

Berechnung Multipler Graph Alignments mittels Evolutionärer Algorithmen

Thomas Fober Eyke Hüllermeier Marco Mernberger

Knowledge Engineering & Bioinformatics Group
Mathematics and Computer Science Department



Bommerholz, 2007

Inhaltsverzeichnis

- 1 Modellieren mit Graphen
- 2 Multiples Graph Alignment
- 3 Algorithmen
- 4 Experimente
- 5 Zusammenfassung und Ausblick

Proteinanalytik

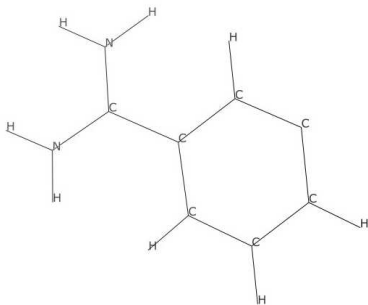
Vergleich von (potentiell sehr vielen) Proteinstrukturen

- Welche strukturelle Gemeinsamkeit besteht?
- Existieren strukturelle Unterschiede?

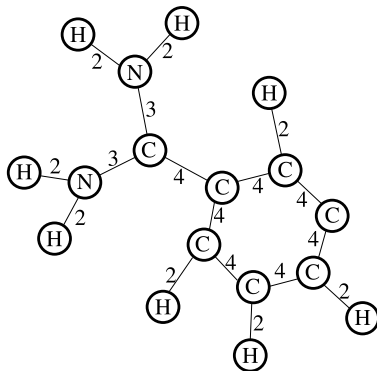
Verfahren wie das Sequenzalignment sind oft nicht ausreichend

- Funktion ist nicht aus Sequenz vorhersagbar
- Räumliche Struktur ist maßgeblich für die Funktion des Proteins

Graphmodellierung eines Moleküls

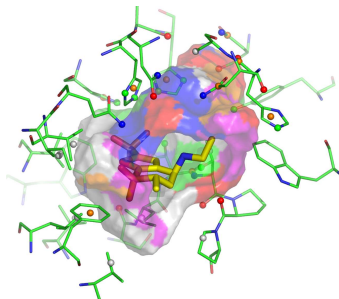
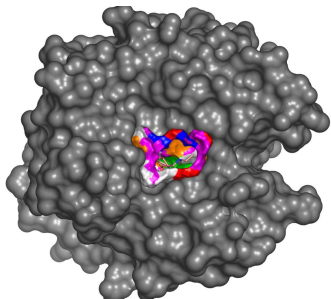


Benzamidin Struktur



Benzamidin Struktur als Graph

Proteinbindetaschen



- Knoten repräsentieren Pseudozentren
- Pseudozentren \rightsquigarrow physiko-chemische Eigenschaften
- Protein - Ligand Wechselwirkungen über Pseudozentren

Übersicht

- 1 Modellieren mit Graphen
- 2 Multiples Graph Alignment**
- 3 Algorithmen
- 4 Experimente
- 5 Zusammenfassung und Ausblick

Als Einstieg: Sequenzalignment

- Wie ähnlich sind Sequenzen s , t und u ?
- Eindeutige Zuordnung von Symbolen zueinander, so dass
 - ▶ Reihenfolge erhalten bleibt
 - ▶ jedes Symbol einer Sequenz genau einem Symbol einer anderen Sequenz oder einer Lücke zugeordnet wird
 - ▶ eine Fehlpaarung einer Mutation entspricht (Editieroperation)
 - ▶ eine Lücke einer Einfüge/Lösch-Operation entspricht (Editieroperation)
- Optimales Alignment: Alignment, das mit möglichst wenigen Editieroperationen auskommt

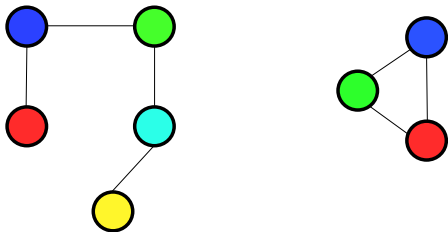
s:	A	C	C	T	G	-	A	T
t:	A	T	G	T	G	C	A	T
u:	G	C	T	A	A	G	C	T

Vom Sequenz- zum Graphalignment

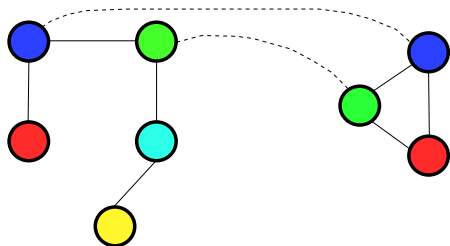
Unterschiede?

