

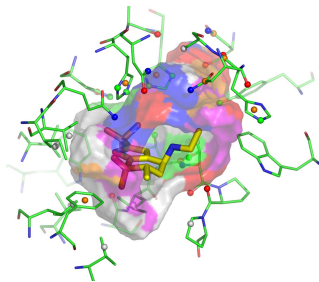
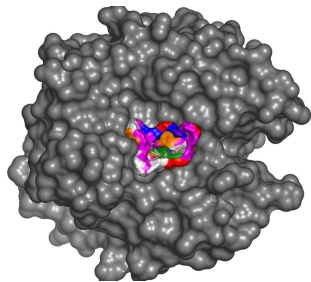
# Fuzzy Modeling of Labeled Point Cloud Superposition for the Comparison of Protein Binding Sites

Thomas Fober    Eyke Hüllermeier

Knowledge Engineering & Bioinformatics Group  
Mathematics and Computer Science Department



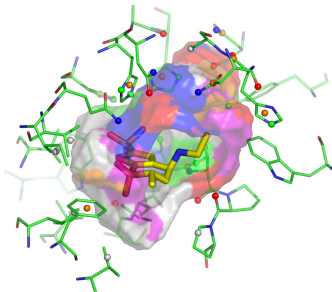
# Protein binding sites



- small cavities on the surface of a protein
- protein consists of amino acids; surface of binding 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 binding sites?

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



# Applications

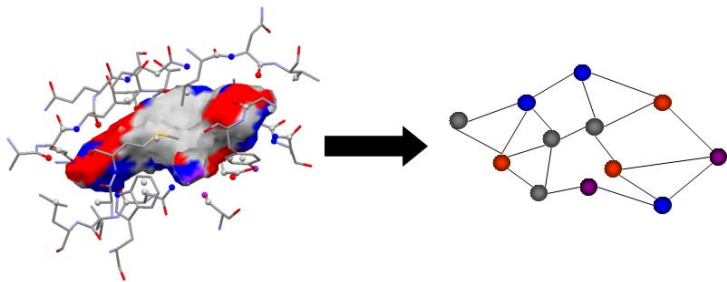
## 1 cross reactivities

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

## 2 prediction of the function

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

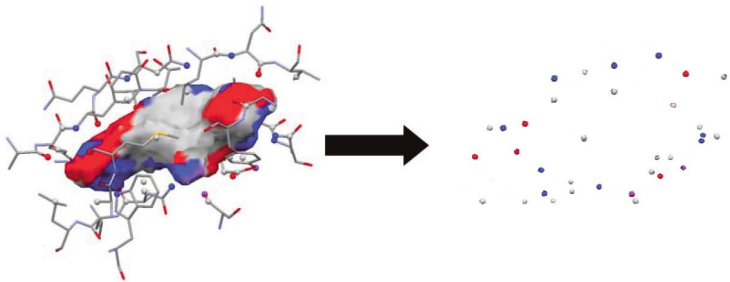
# Graph-representation of protein binding sites



## 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 binding 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

- define an appropriate similarity measure
  - ▶ theory formation in biology founded on similarity-based and analogical reasoning principles
  - ▶ so far: similarity measures on *unlabeled* point clouds
  - ▶ for labeled geometrical data graph based approaches available

# Similarity of labeled 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 cloud
  - ▶ spatially close
  - ▶ same label

# Similarity of labeled 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  $A$

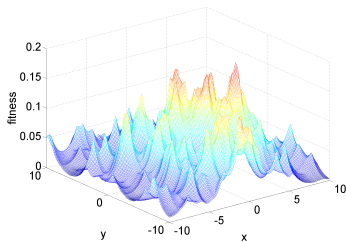
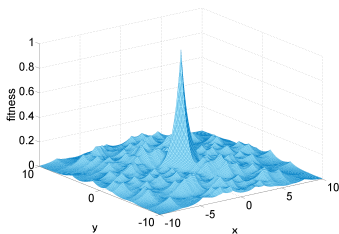
$$\mu_A(y) = \exp \left( -1 \cdot \min_{\substack{x \in A \\ \ell(x) = \ell(y)}} \|y - x\|_1 \right)$$

- 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  $A$ :

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

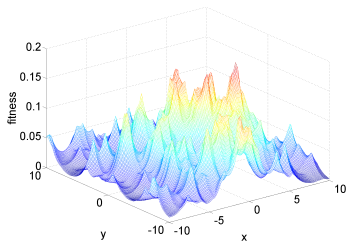
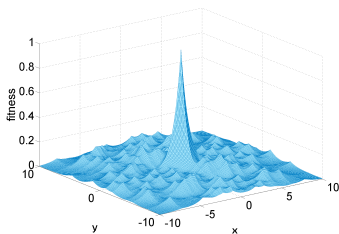


- evolution strategy optimized with SPO

## Optimization problem

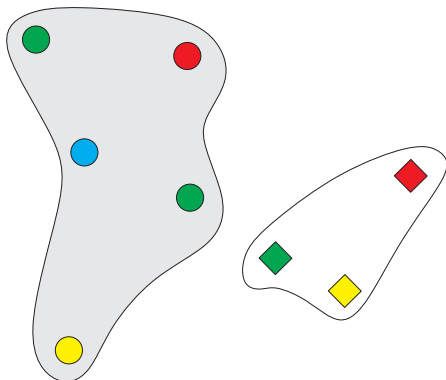
- $\text{TF}(B, t)$  moves  $B$  by  $t = (\theta_1, \theta_2, \theta_3, \delta_1, \delta_2, \delta_3) \in [0, 2\pi]^3 \times \mathbb{R}^3$
- $B^* = \text{TF}(B, t) = \{(y_1^*, \ell(y_1)), \dots, (y_n^*, \ell(y_n))\}$
- position-invariant degree of inclusion of  $B$  in  $A$ :

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

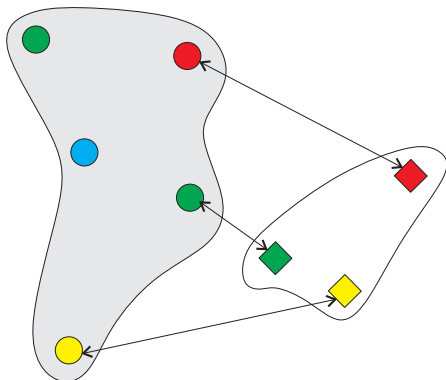


- evolution strategy optimized with SPO

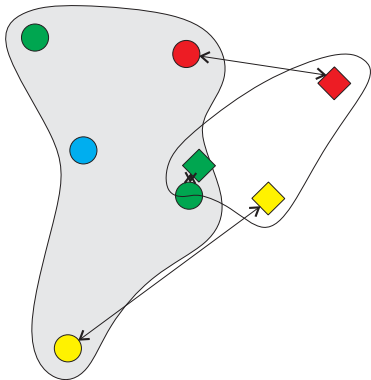
# Example



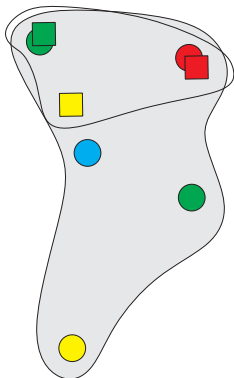
# Example



# Example



# Example



# Data & test procedure

- data set of 355 protein binding sites (NADH/ATP)
- 214 bind NADH-, 141 ATP ligands
- binary classification problem
- use  $k$ -NN classifier
- use leave-one-out cross validation

## 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

# 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