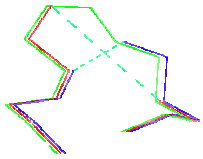
Structural Alignment of Proteins

Thomas Funkhouser
Princeton University
CS597A, Fall 2005

Goal

Align protein structures

```
1  2   3   4   5   6   7   8   9   10  11  12  13  14
PHE ASP ILE CYS ARG LEU PRO GLY SER ALA GLU ALA VAL CYS
PHE ASN VAL CYS ARG THR PRO --- --- --- GLU ALA ILE CYS
PHE ASN VAL CYS ARG --- --- --- THR PRO GLU ALA ILE CYS
```



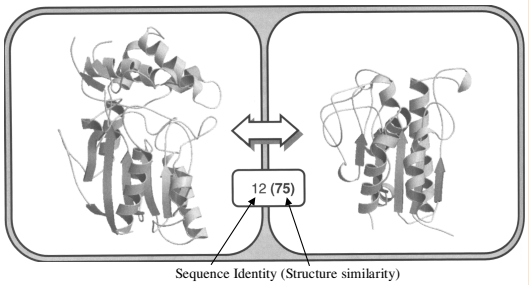
[Marian Novotny]

Terminology

Superposition
- Given correspondences, compute optimal alignment transformation, and compute alignment score

Alignment
- Find correspondences, and then superpose structures

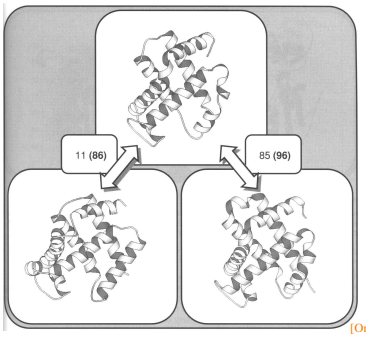
Structure vs. Sequence



Sequence Identity (Structure similarity)

[Orengo04, Fig 6.2]

Structure vs. Sequence

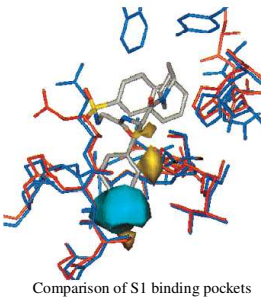


[Orengo04, Fig 6.1]

Applications

Fundamental step in:
- Analysis
- Visualization
- Comparison
- Design

Useful for:
- Structure classification
- Structure prediction
- Function prediction
- Drug discovery



Comparison of S1 binding pockets of thrombin (blue) and trypsin (red).

[Katzenholtz00]

Goals

Desirable properties:

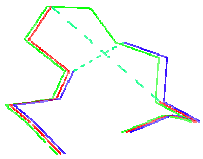
- Automatic
- Discriminating
- Fast

Theoretical Issues

NP-complete problem

- Arbitrary gap lengths
- Global scoring function

```
1  2   3   4   5   6   7   8   9   10  11  12  13  14
PHE ASP ILE CYS ARG LEU PRO GLY SER ALA GLU ALA VAL CYS
PHE ASN VAL CYS ARG THR PRO --- --- --- GLU ALA ILE CYS
PHE ASN VAL CYS ARG --- --- --- THR PRO GLU ALA ILE CYS
```



Methodological Issues

Choices:

- Representation
- Scoring function
- Search algorithm

Methodological Issues

Factors governing choices:

?

Methodological Issues

Factors governing choices:

- Application: homology detection, drug design, etc.
- Granularity: atom, residue, fragment, SSE
- Representation: inter-molecular, intra-molecular
- Scoring: geometric, gaps, chemical, structural, etc.
- Correspondences: sequential, non-sequential
- Gap penalty: expect gaps near loops, etc.
- Flexibility: rigid, flexible
- Target: single protein, representative proteins, PDB

Methodological Issues

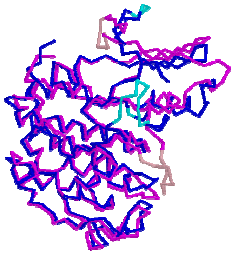
Representations:

- Residue positions
- Local geometry
- Side chain contacts
- Distance matrices (DALI)
- Properties (COMPARER)
- SSEs (SSM, VAST)
- Geometric invariants

Methodological Issues

Scoring functions:

- Distances (RMSD)
- Substitutions
- Gaps



Methodological Issues

Search algorithms:

- Heuristics (CE)
- Monte Carlo (DALI, VAST)
- Dynamic programming (STRUCTAL, SSAP)
- Graph matching (SSM)

Outline

Alignment issues

Example alignment methods ←

Fold prediction experiment

Function prediction experiment

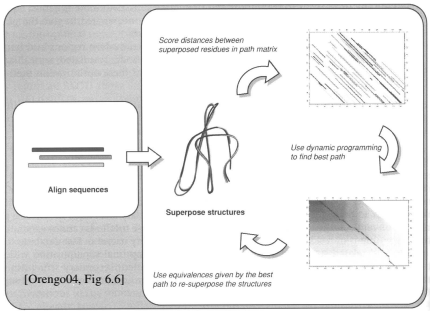
Example Methods

SSAP	Taylor & Orengo, 1989
STRUCTAL	Subbiah, Laurents & Levitt, 1993 Gerstein & Levitt 1998
DALI	Holm & Sander, 1993 Holm & Park, 2000
DEJAVU /LSQMAN	Kleywegt, 1996
CE	Shindyalov & Bourne, 1998
SSM	Krissinel & Henrick, 2003

+ 30 others!

