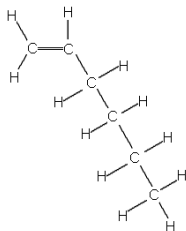


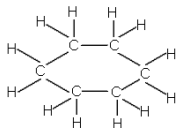
Outline

- 1 Applications
- 2 Multiple graph alignment
- 3 Methods
- 4 Experimental results
- 5 Conclusions

Some applications



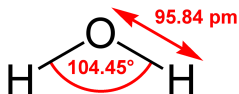
1-Hexen



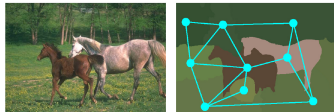
Cyclohexan

- (12, 6, 18) (12, 6, 18)
- feature vectors lead to loss of information
- e.g. counting occurrence of atoms cannot distinguish both molecules

- chemical compounds

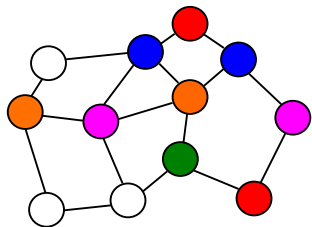
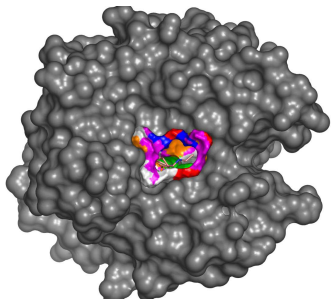


- image classification



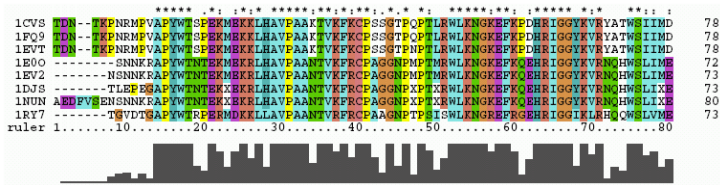
- ...

Protein binding sites



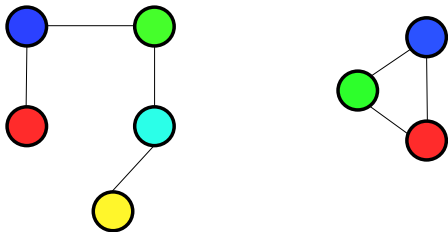
- small cavities on the surface of a protein
- binding sites for small molecules (ligands)
- physico-chemical properties are summarized in point-like pseudocenters

Multiple Sequence Alignment

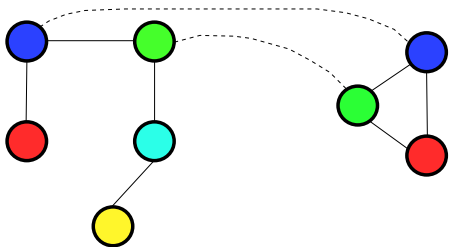


- important concept in bioinformatics
- efficient solution led to a break through in biology