- Symbole \rightarrow Knoten und Kanten
- Keine lineare Ordnung
- Geometrische Information
- Messfehler, konformative Flexibilität (\rightsquigarrow approximative Muster)

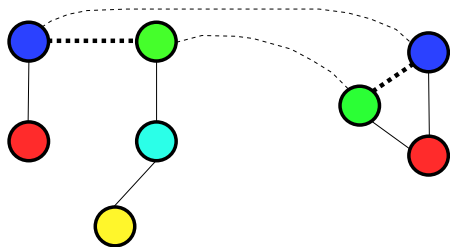
Beispiel



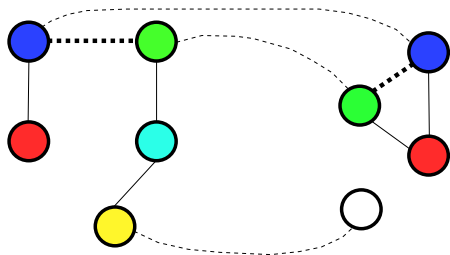
Beispiel



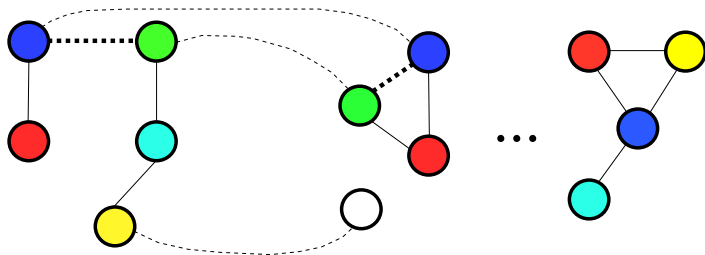
Beispiel



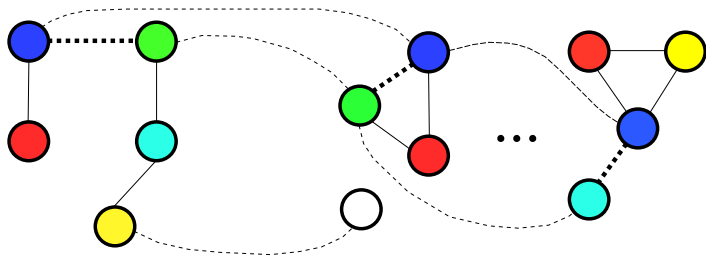
Beispiel



Beispiel



Beispiel



Formale Definitionen

Definition (Multiples Graph Alignment)

$\mathcal{A} \subseteq (V_1 \cup \{\perp\}) \times \cdots \times (V_m \cup \{\perp\})$ ist Alignment der Graphen $G_1 = (V_1, E_1), \dots, G_m = (V_m, E_m)$ gdw.

- 1 für $i = 1 \dots m$ und $v \in V_i$ existiert genau ein $a = (a_1 \dots a_m) \in \mathcal{A}$, so dass $v = a_i$.
- 2 für $a = (a_1 \dots a_m) \in \mathcal{A}$ existiert ein $1 \leq i \leq m$, so dass $a_i \neq \perp$

Definition (Sum-Of-Pairs Maß)

- Jede Editieroperation wird mit Konstanten $c_i < 0$ bewertet
- Ist keine Editieroperation nötig, bewerte Zuordnung mit $c_i > 0$

$$f(\mathcal{A}) = \text{score}_V(\mathcal{A}) + \text{score}_E(\mathcal{A})$$

mit

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

$$\text{score}_E \left(\left(\begin{pmatrix} a_1^j \\ \vdots \\ a_m^j \end{pmatrix} \right), \left(\begin{pmatrix} a_1^l \\ \vdots \\ a_m^l \end{pmatrix} \right) \right) = \sum_{1 \leq k < l \leq m} \begin{cases} c_{emm} & (a_k^j, a_k^l) \in E_k, (a_j^j, a_l^l) \notin E_l \\ c_{emm} & (a_k^j, a_k^l) \notin E_k, (a_j^j, a_l^l) \in E_l \\ c_{em} & d_{kl}^{jj} \leq t_{max} \\ c_{emm} & d_{kl}^{jj} > t_{max} \end{cases}$$