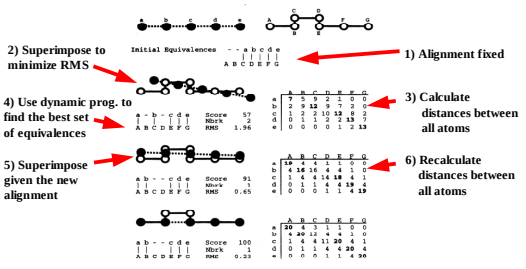
Slide by Rachel Kolodny

STRUCTAL



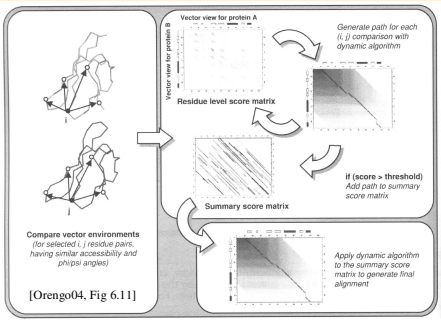
[Subbiah93, Gerstein98]

STRUCTAL



[Subbiah93, Gerstein98]

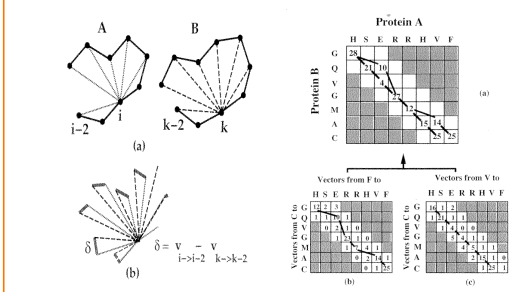
SSAP



[Orengo04, Fig 6.11]

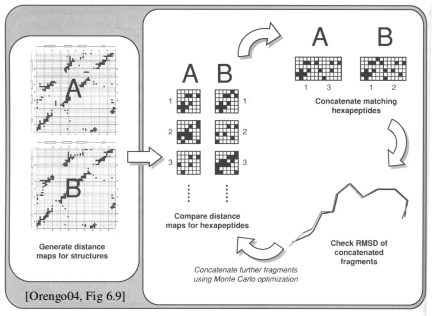
[Orengo96]

SSAP



[Orengo96]

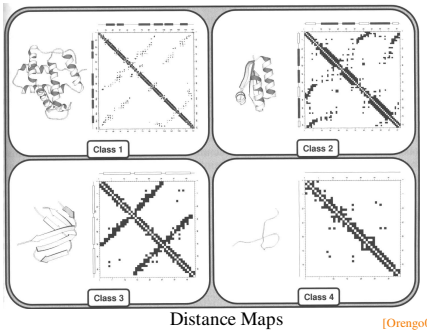
DALI



[Orengo04, Fig 6.9]

[Holm93]

DALI



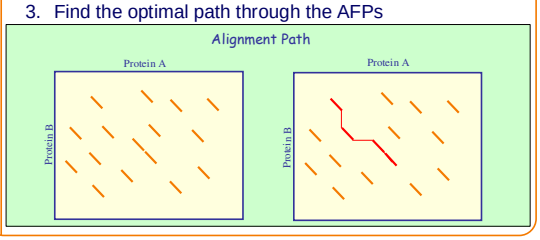
Distance Maps

[Orengo04, Fig 6.7]

CE

Basic steps:

1. Compare octameric fragments to create candidate aligned fragment pairs (AFP)
2. Stitch together AFPs according to heuristics
3. Find the optimal path through the AFPs



SSM

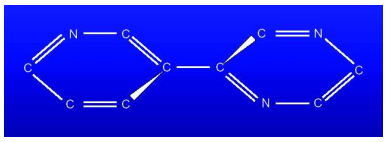
Two-step solution:

1. **Graph representation of structures**
2. **Graph matching**

SSM

Slide by Eugene Krissnel

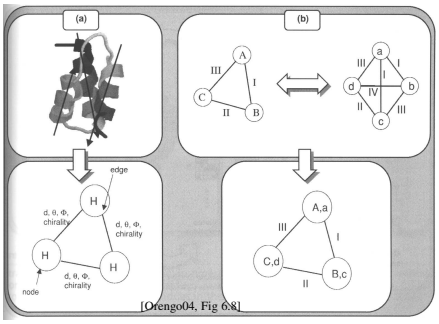
Graph representation of molecular structures



- **Simple and intuitive, however results in intractably large graphs for proteins**
- **Solution: build graphs over stable substructures, such as secondary structure elements (SSEs). Having a correspondence between SSEs, one may use that for the 3D alignment of all core atoms.**

Slide by Eugene Krissnel

SSM



[Orengo04, Fig 6.8]

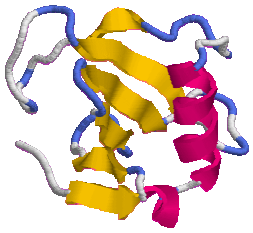
Slide by Eugene Krissnel

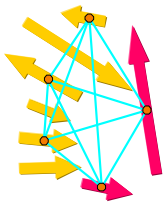
SSM

Slide by Eugene Krissnel

Graph representation of protein SSEs

E. M. Mitchell et al. (1990) J. Mol. Biol. 212:151
A. P. Singh and D. L. Brutlag (1997) ISMB-97 4:284



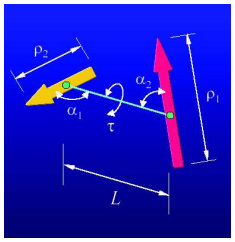


Slide by Eugene Krissnel

SSM

Slide by Eugene Krissnel

Protein graph labeling



Composite label of a vertex

- **type - helix or strand**
- **length r**

Composite label of an edge

- **length *L* (directed if connects vertices from the same chain)**
- **vertex orientation angles a_1 and a_2**
- **torsion angle t**

