

# Multiple Geometrical Alignments for the Analysis of Protein Active Sites

Thomas Fober   Eyke Hüllermeier

Knowledge Engineering & Bioinformatics Group  
Mathematics and Computer Science Department

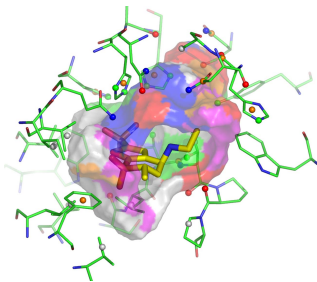
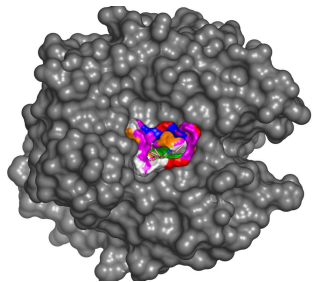
Philipps



Universität  
Marburg



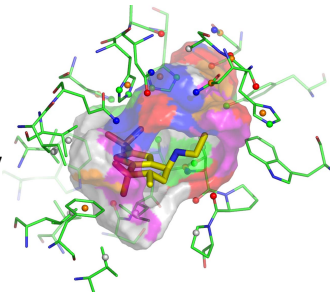
# Protein active sites



- small cavities on the surface of a protein
- protein consists of amino acids; surface of active site, too
- amino acids define physico-chemical properties
  - ▶ predefined rules:  $\mathcal{L} = \{\text{donor, aliphatic, } \dots\}$ ,  $|\mathcal{L}| = 7$
- abstraction: summarize patch into a spatial point

# Why we are interested in active sites?

- small molecules bind to these sites and cause a reaction of the protein
- pharmaceutical chemistry is interested in active sites and similarity measures between them
  - ▶ inhibit active site to cause or suppress a reaction of the protein



# Applications

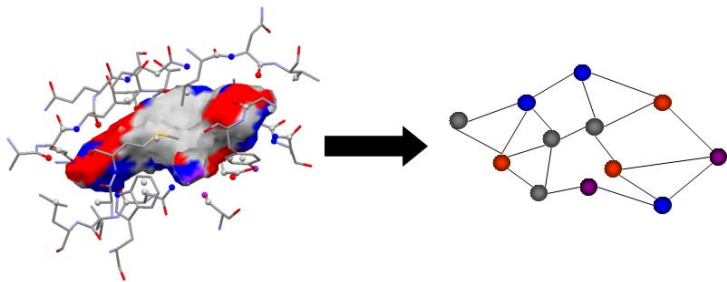
## 1 cross reactivities

- ▶ active site of target is known
- ▶ search for proteins with similar active site
- ▶ these proteins may also be influenced by molecule

## 2 prediction of the function

- ▶ protein with unknown function but known structure
- ▶ search for proteins with known function and similar active site

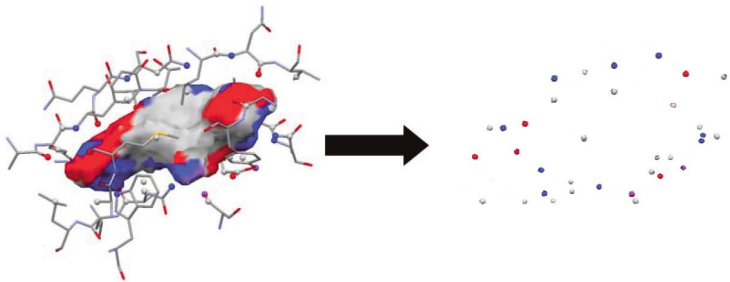
# Graph-representation of protein active sites



## Definition (graph)

$G = (V, E, l_V, l_E)$  is a node-labeled and edge weighted graph, where  $V$  is a finite set of nodes and  $E \subseteq V \times V$  a set of edges, and where  $l_V : V \rightarrow \mathcal{L}_V$  and  $l_E : E \rightarrow \mathbb{R}$  are functions that assign labels and weights, respectively.