Definition (Optimales Alignment)

$$\mathcal{A}^* \in \arg \max_{\mathcal{A}} f(\mathcal{A})$$

(Offensichtlich ein kombinatorisches Maximierungsproblem)

Übersicht

- 1 Modellieren mit Graphen
- 2 Multiples Graph Alignment
- 3 Algorithmen**
- 4 Experimente
- 5 Zusammenfassung und Ausblick

Bisherige Methode

Dekompositionsmethode (Star-Alignment)

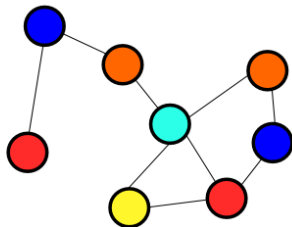
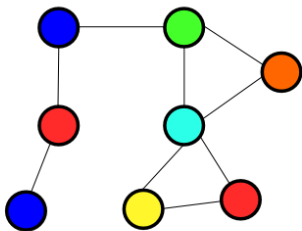
Multiples Alignment $\rightarrow \frac{1}{2} \cdot (m^2 - m)$ paarweise Alignments

- Wähle einen Graphen G_z als Zentrum
- Bestimme paarweise Alignments:
 $G_z || G_1, \dots, G_z || G_m$
- Multiples Alignment: $G_1 || G_2 || \dots || G_m$

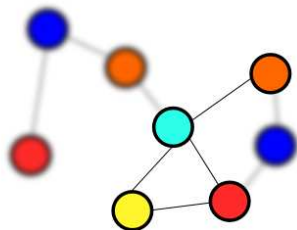
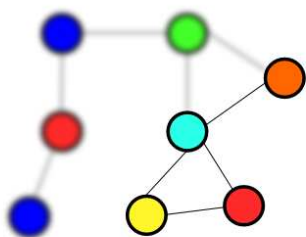
Paarweises Alignment

- Identifiziere gemeinsamen Teilgraphen
- Greedy Erweiterung der Teilgraphen zum Alignment