Vertex and edge labels are matched with thresholds on particular quantities

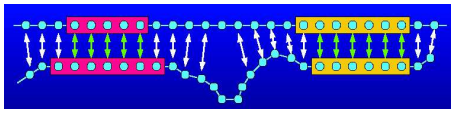
Slide by Eugene Krissnel

SSM

Slide by Eugene Krissnel

C_α alignment

- **SSE-alignment is used as an initial guess for C_α-alignment**
- **C_α-alignment is an iterative procedure based on the expansion of shortest contacts at best superposition of structures**



- **C_α-alignment is a compromise between the alignment length N_a and *r.m.s.d.* The optimised quantity is**

$$Q = \frac{N_a^2}{\left(1 + (r.m.s.d./R_0)^2\right) N_1 N_2}$$

Slide by Eugene Krissnel

SSM

Slide by Eugene Krissnel

Statistical significance of match

- **The overall probability of getting a particular match score by chance is the measure of the statistical significance of the match**

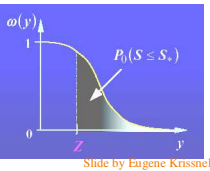
$$P_{value} = 1 - \left(1 - P(S_{N_a}) P(S_{r.m.s.d.}) \prod_{SSE} P(S_{SSE})\right)^{N_{combinations}}$$

- **P_M is traditionally expressed through so-called *Z*-characteristics**

$$P_0 = P(S_{N_a})\, P(S_{r.m.s.d.}) \prod_{SSE} P(S_{SSE})$$

$$P_0 = \int_z^\infty \omega(y)\,dy$$

$$\omega(y) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{1}{2} y^2\right)$$



Slide by Eugene Krissnel

SSM

Slide by Eugene Krissnel

SSM output

- **Table of matched Secondary Structure Elements (SSE alignment)**
- **Table of matched core atoms (C_a - alignment) with dists between them**
- **Rotational-translation matrix of best structure superposition**
- ***R.m.s.d.* of C_a - alignment**
- **Length of C_a - alignment N_a**
- **Number of gaps in C_a - alignment N_g**
- **Quality score *Q***
- **Probability estimate for the match P_M**
- ***Z* - characteristics**
- **Sequence identity**

Slide by Eugene Krissnel

SSM

Slide by Eugene Krissnel

List of matches

Structure Alignment Results

explanation of output

Query: pdb entry **1ldc**, chain A : 478 residues.
L-LACTATE DEHYDROGENASE: CYTOCHROME C OXIDOREDUCTASE 1LDC 3 (FLAVOCYTOCHROME B<2>)
(E.C.1.1.2.3) MUTANT WITH TYR 143 1LDC 4 REPLACED BY PHE (Y143F) COMPLEXED WITH PYRUVATE 1LDC 5

Examined 19295 entries (39511 chains),
Matches 1–14 of 14.

Back to query

reset results

##	Scoring			Rmsd	N_{algn}	N_g	$\%_{seq}$	Query $\%_{sse}$	Target (PDB entry)			
	Q	P	Z						Match	$\%_{sse}$	N_{res}	Title
1	1.00	82.6	27.4	0.00	478	0	100	100	1ldc:A	100	478	L-LACTATE DEHYDROGENASE: CYTOCHROME C OXIDOREDUCTASE 1LDC 3 (FLAVOCYTOCHROME B<2>) (E.C.1.1.2.3) MUTANT WITH TYR 143 1LDC 4 REPLACED BY PHE (Y143F) COMPLEXED WITH PYRUVATE 1LDC 5
2	0.99	62.7	23.8	0.30	478	1	100	100	1lco:A	91	480	L-LACTATE DEHYDROGENASE; 1LCO 4
3	0.98	59.5	23.1	0.41	478	1	100	100	1ltd:A	89	481	FLAVOCYTOCHROME B<2> (E.C.1.1.2.3) COMPLEXED WITH SULFITE 1LTD 3
4	0.94	59.1	23.1	0.56	478	1	100	97	1fcb:A	91	494	FLAVOCYTOCHROME B$b2+ (E.C.1.1.2.3) 1FCB 3
5	0.91	55.4	22.3	0.51	474	2	98	97	1kbi:A	86	504	CRYSTALLOGRAPHIC STUDY OF THE RECOMBINANT FLAVIN-BINDING DOMAIN OF BAKER'S YEAST FLAVOCYTOCHROME B2: COMPARISON WITH THE INTACT WILD-TYPE ENZYME

Slide by Eugene Krissnel

SSM

Slide by Eugene Krissnel

Match details

1LDC X-RAY DIFFRACTION

File Display Colours Options Settings

Match 17 of 22

Back to match list | first match | << | >> | last match

Query PDB 1ldc:A				Alignment			
N_{res}	$\%_{res}$	N_{SSE}	$\%_{SSE}$	Q	P	RMSD	N_{algn}
478	66	31	65	0.514	30.07	1.191	315
L-LACTATE DEHYDROGENASE: CYTOCHROME C OXIDOREDUCTASE 1LDC 3 (FLAVOCYTOCHROME B<2>) (E.C.1.1.2.3) MUTANT WITH TYR 143 1LDC 4 REPLACED BY PHE (Y143F) COMPLEXED WITH PYRUVATE 1LDC 5				$\%_{seq}$	Z	N_{SSE}	N_{gaps}
				33.3	17.30	20	6