# Pointcloud-representation of protein active sites

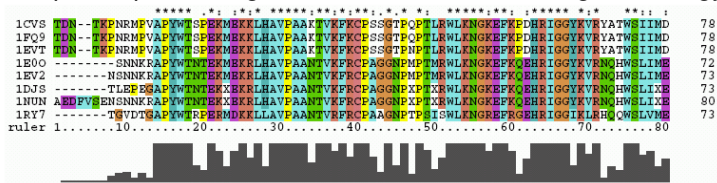


## Definition (labeled geometrical point cloud)

a labeled geometrical point cloud is a set  $P$  of  $n$  points  $p_i$ ,  $i = 1, \dots, n$ , with associated functions  $\ell : P \rightarrow \mathcal{L}$  and  $c : P \rightarrow \mathbb{R}^3$

# Objectives

- 1 define an appropriate similarity measure
  - ▶ theory formation in biology founded on similarity-based and analogical reasoning principles
- 2 calculate a geometrical alignment
  - ▶ multiple sequence alignment leads to break through in biology



- ▶ functional similarity has no strong correspondence to sequence similarity
- ▶ transfer concept of sequence alignment to point clouds
- ▶ so far structural alignments on graph representation



# Similarity of point clouds: intuition

- two labeled point clouds are similar if they can be spatially superimposed
  - ▶ fix position of the first cloud
  - ▶ move the second cloud
  - ▶ no change of the internal arrangement of points
- for each point in one of the structures, there exists a point in the other structure
  - ▶ same label
  - ▶ spatially close

# Similarity of point clouds: formal

- two point clouds

- ▶  $A = \{(a_1, \ell(a_1)), \dots, (a_n, \ell(a_n))\}$

- ★ where  $a_i = (a_{i,1}, a_{i,2}, a_{i,3}) \in \mathbb{R}^3$  and  $\ell(a_i) \in \mathcal{L}$

- ▶  $B = \{(b_1, \ell(b_1)), \dots, (b_m, \ell(b_m))\}$

- ★ where  $b_i = (b_{i,1}, b_{i,2}, b_{i,3}) \in \mathbb{R}^3$  and  $\ell(b_i) \in \mathcal{L}$

- $\mathcal{X}_1 = \mathcal{X}_2 \Leftrightarrow (\mathcal{X}_1 \subseteq \mathcal{X}_2) \wedge (\mathcal{X}_2 \subseteq \mathcal{X}_1)$

- similarity between labeled point clouds

$$\text{SIM}(A, B) = \min\{\text{INC}(A, B), \text{INC}(B, A)\}$$

## Inclusion: point cloud $B$ in point cloud $A$

- $B \subseteq A: \forall y \in B \Rightarrow y \in A$

$$inc(B, A) = \min_{y \in B} (\mu_B(y) \rightsquigarrow \mu_A(y)) = \min_{y \in B} \mu_A(y)$$

- ▶ universal quantification too strict
- ▶ use *fuzzy for most* quantifier

- is fixed point  $y \in B$  presented in  $A$

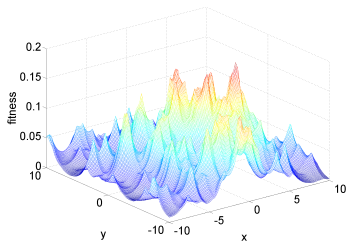
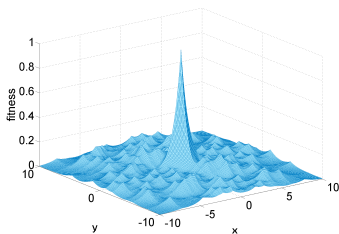
$$\mu_A(y) = \exp \left( -1 \cdot \min_{\substack{x \in A \\ \ell(x) = \ell(y)}} \|y - x\|_1 \right)$$

- degree of inclusion depends on position of point clouds in Euclidean space
  - ▶ find optimal superposition of  $A$  and  $B$  that maximize  $inc(B, A)$
  - ▶ hold  $A$  fix; move  $B$  according to
 
$$t = (\theta_1, \theta_2, \theta_3, \delta_1, \delta_2, \delta_3) \in [0, 2\pi]^3 \times \mathbb{R}^3$$

## Optimization problem

- $\text{TF}(B, t)$  moves  $B$  by  $t = (\theta_1, \theta_2, \theta_3, \delta_1, \delta_2, \delta_3) \in [0, 2\pi]^3 \times \mathbb{R}^3$
- $B^* = \text{TF}(B, t) = \{(y_1^*, \ell(y_1)), \dots, (y_n^*, \ell(y_n))\}$
- position-invariant degree of inclusion of  $B$  in  $A$ :

$$\text{INC}(B, A) = \max_{t \in [0, 2\pi]^3 \times \mathbb{R}^3} \text{inc}(\text{TF}(B, t), A)$$

