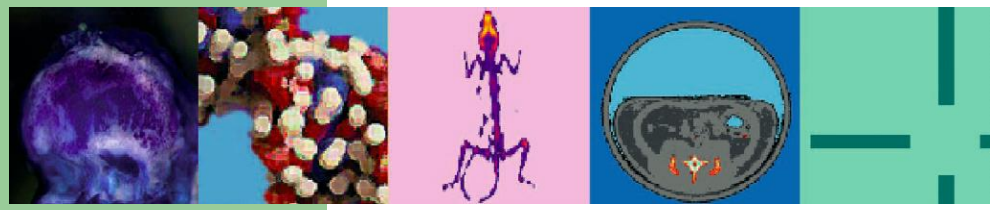
Protein structure description by INVARIANT FEATURES for classification and retrieval in large data bases

Maja Temerinac

Chair for Pattern Recognition and Image Processing
Institute for Computer Science
Albert-Ludwigs-University
Freiburg, Germany



Protein Structure Hierarchy

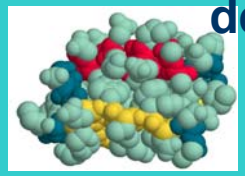
Prologue

4.

quaternary structure

3.

fold

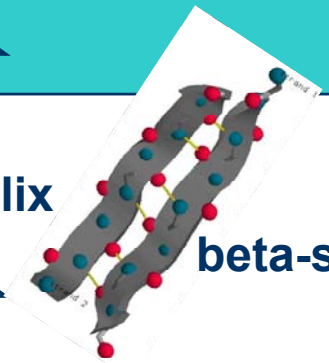


dense packing

tertiary structure

2.

alpha-helix



beta-sheet

secondary structure

1.

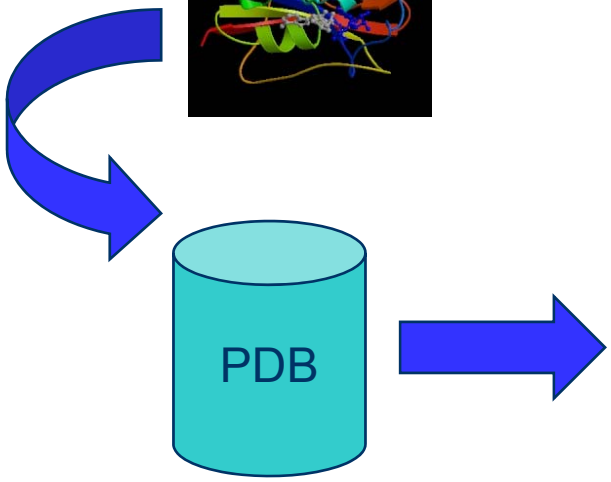
Val Asp Gly Gly Ser His Pro . . .

primary structure

Protein Retrieval

- ***Idea: Perform a structural similarity search***

query protein:1dlr



similarity list:1dlr

Rank	ID		L1-Score	Scop-ID	Class
Query	1dlr	SCP PDB FEAT	-	c.71.1.1	OXIDO-REDUCTASE
1	1dlr	SCP PDB FEAT	0.00	c.71.1.1	OXIDO-REDUCTASE
2	1boz	SCP PDB FEAT	0.96	c.71.1.1	OXIDOREDUCTASE
3	1dls	SCP PDB FEAT	0.98	c.71.1.1	OXIDO-REDUCTASE
4	1s3w	SCP PDB FEAT	1.02	c.71.1.1	OXIDOREDUCTASE
5	1pd8	SCP PDB FEAT	1.03	c.71.1.1	OXIDOREDUCTASE
6	1u72	SCP PDB FEAT	1.04		OXIDOREDUCTASE
7	1hfp	SCP PDB FEAT	1.12	c.71.1.1	OXIDOREDUCTASE
8	1hfr	SCP PDB FEAT	1.12	c.71.1.1	OXIDOREDUCTASE
9	1mvs	SCP PDB FEAT	1.20	c.71.1.1	OXIDOREDUCTASE
10	1pd9	SCP PDB FEAT	1.24	c.71.1.1	OXIDOREDUCTASE