view | download | view superposed | view | download

Slide by Eugene Krissnel

SSM

Slide by Eugene Krissnel

SSE alignment

Secondary Structure Alignment

Query PDB 1ldc:A						Target PDB 1huv:A					
10 |H1	13 |A |ASN	124 |LEU	136 |	<->	1 |H1	13 |A |ASN	7 |LEU	19 |
11 |H1	10 |A |THR	137 |SER	146 |	<->	2 |H1	10 |A |PRO	20 |GLY	29 |
12 |H1	10 |A |GLU	151 |ALA	160 |	<->	3 |H1	10 |A |GLU	34 |VAL	43 |
13 |SD	3 |A |PHE	192 |VAL	194 |	<->	7 |SD	3 |A |LEU	75 |ILE	77 |
14 |H1	10 |A |GLY	208 |GLY	217 |	<->	9 |H1	14 |A |LYS	89 |GLY	102 |
15 |SD	4 |A |GLN	225 |SER	228 |	<->	10 |SD	3 |A |PHE	105 |LEU	107 |
16 |H1	8 |A |SER	234 |ALA	241 |	<->	11 |H1	9 |A |SER	114 |CYS	122 |
17 |SD	5 |A |GLN	249 |LEU	253 |	<->	12 |SD	5 |A |LEU	126 |LEU	130 |
18 |H1	15 |A |ASP	258 |LEU	272 |	<->	13 |H1	15 |A |HIS	134 |THR	148 |
19 |SD	4 |A |LEU	277 |THR	280 |	<->	14 |SD	5 |A |THR	152 |THR	156 |
20 |H1	8 |A |ARG	289 |LYS	296 |	<->	15 |H1	7 |A |ARG	165 |ASN	171 |
21 |H1	12 |A |THR	331 |THR	342 |	<->	18 |H1	12 |A |ASN	213 |TRP	224 |
22 |SD	6 |A |ILE	346 |VAL	351 |	<->	19 |SD	7 |A |LYS	227 |LEU	233 |
23 |H1	11 |A |ARG	353 |ILE	363 |	<->	20 |H1	11 |A |SER	235 |GLU	245 |
24 |SD	4 |A |GLY	367 |LEU	370 |	<->	21 |SD	4 |A |GLY	249 |LEU	252 |
25 |H1	16 |A |ALA	383 |ARG	398 |	<->	23 |H1	10 |A |VAL	269 |GLY	278 |
27 |SD	5 |A |GLU	405 |ASP	409 |	<->	24 |SD	3 |A |VAL	281 |ILE	283 |
28 |H1	11 |A |ARG	414 |LEU	424 |	<->	25 |H1	11 |A |ARG	289 |LEU	299 |
29 |SD	4 |A |GLY	428 |LEU	431 |	<->	26 |SD	4 |A |ALA	303 |LEU	306 |
30 |H1	35 |A |GLY	432 |GLY	466 |	<->	27 |H1	35 |A |GLY	307 |GLY	341 |

OCA | SCOP domain | SCOP family
GeneCensus | FSSP | 3Dee | CATH | PDBsum

OCA | SCOP domain | SCOP family
GeneCensus | FSSP | 3Dee | CATH | PDBsum
SWISS–PROT: MDLB_PSEPU , GOX_SPIOL
ProtoMap: MDLB_PSEPU , GOX_SPIOL

view | download sequence | view superposed | view | download sequence

Slide by Eugene Krissnel

SSM

Slide by Eugene Krissnel

Rotation–translation matrix
(to be applied to the query)

-0.710	0.423	-0.563	X		98.778
-0.471	0.309	0.826	× Y	+	96.485
0.524	0.852	-0.020	Z		-59.445

Rotational-translation matrix of best superposition

3D Structural alignment

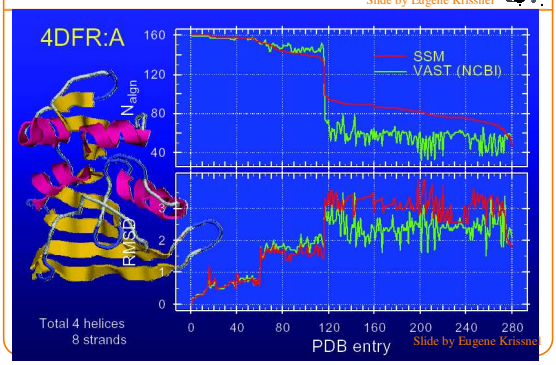
PDB 1ldc:A	Dist. (Å)	PDB 1huv:A
A: GLY 203	1.42 |||	A: LEU 86
A: ASN 204	0.27 |	A: THR 87
A: PSO 205	2.75 |||||	A: GLY 88
A: LEU 206		A: - 88a
A: GLU 207	1.56 |||	A: LYS 89
A: GLY 208	1.66 |||	A: GLY 90
A: GLU 209	0.77 ||	A: ASP 91
A: LYS 210	0.99 ||	A: LEU 92
A: ASP 211	0.91 ||	A: ALA 93
A: VAL 212	0.95 ||	A: LEU 94
A: ALA 213	1.57 |||	A: ALA 95
A: ARG 214	1.65 |||	A: ARG 96
A: GLY 215	1.55 |||	A: ALA 97
A: CYS 216	1.67 |||	A: ALA 98
A: GLY 217	2.01 ||||	A: THR 99
A: GLN 218	2.99 |||||	A: LYS 100
A: GLY 219		
A: VAL 220		
A: THR 221	2.95 |||||	A: ALA 101
A: LYS 222	1.92 ||||	A: GLY 102

C_a - alignment

Slide by Eugene Krissnel

SSM Results

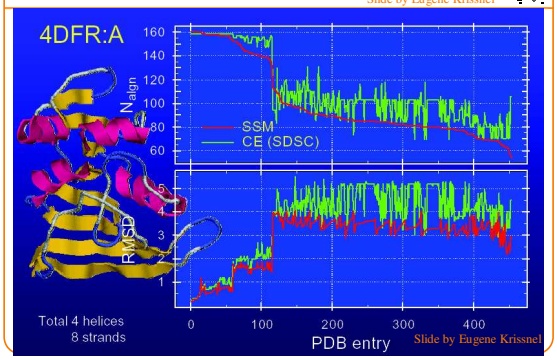
Slide by Eugene Krissnel



Slide by Eugene Krissnel

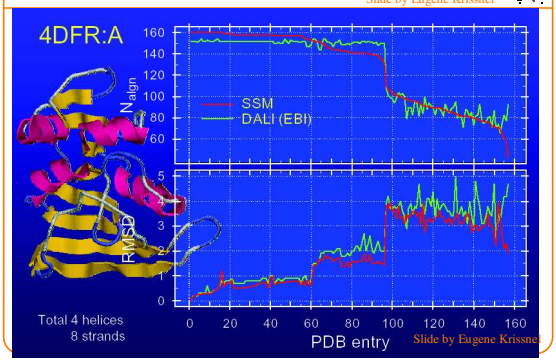
SSM Results

Slide by Eugene Krissnel



SSM Results

Slide by Eugene Krissnel



Outline

Alignment issues

Example alignment methods

Fold prediction experiment ←

Function prediction experiment

Fold Prediction Experiments

Evaluate how useful alignment algorithms are for predicting a protein's fold

How?