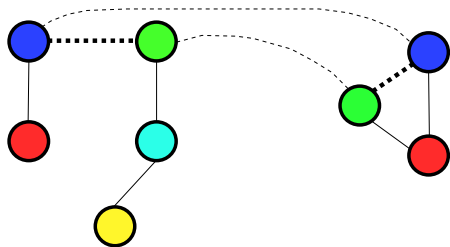
Multiple Graph Alignment



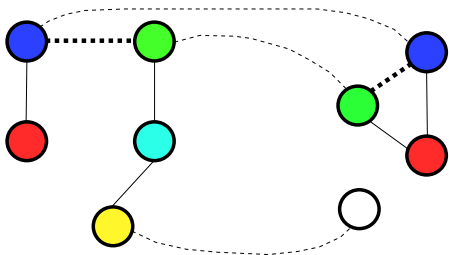
Multiple Graph Alignment



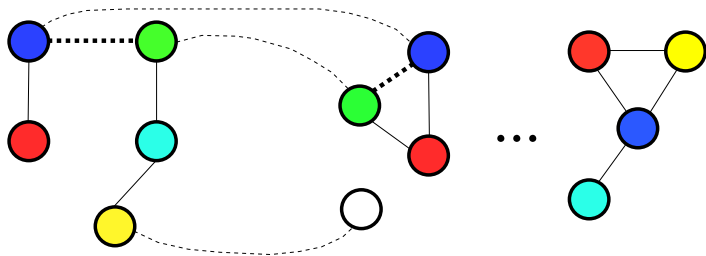
Multiple Graph Alignment



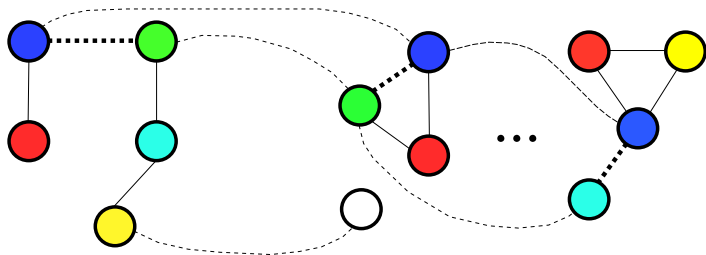
Multiple Graph Alignment



Multiple Graph Alignment



Multiple Graph Alignment



Definition (multiple graph alignment)

$\mathcal{A} \subseteq (V_1 \cup \{\perp\}) \times \dots \times (V_m \cup \{\perp\})$ is a multiple alignment of the graphs $G_1 = (V_1, E_1), \dots, G_m = (V_m, E_m)$ iff

- 1 for all $i = 1 \dots m$ and for each $v \in V_i$ there exists exactly $a = (a_1 \dots a_m) \in \mathcal{A}$, such that $v = a_i$.
- 2 for each $a = (a_1 \dots a_m) \in \mathcal{A}$ there exists at least one $1 \leq i \leq m$, such that $a_i \neq \perp$

Definition (sum-of-pairs measure)

$$f(\mathcal{A}) = \sum_{a_i \in \mathcal{A}} \text{score}_V(a_i) + \sum_{\substack{a_i \in \mathcal{A}, a_j \in \mathcal{A} \\ i < j}} \text{score}_E(a_i, a_j)$$

with

$$\text{score}_V(a_i) = \text{score}_V \begin{pmatrix} a_1^i \\ \vdots \\ a_m^i \end{pmatrix} = \sum_{1 \leq j < k \leq m} \begin{cases} c_{\text{match}} & l(a_j^i) = l(a_k^i) \\ c_{\text{mismatch}} & l(a_j^i) \neq l(a_k^i) \\ c_{\text{dummy}} & a_j^i = \perp, a_k^i \neq \perp \\ c_{\text{dummy}} & a_j^i \neq \perp, a_k^i = \perp \end{cases}$$

calculation of edge-score analog

- edit-operations are penalized by $c_i < 0$
- $c_i \geq 0$ rewards a mapping that does not require an edit operation

Definition (optimal alignment)

$$\mathcal{A}^* \in \mathit{arg\,max}_{\mathcal{A}} f(\mathcal{A})$$

Greedy approach, Weskamp et al. (2007)

decomposition method (star-alignment)

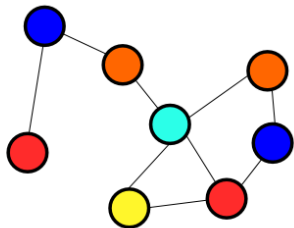
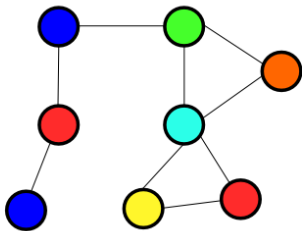
multiple alignment \rightarrow set of pairwise alignments

- choose center-graph G_c
- calculate pairwise alignments: $G_c || G_1, \dots, G_c || G_m$
- multiple alignment: $G_1 || G_2 || \dots || G_m$

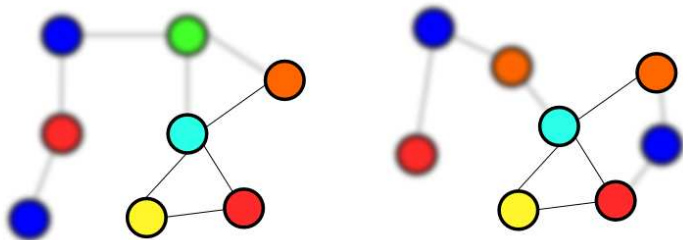
pairwise alignment

- identification of common subgraph
- greedy extension of common subgraph to an alignment