Protein Classification

Query

Protein DB

Feature extraction

Feature DB

Classification Server

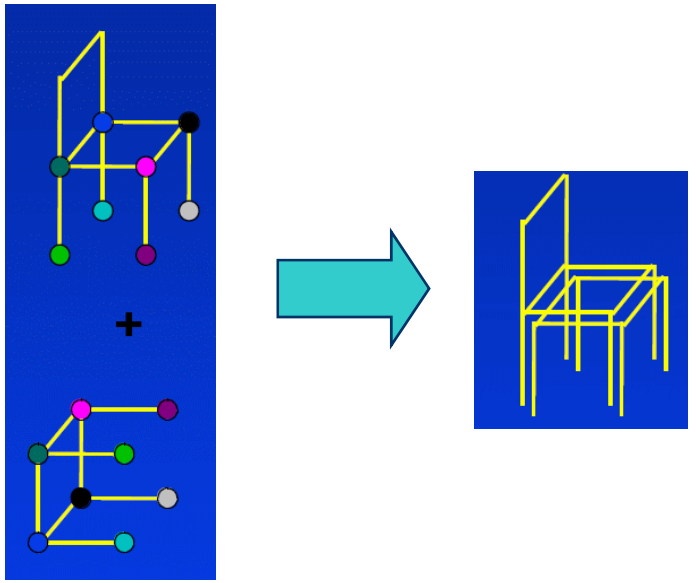
NMR, X-Ray

Feature selection

Evaluation

State-of-the-Art

How to compare two structures?

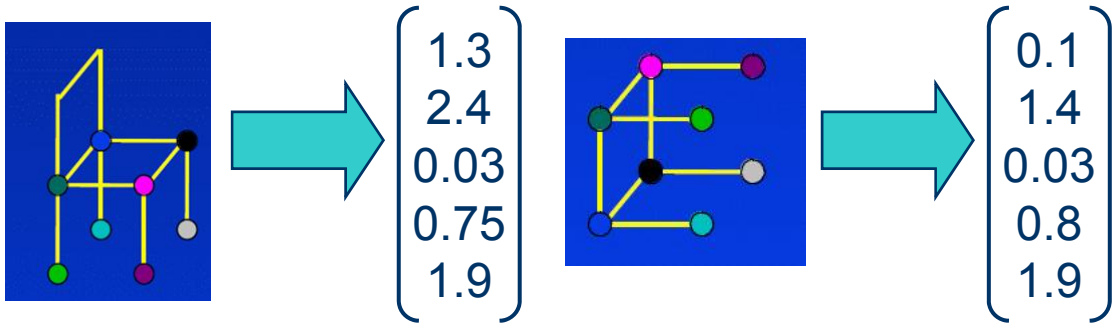


Alignment Methods:

- RMSD: Root Mean Square Error
- CMO: Contact Map Overlap
- DALI: Distance Matrix Alignement

Structural Fingerprint Methods:

- PRIDE: Priority of Identity
- Gauss Integrals



GI Approach

Group Integration for Structural Fingerprints

- Use Invariant Theory to describe the structure
- Combine with Spherical Harmonics and D-Wigner Matrices
- Compare to State-of-the-Art methods

Goal: **Construct a scalable method which provides any wished trade-off between accuracy and complexity**

Incorporating PSD into Group Integration (GI)

- We want to find an invariant function I such that:

$$X_1 \overset{G}{\sim} X_2 \Rightarrow I(X_1) = I(X_2)$$

- We use the Haar-Integral to find an invariant representation for X

$$I_k(X) = \int_G k(gX)\,dg$$

kernel function

In our application

G = Euclidean group

n´=Rn + t

Incorporating PSD into Group Integration (GI)