Fold Prediction Experiments

Kolodny, Koehl, & Levitt [2005]
- ROC curves and geometric measures using CATH

Sierk & Pearson [2004]
- ROC curves using CATH

Novotny et al. [2004]
- Checked a few dozen cases using CATH

Leplae & Hubbard [2002]
- ROC curves using SCOP

Fold Prediction Experiments

Kolodny, Koehl, & Levitt [2005] ←
- ROC curves and geometric measures using CATH

Sierk & Pearson [2004]
- ROC curves using CATH

Novotny et al. [2004]
- Checked a few dozen cases using CATH

Leplae & Hubbard [2002]
- ROC curves using SCOP

Kolodny, Koehl, & Levitt [2005]

Large scale alignment study

- 2,930 structures (all pairs)
- 6 structural alignment algorithms
- 4 geometric scoring functions
- Evaluation with respect to CATH topology level
- 20,000 hours of compute time

Tested Methods

SSAP	Taylor & Orengo, 1989
STRUCTAL	Subbiah, Laurents & Levitt, 1993 Gerstein & Levitt 1998
DALI	Holm & Sander, 1993 Holm & Park, 2000
DEJAVU /LSQMAN	Kleywegt, 1996
CE	Shindyalov & Bourne, 1998
SSM	Krissinel & Henrick, 2003
Best-of-All	Best of above methods

Slide by Rachel Kolodny

Scoring Functions

Consider # aligned residues & geometric similarity:

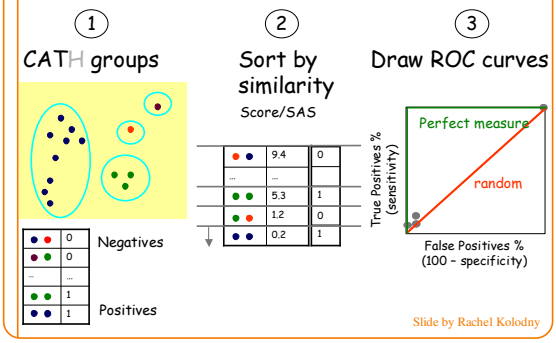
$$SAS = \frac{RMSD \times 100}{N_{mat}}$$

Also penalize gaps:

$$GSAS = \begin{cases} if\,(N_{mat} > N_{gap}) & \frac{RMSD \times 100}{N_{mat} - N_{gap}} \\ else & 99.9 \end{cases}$$

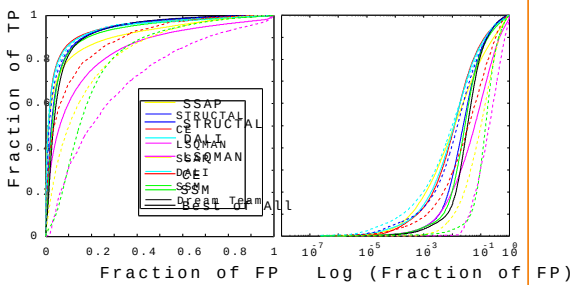
[Kolodny05]

Evaluation Using ROC Curves



Slide by Rachel Kolodny

SAS & Native ROC Curves

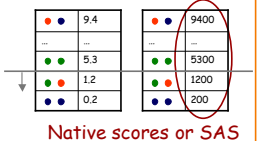


Slide by Rachel Kolodny

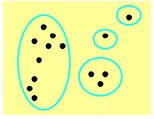
ROC Curve Issues

Uses only internal ordering

- Estimation of similarity can be very wrong

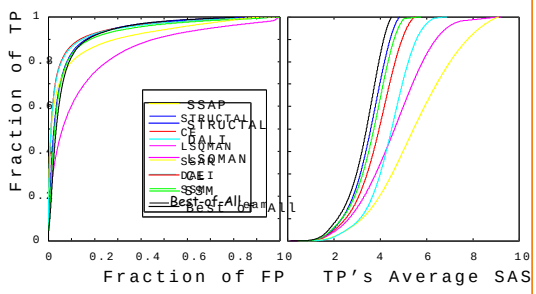


Converts a classification gold standard into binary truth



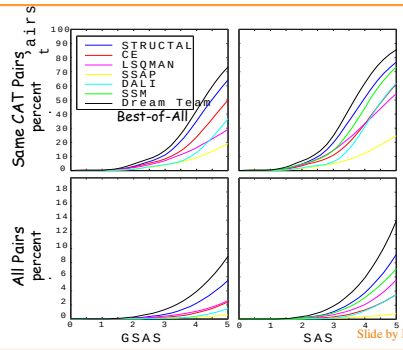
Slide by Rachel Kolodny

Comparing SAS Values Directly



Slide by Rachel Kolodny

GSAS & SAS Distributions



Slide by Rachel Kolodny

Contributions to "Best-of-All"

	Total	SSAP	STRUCTAL	DALI	LSQMAN	CE	SSM
GSAS≤5 Å	275,547 (100%)	832 (0.3%)	**199,071** (69%)	5868 (2.1%)	54,606 (20%)	24,370 (8.8%)	–
SAS≤5 Å	539,755 (100%)	498 (0.09%)	**286,972** (53%)	15,648 (2.9%)	103,408 (19.2%)	15,844 (2.9%)	117,385 (21.8%)
SI≤5 Å	978,531 (100%)	3745 (0.4%)	**497,330** (51%)	24,767 (2.5%)	201,202 (21%)	17,142 (1.8%)	234,345 (24%)
MI≤0.8	880,503 (100%)	4579 (0.5%)	**573,542** (65%)	31,402 (3.6%)	63,088 (7.2%)	72,974 (8.3%)	134,918 (15.3%)

The absolute number of alignments contributed by each method is listed and the percentage of alignments is given in parentheses. The largest contributor is shown in bold.