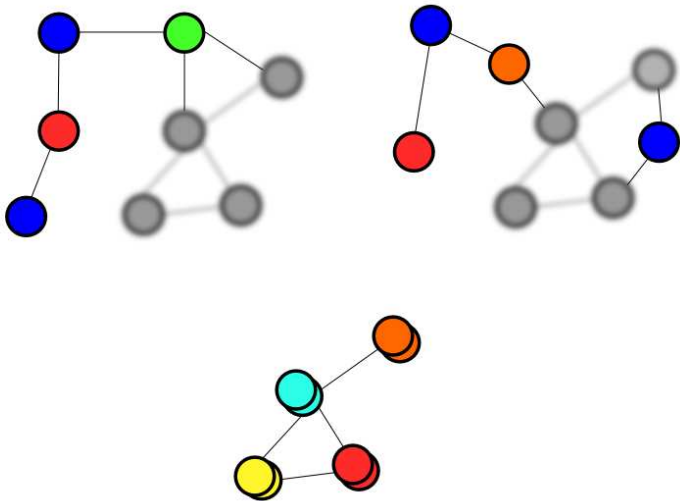
Example



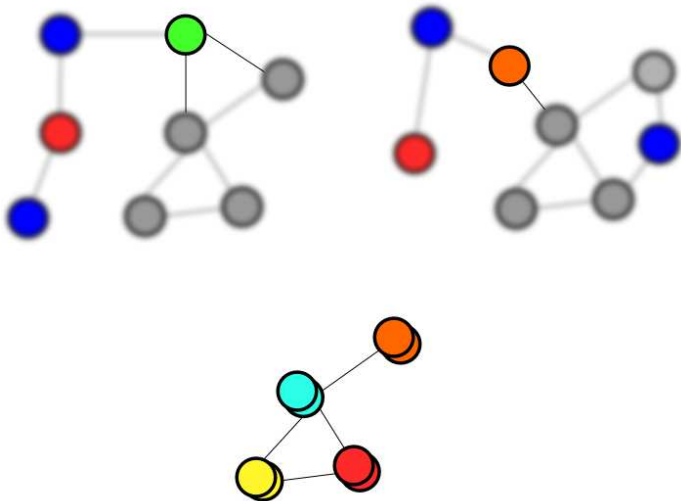
Example



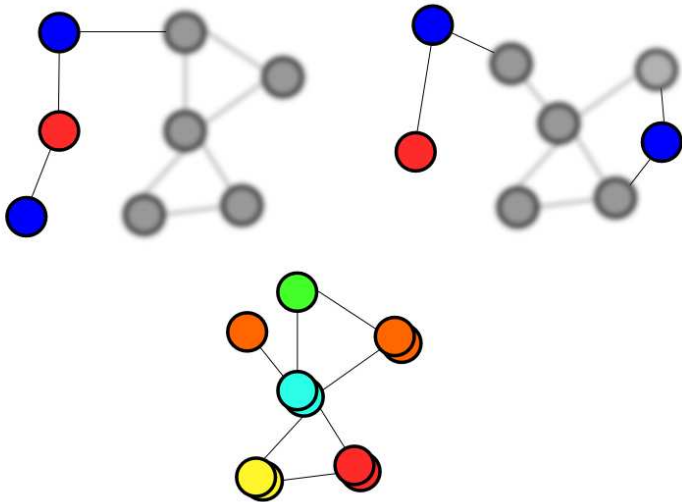
Example



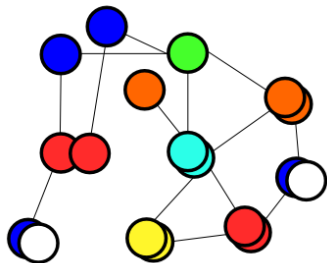
Example



Example



Example

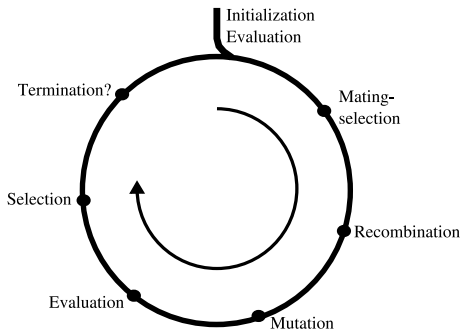


Drawbacks

- decision cannot be revised later
- this can lead to missing the global optimum
- Weskamp et al. approach consists of two greedy heuristics
- finding seed solution requires the calculation of a product graph