- choosing the kernel function

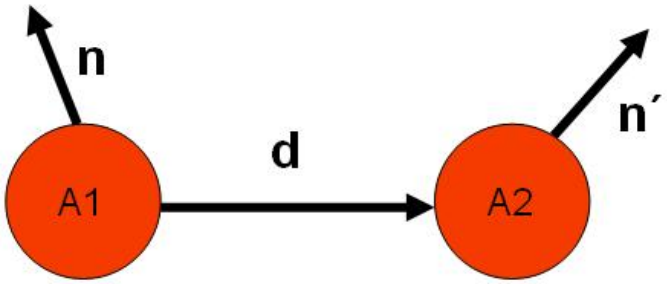
$$k_d(X) = X(0) \cdot X(d)$$

- keeping more information

$$k_d(X, \nabla X) = h_n(\nabla X(0)) \cdot h_{n'}(\nabla X(d))$$

where

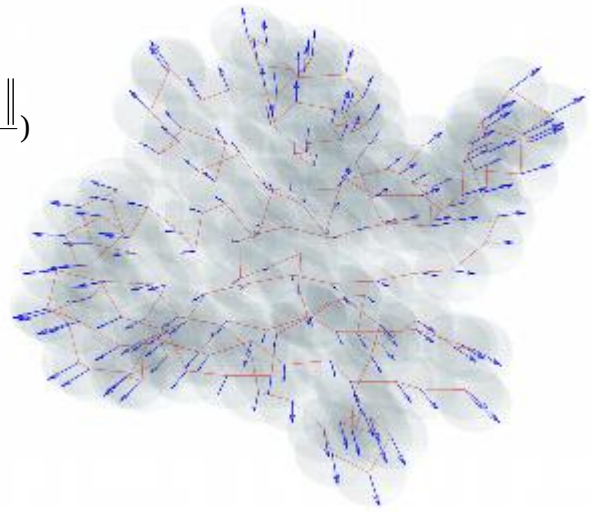
$$h_n(v) = |v| \cdot \delta_1\left(\frac{|v^T n|}{|v|}\right)$$



Incorporating PSD into Group Integration (GI)

- computing the gradient for proteins

$$\nabla X(r) = \sum_i \delta_{u_i}(r) \frac{2}{\sigma^2} \sum_j (u_i - u_j) e^{-2\left(\frac{\|u_i - u_j\|}{\sigma}\right)}$$



- computing the group integral

$$I_\Pi = \sum_{i,k} \theta_\Pi \cdot \delta_d\left(\left|u_i - u_k'\right|\right) \cdot \left|\nabla X(u_i)\right| \cdot \left|\nabla X(u_k')\right|$$

GI Approach

Group Integration Algorithm

1. **Initialize** $I_\Pi = 0$
2. **for** i = 1…n **do**

for k = 1…n **do**

Compute $\Pi = \{\alpha, \beta, \gamma, \Delta\}$

$$\alpha = \frac{\nabla x(u_i)}{|\nabla x(u_i)|} \frac{u_i - u_k}{|u_i - u_k)|}$$

$$\beta = \frac{\nabla x(u_k)}{|\nabla x(u_k)|} \frac{u_i - u_k}{|u_i - u_k)|}$$

$$\gamma = \frac{\nabla x(u_i)}{|\nabla x(u_i)|} \frac{\nabla x(u_k)}{|\nabla x(u_k)|}$$

$$\Delta = |u_i - u_k|$$

Update $I_\Pi \rightarrow I_\Pi + |\nabla x(u_i)| \cdot |\nabla x(u_k)|$

end for

end for

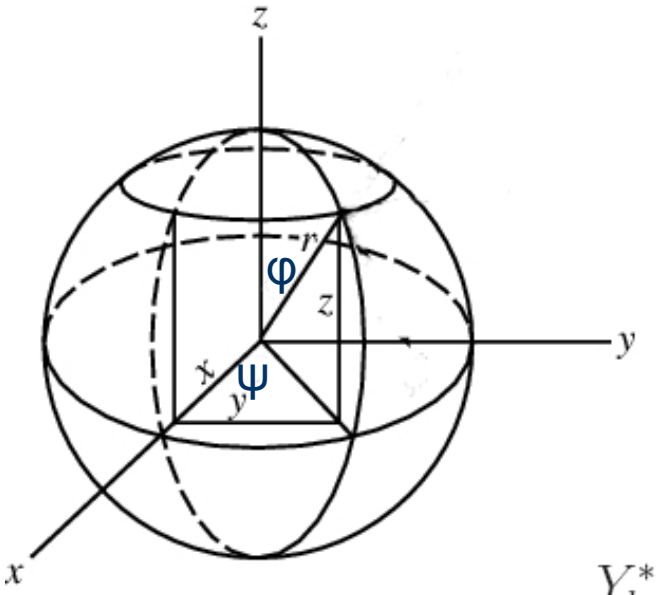
GI Approach

Extending GI with Spherical Harmonics

- Definition of spherical harmonics

f is defined on a 2-sphere

$$\mathbf{f}(\phi,\psi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_m^l Y_m^l(\phi,\psi)$$