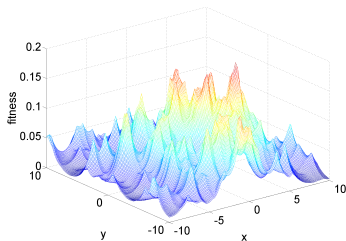
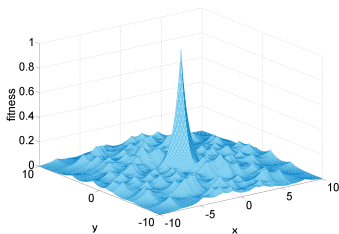


- evolution strategy optimized with SPO

## Optimization problem

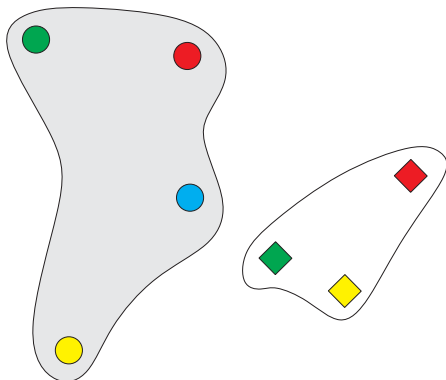
- $\text{TF}(B, t)$  moves  $B$  by  $t = (\theta_1, \theta_2, \theta_3, \delta_1, \delta_2, \delta_3) \in [0, 2\pi]^3 \times \mathbb{R}^3$
- $B^* = \text{TF}(B, t) = \{(y_1^*, \ell(y_1)), \dots, (y_n^*, \ell(y_n))\}$
- position-invariant degree of inclusion of  $B$  in  $A$ :

$$\text{INC}(B, A) = \max_{t \in [0, 2\pi]^3 \times \mathbb{R}^3} \text{inc}(\text{TF}(B, t), A)$$

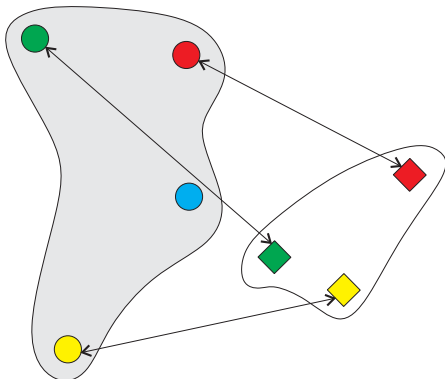


- evolution strategy optimized with SPO

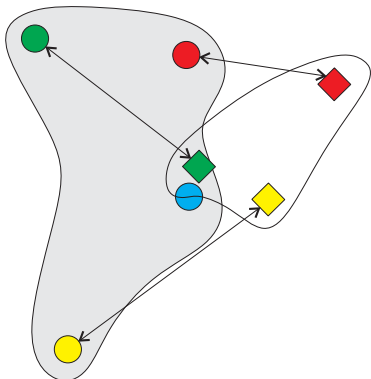
# Example



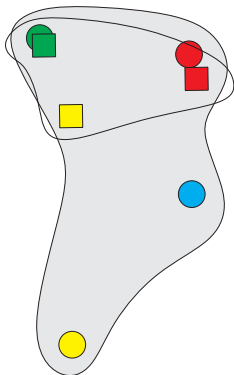
# Example



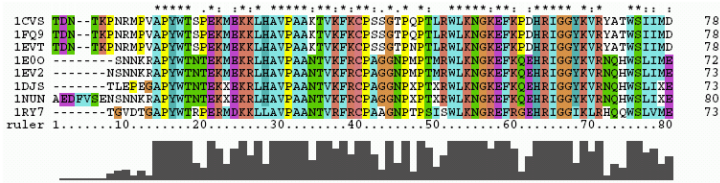
# Example



# Example



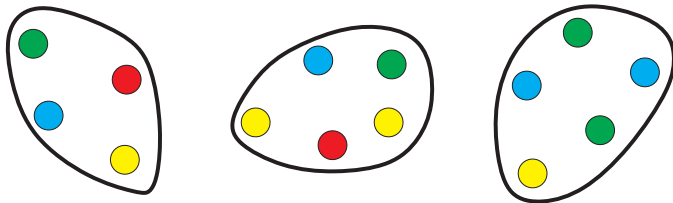
- 1 define an appropriate similarity measure
  - ▶ theory formation in biology founded on similarity-based and analogical reasoning principles
- 2 calculate a geometrical alignment
  - ▶ multiple sequence alignment leads to break through in biology