[Kolodny05]

Outline

Alignment issues

Example alignment methods

Fold prediction experiment

Function prediction experiment ←

Function Prediction Experiment

Evaluate how useful alignment methods are for predicting a protein's molecular function

How?

Data Set

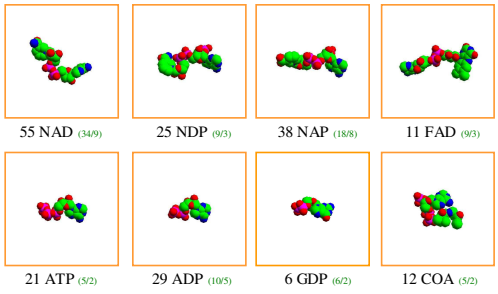
Proteins crystallized with bound ligands
- PDB file must have resolution ≤3 Angstroms
- Ligands must have ≥20 HETATOMS

Classified by reaction/reactant
- PDB file must have an EC number (enzymes only)
- EC number must have a KEGG reaction with a reactant whose graph closely matches ligand in PDB file

Non-redundant
- No two ligands contacting domains with same CATH S95
- No two ligands contacting domains with same SCOP SP
- No two ligands from same PDB file

Data Set

351 proteins / 58 Reactions (189 outliers)



55 NAD (34/9) | 25 NDP (9/3) | 38 NAP (18/8) | 11 FAD (9/3)

21 ATP (5/2) | 29 ADP (10/5) | 6 GDP (6/2) | 12 COA (5/2)

Data Set

REACTION	NAME	#
R00145	NAD	2
R00214	NAD	2
R00342	NAD	7
R00538	NAD	3
R00623	NAD	5
R00703	NAD	5
R01061	NAD	5
R01403	NAD	2
R01778	NAD	3
R00112	NAP	2
R00343	NAP	2
R00625	NAP	2
R00939	NAP	2
R01041	NAP	4
R01058	NAP	2
R01195	NAP	2
R02477	NAP	2
R00703	NAI	2
R00939	NDP	5
R01063	NDP	2
R01195	NDP	2
MISC	NAD	21
MISC	NAP	20
MISC	NAH	2
MISC	NAI	2
MISC	NDP	16

REACTION	NAME	#
R00162	ATP	3
R03647	ATP	2
R00124	ADP	2
R00497	ADP	2
R00756	ADP	2
R01512	ADP	2
R02412	ADP	2
R03647	AMP	2
R00330	GDP	2
R01135	GDP	4
R01130	IMP	3
R02094	TMP	2
R02101	UMP	6
R00965	USP	2
R00966	USP	2
R01229	5GP	2
MISC	ATP	18
MISC	ADP	19
MISC	AMP	10
MISC	A3P	5
MISC	GTP	2
MISC	UDP	4
MISC	UMP	1
MISC	5GP	1

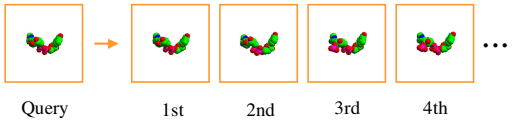
REACTION	NAME	#
R00408	FAD	5
R00924	FAD	2
R01175	FAD	2
MISC	FAD	2
R00351	COA	3
R03552	COA	2
MISC	COA	7
R02961	SAM	3
MISC	SAM	3
R03552	ACO	2
R00291	GDU	2
R03522	GTT	12
R01146	PQQ	3
R00190	PRP	2
R01402	MTA	2
R03435	I3P	2
R02986	CBI	4
R01590	ACD	2
R00529	ADX	2
R03491	SIA	2
R00137	NMN	3
R03992	MYA	2
R03509	137	2
MISC	etc	etc

Evaluation Method

"Leave-one-out" classification experiment

ØMatch every ligand against all the others in data set

- Log a "hit" when best match performs same reaction
- Report percentage of hits (correctly classified ligands)

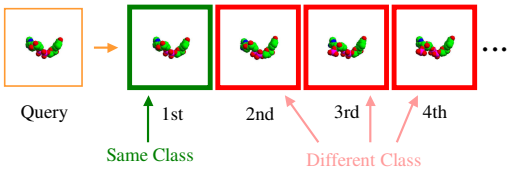


Evaluation Method

"Leave-one-out" classification experiment

ØMatch every ligand against all the others in data set

- Log a "hit" when best match performs same reaction
- Report percentage of hits (correctly classified ligands)



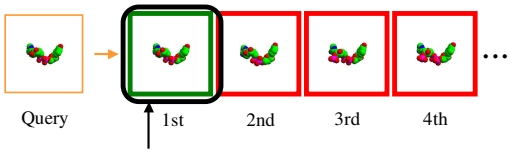
Evaluation Method

"Leave-one-out" classification experiment

- Match every ligand against all the others in data set

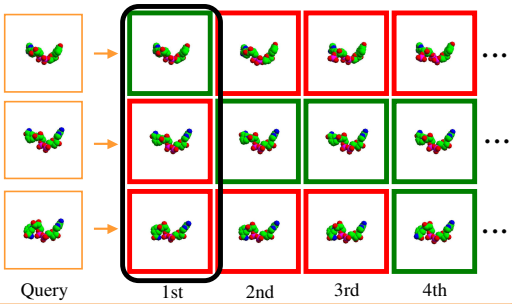
ØLog a "hit" when best match performs same reaction