$$a_m^l - \int_{S^2} f(\phi,\psi) Y_{lm}^*(\phi,\psi) d\Omega$$

φ - colatitudinal
Ψ - longitudinal

$$Y_{lm}^*(\phi,\psi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} \cdot P_l^m(\cos\phi) \cdot e^{im\psi}$$

associated Legendre polynomial

Spherical Harmonics Algorithm

1. **Initialize** $I_{\Pi} = 0$

2. **for** i = 1…n **do**

   for k = 1…n **do**

   Compute $\Pi = \{\alpha, \beta, \gamma, \Delta\}$

   Update $I_{\Pi}^{lm} \rightarrow I_{\Pi}^{lm} + Y_m^l \left(\frac{u_i - u_k}{|u_i - u_k|} \right) \cdot |\nabla x(u_i)| \cdot |\nabla x(u_k)|$

   end for

   end for

3. **Make invariant** $I_{\Pi}^{l} = \sum_{m=-l}^{l} \left| I_{\Pi}^{lm} \right|^2$

Data Set Overview

sccs
?.1.1.1
a.?.?.?

dataset	# of domains	classification level	# of classification classes
all-classes	2,650	SCOP-class	7
all-alpha	3,680	SCOP-fold	172
27fold	685	SCOP-fold	27
cath	20,937	CATH-homology	2147

Data&Results

Experiments with GI features

Results SCOP classes and folds

Feature	1NN	1T	2T	EM	DCG
noSH	99.8	86.8	91.4	13.4	96.7
SH	99.8	87.6	92.5	13.4	97.2
D-Wigner	99.5	86.1	89.9	13.3	96.3

Table 6.3: **Results 'all-classes'.** Results on the 'all-classes'-dataset with GI, SH and D-Wigner features.

Feature	1NN	1T	2T	EM	DCG
noSH	97.4	84.8	88.6	35.6	94.4
SH	97.8	89.3	92.2	37.4	96.0
D-Wigner	97.4	87.5	90.4	36.8	95.2

Table 6.4: **Results 'all-alpha'.** Results on the 'all-alpha'-dataset with GI, SH and D-Wigner features.

Classification into classes is better than classification into folds

Experiments with GI features

Results 27folds data set

Feature	1NN	1T	2T	EM	DCG
noSH	77.3	31.0	41.2	27.2	67.9
SH	78.8	32.4	44.7	28.7	69.3
Dwigner	77.8	29.5	39.1	26.2	66.8

Table 6.5: **Results '27fold'.** Results on the 'all-classes'-dataset with GI, SH and D-Wigner features.

- Difficult for classification
- SH improve the results for 1.5%
- D-Wigner are worse than SH

Data&Results

Comparison to State-of-the-Art methods

Group Integrals vs. DALI (Alignment)

Feature	1NN	1T	2T	EM	DCG
SH	78.8	32.4	44.7	28.7	69.3
DALI	85.1	59.1	67.8	45.0	82.8

Table 6.13: **Comparison of results with DALI.** Comparison of the results on the '27folds'-dataset computed by DALI and by the new method.

DALI is better for 6.3%

Time consumption!

SH ~ 2 min

DALI ~ 1 week

Comparison to State-of-the-Art methods

Group Integrals vs. PRIDE (Structural Fingerprint)

dataset	Feature	1NN	1T	2T	EM	DCG
all-classes	SH	99.8	87.6	92.5	13.4	97.2
all-classes	PRIDE	99,7	84.8	88.2	13.3	96
all-alpha	SH	97.8	89.3	92.2	37.4	96.0
all-alpha	PRIDE	96.8	80.7	85	34.3	92.7
27folds	SH	78.8	32.4	44.7	28.7	69.3
27folds	PRIDE	70.7	29.4	38.9	25.9	65.1
cath	SH	98.9	72.6	77.7	41.2	91.1
cath	PRIDE	98.8	66.8	73.2	39.1	88.8

Table 6.14: **Comparison with PRIDE features.** Comparison of the results on the '27folds'-dataset computed by PRIDE and by the new method.