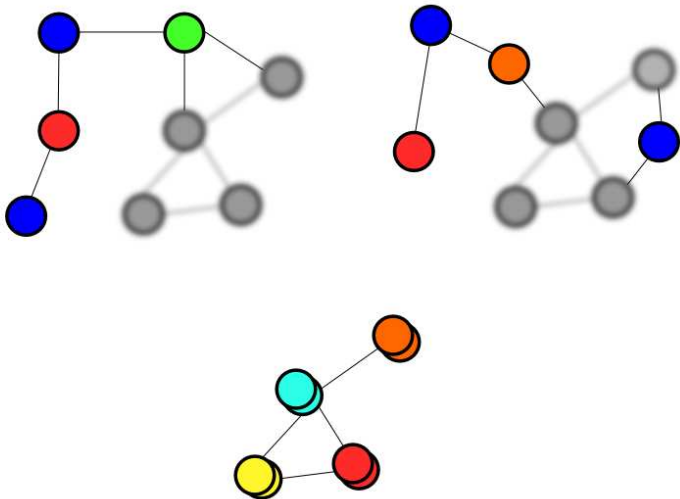
Beispiel



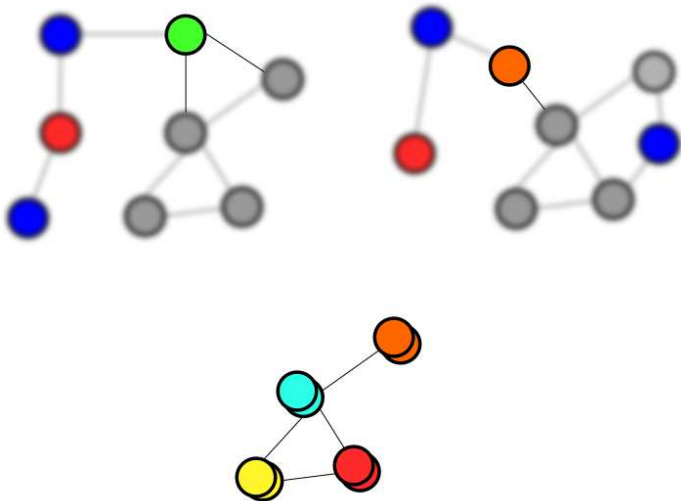
Beispiel



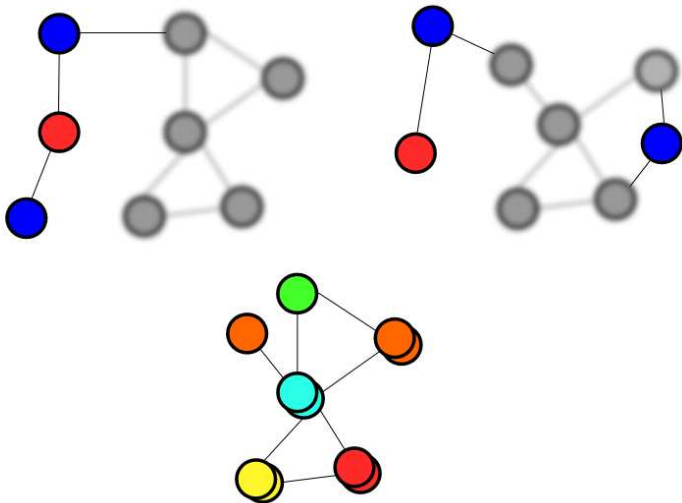
Beispiel



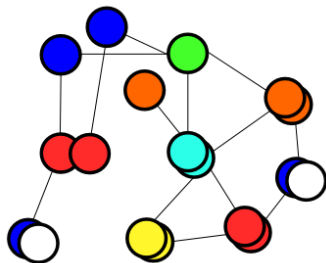
Beispiel



Beispiel



Beispiel

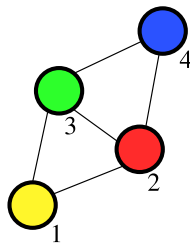
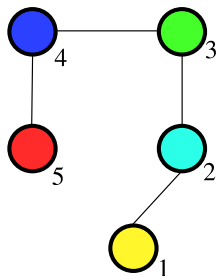


EA: Repräsentation

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

A: 1 3 5 2 4 .

B: 3 4 . 2 1 .

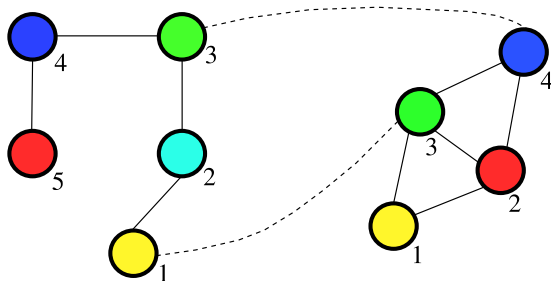


EA: Repräsentation

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

A: 1 3 5 2 4 .

B: 3 4 . 2 1 .

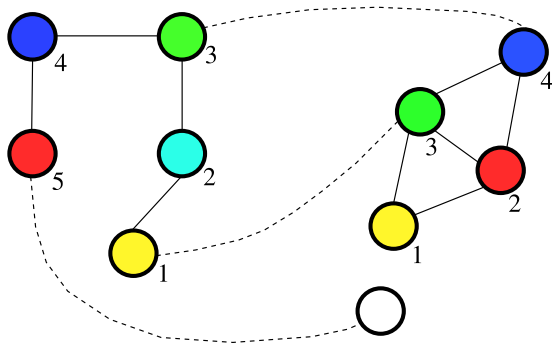


EA: Repräsentation

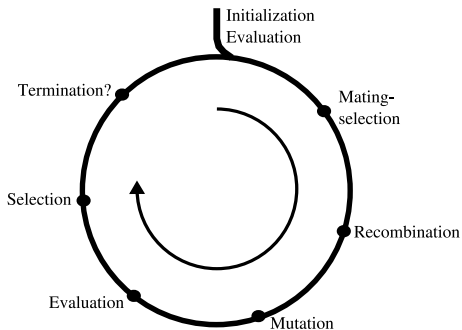
$m \times L$ Matrix, $L = \max_{i=1, \dots, m} \{|V_i|\} + 1$

A: 1 3 5 2 4 .