- Report percentage of hits (correctly classified ligands)



Evaluation Method

Classification rate is 33% is this example



Sequence Alignment Method

Use FASTA to compute Smith-Waterman score for every pair of SCOP domains contacting ligand

```
> fasta34 d1gv0a d1guya

               10        20        30        40        50        60
d1gv0a AGVLDSARFRSFIAMELGVSMQDVTACVLGGHGDAMVPVVKYTTVAGIPVADLISAERIA
       ::::::::::::::::: ::::::: :  ::::::: :::  ::::::::::::::::: :::
d1guya AGVLDAARYRTFIAMEAGVSVEDVQAMLMGGHGDEMVPLPRFSTISGIPVSEFIAPDRLA
               10        20        30        40        50        60

               70        80        90       100       110       120
d1gv0a ELVERTRTGGAEIVNHLKQGSAFYSPATSVVEMVESIVLDRKRVLTCAVSLDGQYGIDGT
       ..:::::: ::.:::: :: :::.:.::.....:::.... :.::::.  :. : ::::::.
d1guya QIVERTRKGGGEIVNLLKTGSAYYAPAAATAQMVEAVLKDKKRVMPVAAYLTGQYGLNDI
               70        80        90       100       110       120

              130       140       150       160
d1gv0a FVGVPVKLGKNGVEHIYEIKLDQSDLDLLQKSAKIVDENCKML
       . :::: :: .:::.: :. :.. .. ::. ::: :
d1guya YFGVPVILGAGGVEKILELPLNEEEMALLNASAKAVRATLDTL
              130       140       150       160

54.487% identity
156 out of 163 amino acids overlap
Smith-Waterman score: 588
```

Sequence Alignment Method

Use FASTA to compute Smith-Waterman score for every pair of SCOP domains contacting ligand

```
> fasta34 d1gv0a d1guya

               10        20        30        40        50        60
d1gv0a AGVLDSARFRSFIAMELGVSMQDVTACVLGGHGDAMVPVVKYTTVAGIPVADLISAERIA
       ::::::::::::::::: ::::::: :  ::::::: :::  ::::::::::::::::: :::
d1guya AGVLDAARYRTFIAMEAGVSVEDVQAMLMGGHGDEMVPLPRFSTISGIPVSEFIAPDRLA
               10        20        30        40        50        60

               70        80        90       100       110       120
d1gv0a E                                                        GIDGT
       .                                                        .....
d1guya Q                                                        GLNDI
               70        80        90       100       110       120

              130       140       150       160
d1gv0a FVGVPVKLGKNGVEHIYEIKLDQSDLDLLQKSAKIVDENCKML
       . :::: :: .:::.: :. :.. .. ::. ::: :
d1guya YFGVPVILGAGGVEKILELPLNEEEMALLNASAKAVRATLDTL
              130       140       150       160

54.487% identity
156 out of 163 amino acids overlap
Smith-Waterman score: 588
```

54.487% identity
156 out of 163 amino acids overlap
Smith-Waterman score: 588

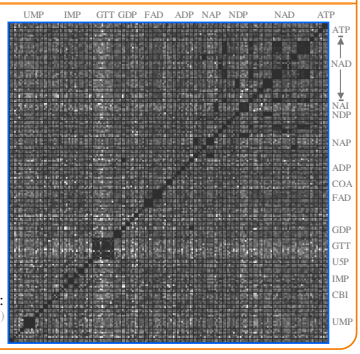
Sequence Alignment Method

Use FASTA to compute Smith-Waterman score for every pair of SCOP domains contacting ligand

$$D(A,B) = 1/\max SmithWaterman(A_i, B_j)_{A_i \in A, B_j \in B}$$

Sequence Alignment Results

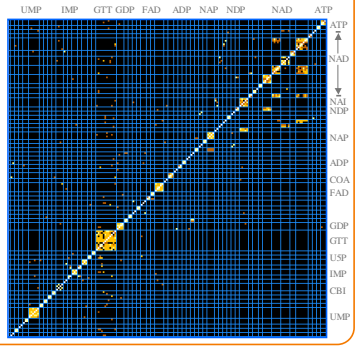
Similarity matrix:



1/SmithWaterman Score:
(Darker means better match)

Sequence Alignment Results

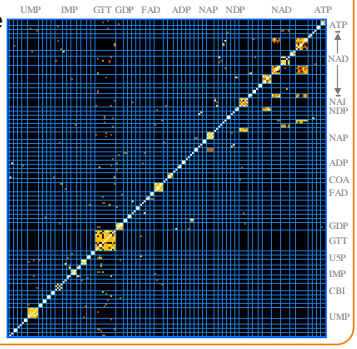
Tier matrix:



Best Matches:
(Beige = Best match)
(Yellow = 1st tier match)
(Orange = 2nd tier match)

Sequence Alignment Results

Classification rate
FASTA = 68%
Random = <1%



Best Matches:
(Beige = Best match)
(Yellow = 1st tier match)
(Orange = 2nd tier match)

Structure Alignment Method

Use CE to compute similarity of protein structures

```
CE - ~/ebi/data/pdbs/1jsu.pdb A ~/ebi/data/pdbs/1hcl.pdb _ scratch
Structure Alignment Calculator, version 1.02, last modified: Jun 15, 2001.
CE Algorithm, version 1.00, 1998
Aligne
Rmsd =
Z-Scor
Gaps =
CPU =
Sequen
```