On the 27folds data set SH are better by 8.1%

Comparison to State-of-the-Art methods

Group Integrals vs. Gauss Integrals (Structural Fingerprint)

dataset	Feature	1NN	1T	2T	EM	DCG
all-classes	SH	99.8	87.6	92.5	13.4	97.2
all-classes	Gauss	99.2	73.3	81.2	12.1	93.6
all-alpha	SH	97.8	89.3	92.2	37.4	96.0
all-alpha	Gauss	94.2	63.8	72.9	29.5	87.0
27folds	SH	78.8	32.4	44.7	28.7	69.3
27folds	Gauss	67.6	26.1	35.5	23.2	63.3
cath	SH	98.9	72.6	77.7	41.2	91.1
cath	Gauss	98.4	69.8	76.4	40.2	90.0

Table 6.15: **Comparison with Gauss Integrals.** Comparison of the results on the '27folds'-dataset computed by Gauss Integrals and by the new method.

On the 27folds data set SH are better by 11.2%

Time requirements

dataset	size	Time
27folds	685	2min
all-classes	2,650	40 min
all-alpha	3,680	1h
cath	20,937	2h

Table 6.16: **Time requirements new method.** Time requirements of the new method on different datasets.

Method	Time
New Method	2 min
PRIDE	2min
Gauss	2min
DALI	1 week

Table 6.17: **Comparison of time requirements.** Comparison of the time requirements on the '27folds' dataset with different methods.

Summary

- ***Introduced automatic structural classification for proteins***
- ***Found a good set of features for the protein structure***
- ***Comparison with DALI:***
  8 % lower accuracy in classification
  1000 times faster computation time
- ***Comparison with PRIDE and Gauss:***
  10 % higher accuracy in classification
  same computation time
- ***Appropriate for fast pre-classification***

Supplementary Slides

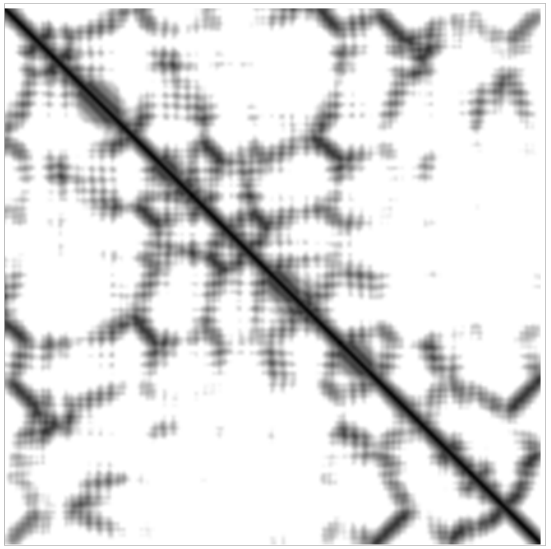
State-of-the-Art

Distance Matrix D_ij

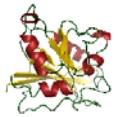
- Protein Retrieval & Classification by Distance Matrices
- Distance between **Cα-atoms (Angstrom A°)**

	$C\alpha^1$	$C\alpha^2$	$C\alpha^3$	$C\alpha^4$
$C\alpha^1$	0	10	20	15
$C\alpha^2$	10	0	12	30
$C\alpha^3$	20	12	0	3
$C\alpha^4$	15	30	3	0

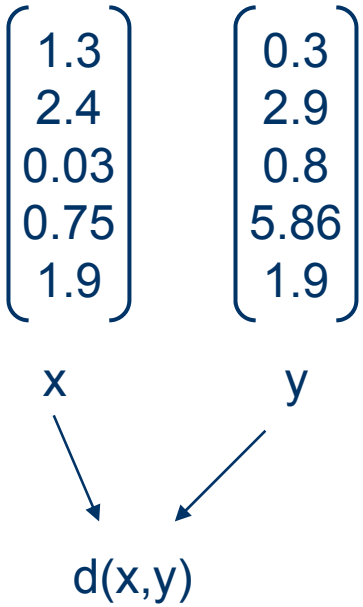
Example: Distance matrix of protein with 4 **Cα-atoms** .



Example: Distance matrix of 1dlr protein with 186 **Cα-atoms**.



