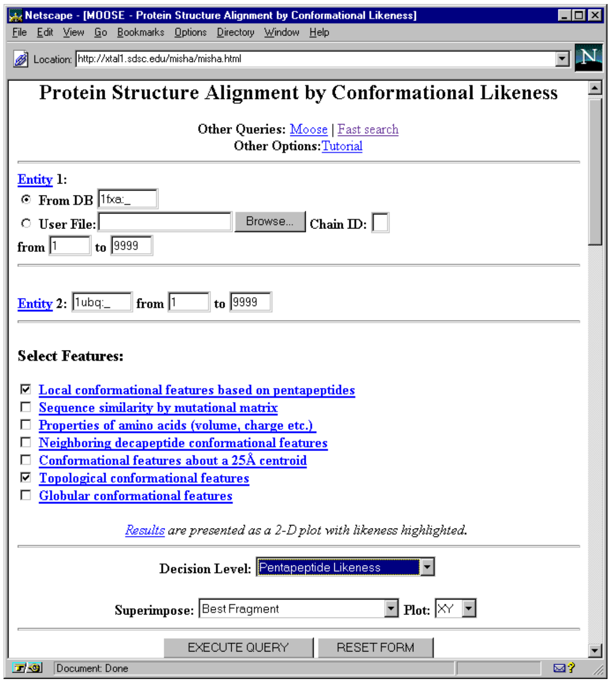
Figure 3. Partial Display of Structure Alignment.

The sequence alignment is given first and this is, of course, merely a reflection of the structure alignment. The structure alignment is displayed in stereo and color coded so that red associated with white represents the best alignment of the two structures. From this display it is evident that differences between open and closed conformations result predominantly from a movement of the small upper lobe, while the larger bottom lobe remains almost unchanged.

3.2 How sensitive is the method in detecting the unusual similarity reported by Vriend and Sander (Proteins 11:52-58, 1991) between ferredoxin (2Fe-2S) and ubiquitin?

Ubiquitin is involved in protein breakdown via covalent conjugates, whereas ferredoxin in an electron carrier in the photoreduction of cytochrome c. That is, there is no apparent functional similarity and no significant sequence similarity between these two structures.

Figure 4 shows the input screen for a local and topological comparison between the two proteins to detect a similarity. The structure superposition is on the best fragment and not the whole structure.



Netscape - [MOOSE - Protein Structure Alignment by Conformational Likeness]

Location: http://xtal1.sdsc.edu/misha/misha.html

Protein Structure Alignment by Conformational Likeness

Other Queries: Moose | Fast search
Other Options: Tutorial

Entity 1:
● From DB 1fxa:_
○ User File: Browse... Chain ID:
from 1 to 9999

Entity 2: 1ubq:_ from 1 to 9999

Select Features:

☑ Local conformational features based on pentapeptides
☐ Sequence similarity by mutational matrix
☐ Properties of amino acids (volume, charge etc.)
☐ Neighboring decapeptide conformational features
☐ Conformational features about a 25Å centroid
☑ Topological conformational features
☐ Globular conformational features

Results are presented as a 2-D plot with likeness highlighted.

Decision Level: Pentapeptide Likeness

Superimpose: Best Fragment Plot: XY

EXECUTE QUERY RESET FORM

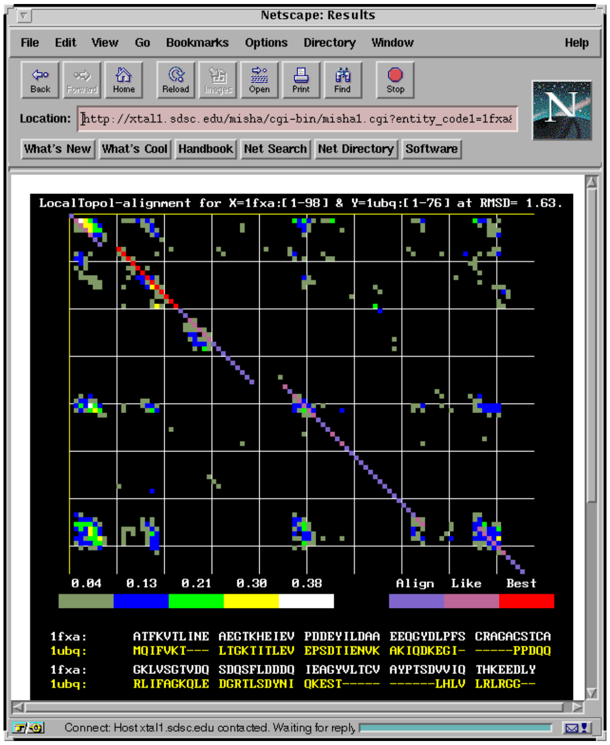
Document: Done

Figure 4 Determining Alignment Between Ubiquitin and Ferredoxin (2Fe-2S).

The results of this alignment are shown in Fig. 5.

Each point on the likeness matrix is color-coded according to the degree of structural similarity at the respective amino acid positions. Thus, white is the strongest similarity and green/blue/grey weaker similarity. The best alignment based on dynamic programming (Needleman and Wunsch, 1970) is shown by the lines from top left to bottom right. Red represents the strongest alignment and corresponds closely to the common fold found by Vriend and Sander. Given the low sequence similarity and apparent lack of common functionality between these two structures this likeness may be attributable to a favorable folding arrangement.

Additional examples can be found in the on-line tutorial pages.