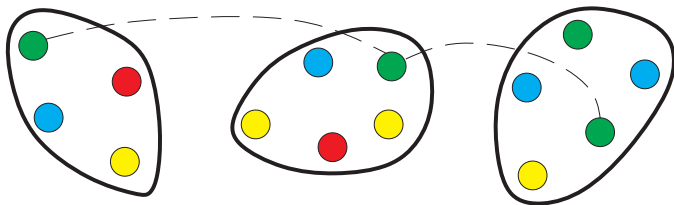
- ▶ functional similarity has no strong correspondence to sequence similarity
- ▶ transfer concept of sequence alignment to point clouds
- ▶ so far structural alignments on graph representation



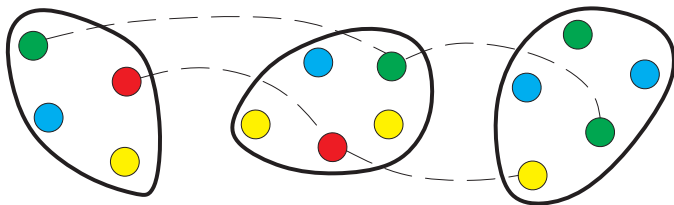
# Geometrical alignment: Example



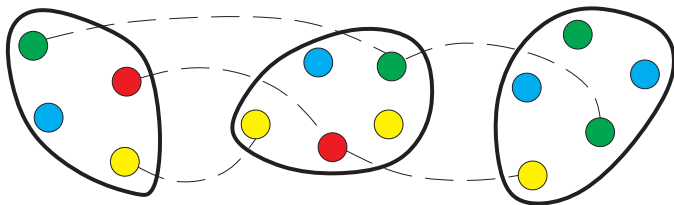
# Geometrical alignment: Example



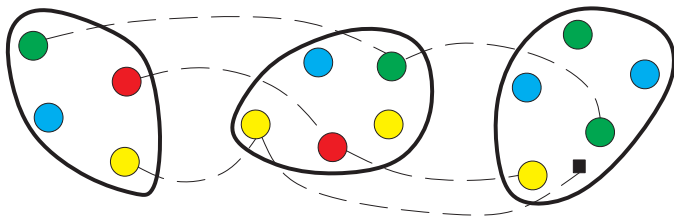
# Geometrical alignment: Example



# Geometrical alignment: Example



# Geometrical alignment: Example

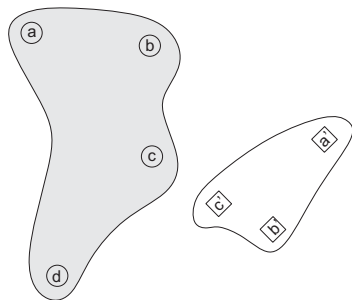


# Optimal alignment

- definition leads to a set of valid alignments
- goal: find optimal alignment
- definition of an optimal alignment?
- combinatorial problem, huge search space
- instead: use the optimal superposition

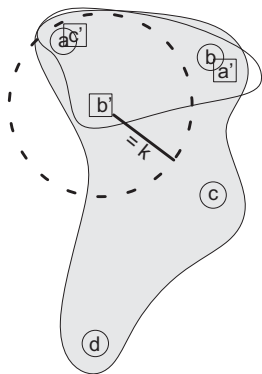
- in a preliminary step calculate the optimal superposition
- determine distances and store them in matrix
- include additional dummy points, distance to existing points given by  $k$
- dummy-dummy distance is zero

	$a'$	$b'$	$c'$	$\perp$	$\perp$	$\perp$	$\perp$
$a$	$d(a, a')$	$d(a, b')$	$d(a, c')$	$k$	$k$	$k$	$k$
$b$	$d(b, a')$	$d(b, b')$	$d(b, c')$	$k$	$k$	$k$	$k$
$c$	$d(c, a')$	$d(c, b')$	$d(c, c')$	$k$	$k$	$k$	$k$
$d$	$d(d, a')$	$d(d, b')$	$d(d, c')$	$k$	$k$	$k$	$k$
$\perp$	$k$	$k$	$k$	$0$	$0$	$0$	$0$
$\perp$	$k$	$k$	$k$	$0$	$0$	$0$	$0$
$\perp$	$k$	$k$	$k$	$0$	$0$ <td $0$	$0$	