Computing the Distance



Distance measure	abbreviation	Formula
Manhattan Distance	L1	$d_{L1}(\mathrm{x},\mathrm{y}) = \sum_{i=0}^{n} \lvert x_i - y_i \rvert$
Euclidean Distance	L2	$d_{L2}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$
Maximum Distance	L_∞	$d_\infty(\mathrm{x},\mathrm{y}) = \max_i \lvert x_i - y_i \rvert$

Table 5.1: The L-distance measures used for feature vector comparison.

Distance measure	abbreviation	Formula
χ_1^2- Distance	χ_1^2	$d_{\chi_1^2}(x,y) = \sum_{i=0}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}$
χ_2^2- Distance	χ_2^2	$d_{\chi_2^2}(x,y) = \sum_{i=0}^{n} \frac{(x_i - y_i)^2}{x_i}$

Table 5.2: The χ^2-distance measures used for feature vector comparison.

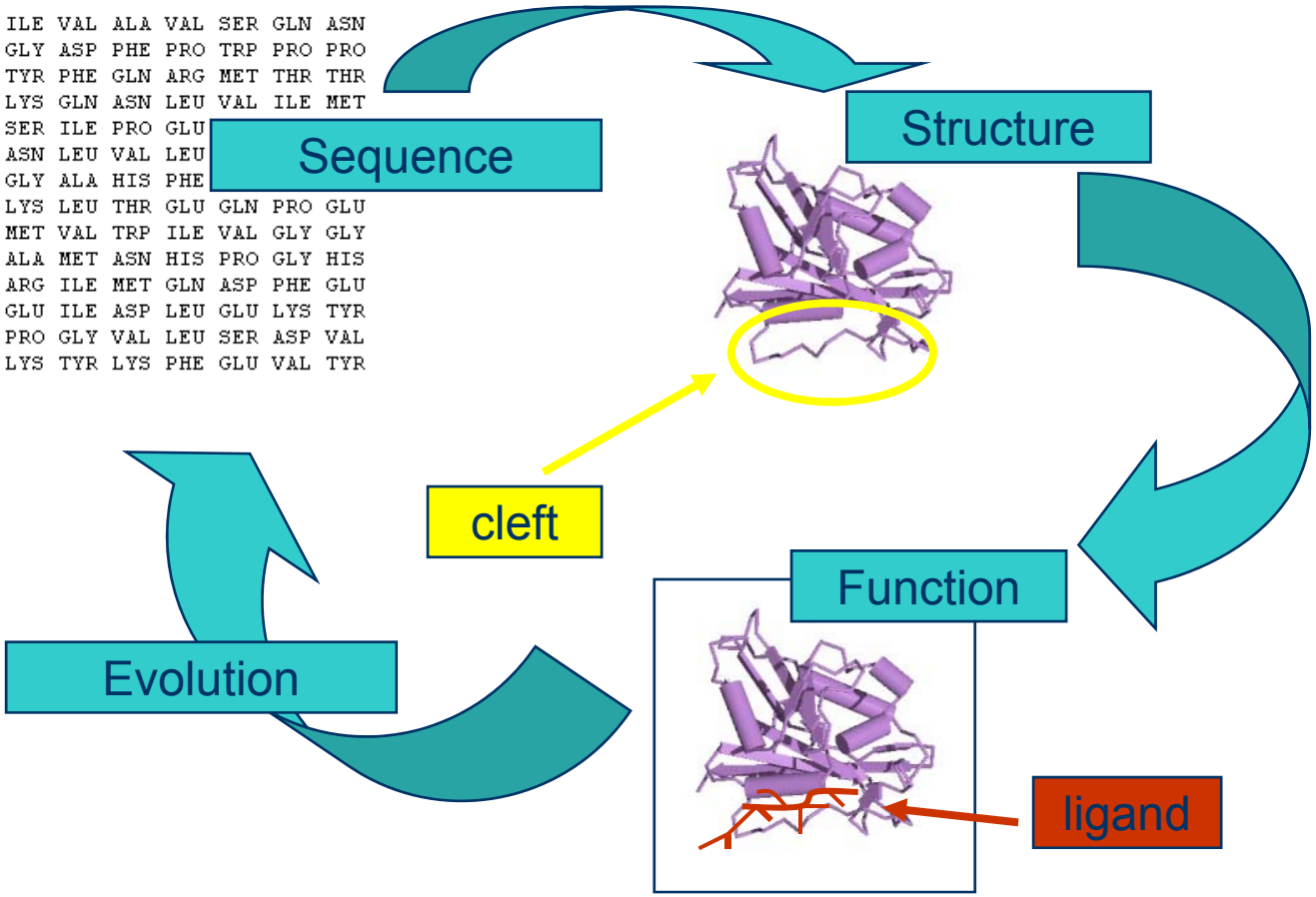
Princeton Shape Benchamark

- Standard for evaluation of retrieval for 3D objects
- 5 statistical measures

• Nearest Neighbor	the percentage of the closest matches that belong to the same class as the query
• First Tier	the percentage of models in the query's class that appear within the top K matches, where K depends on the size of the query's class. Specifically, for a class with |C| members, K = |C| − 1 for the first tier, and K = 2 (|C| − 1) for the second tier.
• Second Tier	
• E-Measure	a composite measure of the precision and recall for a fixed number of retrieved results
• Discounted Cumulative Gain	results near the front of the list weigh more than correct results later in the ranked list

Cycle of Life

```
VAL GLY SER LEU ASN CYS ILE VAL ALA VAL SER GLN ASN
MET GLY ILE GLY LYS ASN GLY ASP PHE PRO TRP PRO PRO
LEU ARG ASN GLU PHE ARG TYR PHE GLN ARG MET THR THR
THR SER SER VAL GLU GLY LYS GLN ASN LEU VAL ILE MET
GLY LYS LYS THR TRP PHE SER ILE PRO GLU
PRO LEU LYS GLY ARG ILE ASN LEU VAL LEU
LEU LYS GLU PRO PRO GLN GLY ALA HIS PHE