Netscape: Results

Location: http://xtal1.sdsc.edu/misha/cgi-bin/misha1.cgi?entity_code1=1fxa&

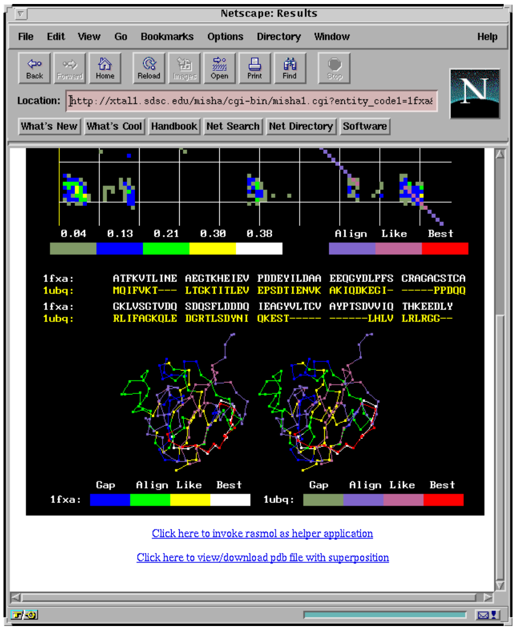
LocalTopol-alignment for X=1fxa:[1-98] & Y=1ubq:[1-76] at RMSD= 1.63.

```
1fxa:   ATFKVTLINE AEGTKHEIEV PDDEYILDAA EEQGYDLPFS CRAGACSTCA
1ubq:   MQIFVKT--- LTGKTITLEV EPSDTIENVK AKIQDKEGI- -----PPDQQ
1fxa:   GKLVSGTVDQ SDQSFLDDDQ IEAGYVLTCV AYPTSDVVIQ THKEEDLY
1ubq:   RLIFAGKQLE DGRTLSDYNI QKEST----- ------LHLV LRLRGG--
```

Connect: Host xtal1.sdsc.edu contacted. Waiting for reply

File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
What's New What's Cool Handbook Net Search Net Directory Software
0.04 0.13 0.21 0.30 0.38
Align Like Best

Figure 5a. Likeness Matrix for Ubiquitin and Ferredoxin (2Fe-2S).



Netscape: Results

Location: http://xtal1.sdsc.edu/misha/cgi-bin/misha1.cgi?entity_code1=1fxa&

```
1fxa:   ATFKVTLINE AEGTKHEIEV PDDEYILDAA EEQGYDLPFS CRAGACSTCA
1ubq:   MQIFVKT--- LTGKTITLEV EPSDTIENVK AKIQDKEGI- -----PPDQQ
1fxa:   GKLVSGTVDQ SDQSFLDDDQ IEAGYVLTCV AYPTSDVVIQ THKEEDLY
1ubq:   RLIFAGKQLE DGRTLSDYNI QKEST----- ------LHLV LRLRGG--
```

Click here to invoke rasmol as helper application

Click here to view/download pdb file with superposition

File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
What's New What's Cool Handbook Net Search Net Directory Software
0.04 0.13 0.21 0.30 0.38
Align Like Best
Gap Align Like Best
1fxa:
Gap Align Like Best
1ubq:

Figure 5b. Sequence and Structure Superposition for Ubiquitin and Ferredoxin (2Fe-2S)

4.0 Discussion

Much more work is needed on the application of this technology, both in interpreting results and in its application.. The large variety of parameters that can be used in determining substructure similarity is both a strength and a weakness. It is a strength for providing a large variety of ways for examining structure similarity; it is a weakness for, in many cases, not providing definitive conclusions about the likeness between two structures. However, this is as much a reflection on nature as it is on the methodology.

References

N.N Alexandrov (1996) *Protein. Engineering* **9**, 727-732.

M. Dayhoff (1978) *Atlas of Protein Sequence and Structure Supp. 3* National Biomedical Research Foundation, Washington DC.

A. Godzik, J. Skolnick, and A. Kolinski (1993) *Protein Engineering* **6**, 801-810.

W.A. Hendrickson (1979) *Acta Cryst*. **A35,** 158-163.

L. Holm and C. Sander (1996) *Nucleic Acids Res.* **24**, 206-209.

L. Holm and C. Sander (1995) *Trends Biochem Sci.* **20**, 478-480.

W. Kabsch and C. Sander (1983) *Biopolymers* **22**, 2577-2637

S. Lee and F. Richards (1971) *J. Mol. Biol.,* **55,** 379-400

K. Mizuguchi and N. Go (1995) *Protein Engineering* **8,** 353-362.

S.B. Needleman and C.D. Wunsch (1970) *J. Mol. Biol.* **48**, 443-453.

R. Nussinov and D.H. Wolfson (1991) *PNAS* **88**, 10495 -10499.

C. Orengo and W. Taylor (1993) *J. Mol. Biol.* **233,** 488-497.

C. Orengo, T. Flores, W. Taylor and J. Thornton (1993) *Protein Engineering* **6**, 485-500.

I.N. Shindyalov and P.E. Bourne (1996) *Acta Cryst. Sup*. 1996, **C78** C-78.

I.N. Shindyalov and P.E. Bourne (1997) *CABIOS* submitted.

W. Taylor, T. Flores and C. Orengo (1994) *Protein Sci.* **3,** 1858-1870.

K. Zhang and D. Eisengberg (1994) *Protein Sci*. **3**, 687-695.

This work was supported by NSF Grants BIR-9507625 and ASC 8902825 and uses equipment donated by Digital Equipment Corp.