- in a preliminary step calculate the optimal superposition
- determine distances and store them in matrix
- include additional dummy points, distance to existing points given by  $k$
- dummy-dummy distance is zero

	$a'$	$b'$	$c'$	$\perp$	$\perp$	$\perp$	$\perp$
$a$	$d(a, a')$	$d(a, b')$	$d(a, c')$	$k$	$k$	$k$	$k$
$b$	$d(b, a')$	$d(b, b')$	$d(b, c')$	$k$	$k$	$k$	$k$
$c$	$d(c, a')$	$d(c, b')$	$d(c, c')$	$k$	$k$	$k$	$k$
$d$	$d(d, a')$	$d(d, b')$	$d(d, c')$	$k$	$k$	$k$	$k$
$\perp$	$k$	$k$	$k$	$0$	$0$	$0$	$0$
$\perp$	$k$	$k$	$k$	$0$	$0$	$0$	$0$
$\perp$	$k$	$k$	$k$	$0$	$0$ <td $0$	$0$	



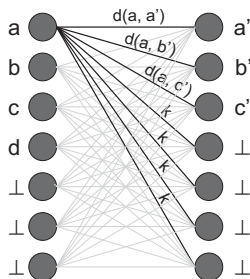
- $G = (V, E)$  where  $V = V_1 \cup V_2, (V_1 \cap V_2) = \emptyset$
- $E = \{(v_1, v_2) \mid v_1 \in V_1, v_2 \in V_2\}$
- each  $e \in E$  has associated costs  $d(e)$
- goal: find  $M \subseteq E$  so that

$$\sum_{e \in M} d(e) \rightarrow \min$$

subject to:

$$\bigcup_{(v_1, v_2) \in M} \{v_1\} = V_1$$

$$\bigcup_{(v_1, v_2) \in M} \{v_2\} = V_2$$



# Multiple alignments

- so far: pairwise alignments
- use the star-alignment approach

# Data sets

- benzamidine
  - ▶ set of 87 chemical compounds
  - ▶ contain 47 – 100 atoms
  - ▶ all compounds share a common core fragment
  - ▶ data set has a clear and unambiguous pattern
- NADH/ATP
  - ▶ 355 protein active pockets
  - ▶ 214 binds to NADH-, 141 to ATP molecules
  - ▶ binary classification problem

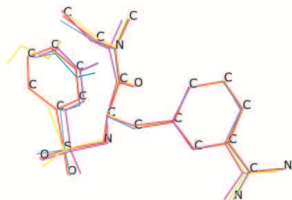
## Results on classification

- concept of similarity vague
- use leave-one-out cross validation and  $k$ -NN classifier
  - ▶ idea: the better the classification rate the better the measure
- for comparison: state-of-the-art graph based approaches

	RW	SP	graph edit distance	LPCS
acc	59.70%	60.60%	76.62%	92.11%
rt	65.5 ± 89.1	9.8 ± 97.8	74.2 ± 85.6	20.0 ± 24.7

# Results on structure retrieval

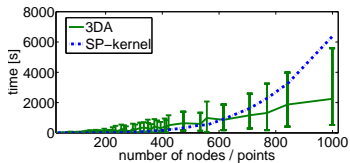
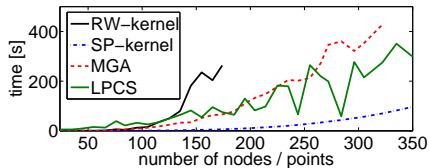
- search for conserved pattern in alignment of  $m$  structures
- for each  $m$  use 100 trials



$m$	graph-based approach	geometrical alignment
2	58%	96%
4	38%	92%
8	14%	80%

# Results on runtime

- RW-kernel:  $\mathcal{O}(n^6)$
- SP-kernel:  $\mathcal{O}(n^4)$



# Conclusions

- working directly on point clouds leads not to loss of information
- efficient and effective similarity measure on point clouds
- six-dimensional optimization problem, independent of point cloud size
- construction of alignment in polynomial time
- better quality of alignment in comparison to graph based approaches