→ use a global optimizer that is able to calculate an MGA directly

EA: Loop and parameters



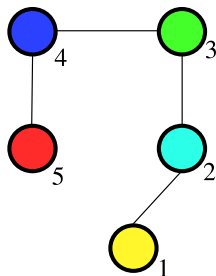
parameter	μ	ν	rec	adjustProb
meaning	population-size	selective-pressure	recombination	prob. of insertion /deletion

EA: Representation

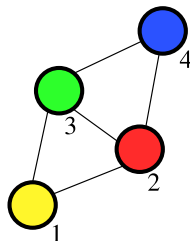
$m \times L$ matrix, $L = \max_{i=1, \dots, m} \{|V_i|\} + 1$

A: 1 3 5 2 4 .

B: 3 4 . 2 1 .



A



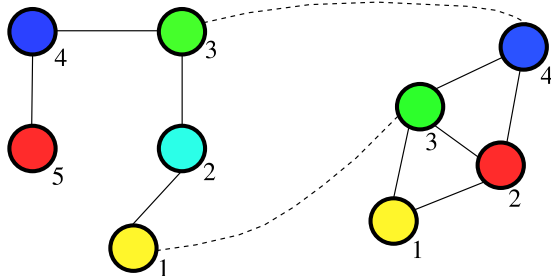
B

EA: Representation

$m \times L$ matrix, $L = \max_{i=1, \dots, m} \{|V_i|\} + 1$

A: 1 3 5 2 4 .

B: 3 4 . 2 1 .



A

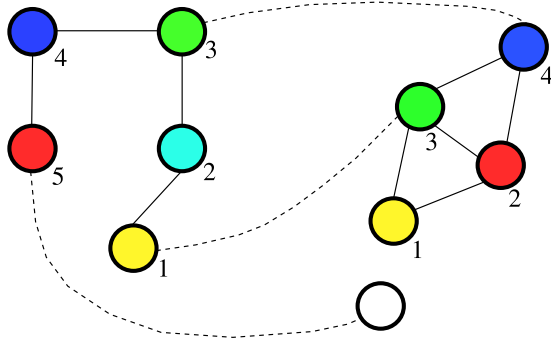
B

EA: Representation

$m \times L$ matrix, $L = \max_{i=1, \dots, m} \{|V_i|\} + 1$

A: 1 3 5 2 4 .

B: 3 4 . 2 1 .



A

B

EA: Initialization and fitness evaluation

- initialization

- ▶ given m graphs; determine $L = \max_{i=1, \dots, m} \{|V_i|\} + 1$
- ▶ in each row j :
 - ★ random permutation over $\{1, \dots, L\}$
 - ★ substitute elements $x > |V_j|$ by dummy nodes

- evaluation

- ▶ sum-of-pairs measure
- ▶ ignore dummy columns

EA: Recombination, $rec : I^2 \rightarrow I$

- recombination of 2 parents
- mechanism required that establishes an ordering
 - choose 1 pivot-graph G_c
 - columns are arranged according to G_c

2	.	1	3	4	3	.	2	4	1	2	.	1	3	4
5	3	1	2	4	1	3	2	4	5	5	3	1	2	4
3	.	2	1	4	2	.	1	4	3	3	.	2	1	4
4	.	2	3	1	1	.	4	3	2	2	.	1	4	3
2	1	4	.	3	2	4	1	.	3	3	4	2	1	.