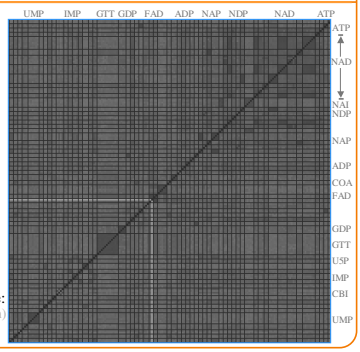
Rmsd = 2.28A
Z-Score = 6.8
Gaps = 30(11.5%)

```
X2 = ( 0.997420)*X1 + ( 0.071548)*Y1 +
     ( 0.005923)*Z1 + (  -93.687386)
Y2 = ( 0.059473)*X1 + (-0.777232)*Y1 +
     (-0.626397)*Z1 + (  119.695427)
Z2 = (-0.040214)*X1 + ( 0.625133)*Y1 +
     (-0.779482)*Z1 + (   84.334198)
```

Image from Shindyalov and Bourne (1998)

Structure Alignment Results

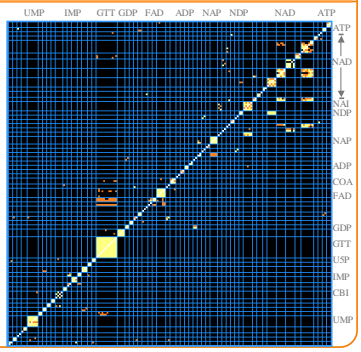
Similarity matrix:



1/CE -Z-Score:
(Darker means better match)

Structure Alignment Results

Tier matrix:



Best Matches:
(Beige = Best match)
(Yellow = 1st tier match)
(Orange = 2nd tier match)

Structure Alignment Results

Classification rate:

FASTA	= 68%
CE	= 65%
Random	= <1%

Structure Alignment Results

Classification rate:

FASTA	= 68%
CE	= 65%
Random	= <1%

When Smith-Waterman ≥ 500:

Sequence	= 80%
CE	= 72%
Random	= <1%

When Smith-Waterman < 500:

CE	= 53%
FASTA	= 44%
Random	= <1%

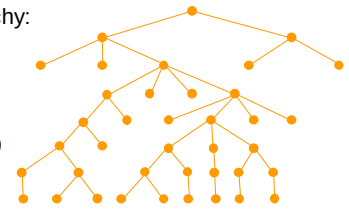
CATH Matching Method

Distance measure is proximity in CATH hierarchy

- D(A,B) = least #levels to common ancestor in hierarchy for any pair of contacting chains

CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S95
- S100



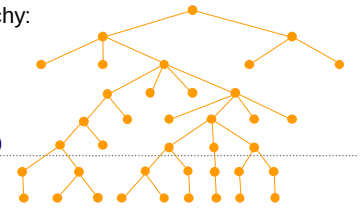
CATH Matching Method

Distance measure is proximity in CATH hierarchy

- D(A,B) = least #levels to common ancestor in hierarchy for any pair of contacting chains

CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S95
- S100



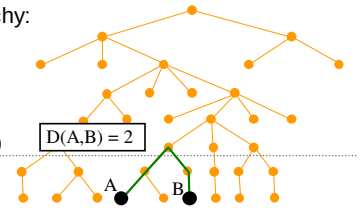
CATH Matching Method

Distance measure is proximity in CATH hierarchy

- D(A,B) = least #levels to common ancestor in hierarchy for any pair of contacting chains

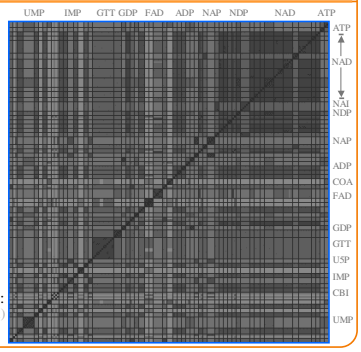
CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S95
- S100



CATH Matching Results

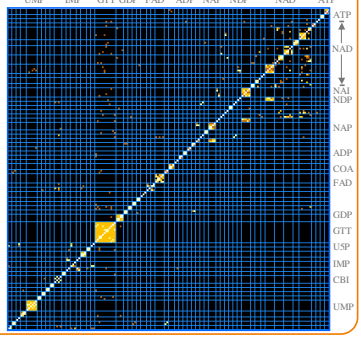
Similarity matrix:



CATH Distance:
(Darker means better match)

CATH Matching Results

Tier matrix:



Best Matches:
(Beige = Best match)
(Yellow = 1st tier match)
(Orange = 2nd tier match)

CATH Matching Results

Classification rate:

FASTA	= 68%
CE	= 65%
CATH	= 58%
Random	= <1%

When Smith-Waterman ≥ 500:

FASTA	= 80%
CE	= 72%
CATH	= 65%
Random	= <1%

When Smith-Waterman < 500:

CE	= 53%
CATH	= 44%
FASTA	= 44%
Random	= <1%

SCOP Matching Results

Classification rate:

FASTA	= 68%
CE	= 65%
SCOP	= 64%
CATH	= 58%
Random	= <1%

When Smith-Waterman ≥ 500:

FASTA	= 80%
CE	= 72%
SCOP	= 72%
CATH	= 65%
Random	= <1%

When Smith-Waterman < 500:

CE	= 53%
SCOP	= 47%
CATH	= 44%
FASTA	= 44%
Random	= <1%

Conclusion

Many algorithms for structural alignment, differing according to

- Application: homology detection, drug design, etc.
- Granularity: atom, residue, fragment, SSE
- Representation: inter-molecular, intra-molecular
- Scoring: geometric, gaps, chemical, structural, etc.
- Correspondences: sequential, non-sequential
- Gap penalty: expect gaps near loops, etc.
- Flexibility: rigid, flexible
- Target: single protein, representative proteins, PDB

None seems best for all situations
All probably provide some benefit over sequence