SER LEU ASP ASP ALA LEU LYS LEU THR GLU GLN PRO GLU
LEU ALA ASN LYS VAL ASP MET VAL TRP ILE VAL GLY GLY
SER SER VAL TYR LYS GLU ALA MET ASN HIS PRO GLY HIS
LEU LYS LEU PHE VAL THR ARG ILE MET GLN ASP PHE GLU
SER ASP THR PHE PHE PRO GLU ILE ASP LEU GLU LYS TYR
LYS LEU LEU PRO GLU TYR PRO GLY VAL LEU SER ASP VAL
GLN GLU GLU LYS GLY ILE LYS TYR LYS PHE GLU VAL TYR
GLU LYS ASN ASP
```



GI Approach

Extending GI with D-Wigner Matrices

- Definition of D-Wigner Matrices:

R is a Rotation Matrix

$$f(R) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sum_{m'=-l}^{l} b^l_{m,m'} (D^l_{m,m'}(R))$$

f is defined on SO(3)

$$b^l_{m,m'} = \frac{2l+1}{8\pi^2} \langle D^l_{m,m'}, f \rangle$$

$$\langle X, X' \rangle = \int_{SO(3)} X^*(g) X'(g) dg$$

GI Approach

- D-Wigner Algorithm

1. **Initialize** $I_{\Pi} = 0$

2. **for** i = 1…n **do**

   for k = 1…n **do**

   Compute $\Pi = \{\alpha, \beta, \gamma', \Delta\}$ **and** $R = MV^{-1}_{\alpha,\beta,\gamma'}$

   Update $I_{\Pi}{}^{l} \rightarrow I_{\Pi}{}^{l} + \left|\nabla x(u_i)\right| \cdot \left|\nabla x(u_k)\right| \cdot D^{l}(R)$

   end for

   end for

3. **Make invaraint: Take norm of each column of the matrix.**

Default Parameter Set Overview

Gradient computation	σ	400
Coordinate Distance Scaling	DScale	0.02
Sequence Distance Scaling	SeqDScale	40
Histogram Bin Dimension	hist Π	[16,2,2,8]
Spherical Harmonics Coefficient	l_{sharm}	1
D-Wigner Matrix Coefficient	l_{dwig}	1

Outlook

- ***Application of GI for clustering***
- ***Improvement of accuracy by moderate increasing of computation time:***
  - •Use chemical information (hydrophobicity)
  - •Use other atoms besides Cα-atoms
  - •Include secondary structure information
- ***Find an algorithm for domain definition***
- ***Classify structures which were not yet published***