EA: Mutation, $mut : I \rightarrow I$

● Reachability

- ▶ each point in the search space should be reachable
- ▶ choose one row and two columns at random
- ▶ swap both cells

1	.	2	3	4
1	2	5	.	3
1	.	2	3	4
1	2	4	.	3

↓

● adapt length L of individual with probability p_l

- ▶ $k > 1$ dummy columns \rightarrow remove $k - 1$ of these columns
- ▶ no dummy columns \rightarrow insert a new dummy column

1	.	2	3	4
1	2	3	.	5
1	.	2	3	4
1	2	3	.	4

EA: Other operators

- mating selection, $I^\mu \rightarrow I^2$: uniformly random
- selection, $I^{\mu+\lambda} \rightarrow I^\mu$: deterministic plus-selection
- termination: after T generations without improvement

Variant: EA*

observation:

- calculation of graphs alignments...
 - ... leads to very good results
 - ... in a short runtime
 - ▶ for the pairwise case

idea:

- use star-alignment
- solve pairwise alignments with EA

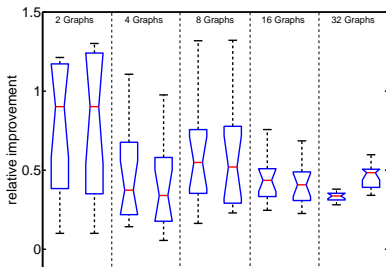
Data sets

- thermolysin
 - ▶ 74 cavities of bacterial proteases
 - ▶ proteases are evolutionary related
 - ▶ highly conserved structures ought to be present
 - ▶ cavities contain 30 – 90 pseudocenters
- benzamidine
 - ▶ 87 chemical compounds
 - ▶ graphs descriptors contain 47 – 100 nodes
 - ▶ all compounds share a common core fragment
 - ▶ data set has a clear and unambiguous pattern
- NADH / ATP
 - ▶ data set consists of 355 protein binding sites
 - ▶ 241 pockets bind to ATP ligands
 - ▶ 141 pockets bind to NADH ligands
 - ▶ 2 class classification problem

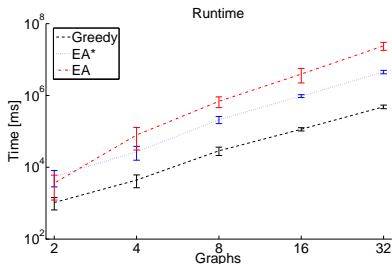
Relative improvement

$$ri = \frac{\text{score}(EA) - \text{score}(Greedy)}{\min\{|\text{score}(EA)|, |\text{score}(Greedy)|\}}$$

thermolysin

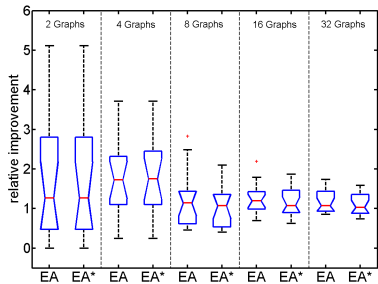


(a) rel. improvement

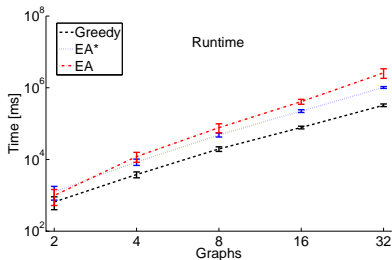


(b) runtime

benzamidine



(c) rel. improvement



(d) runtime

Classification

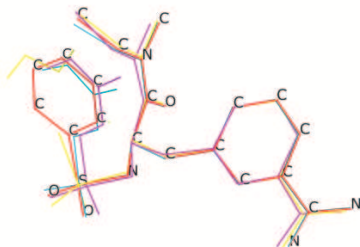
- use MGA as similarity measure (pairwise case)
- better alignment \rightarrow better similarity score \rightarrow better classification rate
- two-class ATP/NADH data set
- predict to which ligand a pocket binds
- k -NN classifier
- leave-one-out cross validation

k	EA	Greedy
1	78.87%	76.62%
3	76.62%	71.83%
9	76.62%	71.27%

Structure retrieval

- 87 compounds of benzamidine derivates
- all derivates share the benzamidine core fragment
- search for conserved structures

m	EA	Greedy
2	100%	58%
4	96%	38%
8	94%	14%
16	98%	4%
32	100%	2%



Conclusions

- 1 MGA: novel method to analyze biomolecules on structural level
- 2 discover approximately conserved patterns in a set of graphs
- 3 alignments constructed with EA have better scores and can retrieve more conserved structures compared to existing greedy solution
- 4 using the EA alignments as similarity measure leads to a better classification rate