B: 3 4 . 2 1 .



EA: Schleife und Parameter



Parameter	μ	ν	rec	selfadapt	σ_0
Bedeutung	Populationsgröße	Selektionsdruck	Rekombination	Selbstadaptation	Initialschrittweite

EA: Initialisierung und Fitnessfunktion

● Initialisierung

- ▶ Gegeben m Graphen; bestimme $L = \max_{i=1, \dots, m} \{|V_i|\} + 1$
- ▶ Erzeuge $m \times L$ Matrix
- ▶ In Zeile j :
 - ★ Zufällige Permutation über $\{1, \dots, L\}$
 - ★ Ersetze Elemente $x > |V_j|$ durch Dummy-Knoten

● Bewertung

- ▶ Verwendung des Sum-of-Pairs Maß
- ▶ Dummy-Spalten ignorieren

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

- Ausschließlich zwei Eltern ($\rho = 2$)
- Keine feste Ordnung der Spalten
- Ordnung erzeugender Mechanismus erforderlich:
 - ▶ Wähle Pivot-Graphen G_z
 - ▶ Richte Spalten entsprechend G_z aus

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

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

- Ausschließlich zwei Eltern ($\rho = 2$)
- Keine feste Ordnung der Spalten
- Ordnung erzeugender Mechanismus erforderlich:
 - ▶ Wähle Pivot-Graphen G_Z
 - ▶ Richte Spalten entsprechend G_Z aus

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

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

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

- Phänotypische Veränderung des Individuums

- ▶ Zufällige Wahl einer Zeile und zweier Spalten
- ▶ Austausch beider Zellen
- ▶ $\lceil \sigma \rceil$ -fache Wiederholung

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

↓

- Anpassung der Länge L des Individuums mit geringer Wahrscheinlichkeit

- ▶ Überflüssige Dummy \rightarrow Dummy-Spalte
- ▶ Keine Dummy-Spalte \rightarrow Zu wenige Dummies vorhanden

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

EA: Übrige Operatoren

- Paarungsselektion, $I^\mu \rightarrow I^2$: Gleichverteilt zufällig
- Umweltselektion, $I^{\mu+\lambda} \rightarrow I^\mu$: Deterministische Plus-Selektion
- Abbruchbedingung: Nach T Generationen ohne Verbesserung

Variante: EA*

Beobachtung:

- Für paarweises Alignment liefert der EA...
 - ... sehr gute Ergebnisse
 - ... in kurzer Zeit

Idee:

- Verwende Star-Alignment
- Löse paarweise Alignments mit EA

Übersicht

- 1 Modellieren mit Graphen
- 2 Multiples Graph Alignment
- 3 Algorithmen
- 4 Experimente**
- 5 Zusammenfassung und Ausblick

Verwendete Datensätze

Benzamidin

- 88 Benzamide Derivate
- Jeweils bis zu 100 Atome
- Gemeinsame Grundstruktur

Thermolysin

- 74 Proteinbindetaschen der Proteinfamilie Thermolysin
- Jeweils 30 - 90 Pseudozentren
- Keine definierte Grundstruktur

Optimierung der exogenen Parameter

μ	ν	σ_0	rec	sa
?	?	?	?	?

Welche Parametrisierung soll gewählt werden?

- Einsatz der *Sequential Parameter Optimization Toolbox*
- Suche EA Parametrisierung, so dass:
 - 1 optimale Lösung gefunden wird
 - 2 dafür benötigte Zeit gering ist

~> Betrachtung der Zeit zum Erreichen des Optimums

Optimierung der exogenen Parameter

μ	ν	σ_0	rec	sa
4	20	—	'on'	'off'

Welche Parametrisierung soll gewählt werden?

- Einsatz der *Sequential Parameter Optimization Toolbox*
- Suche EA Parametrisierung, so dass:
 - 1 optimale Lösung gefunden wird
 - 2 dafür benötigte Zeit gering ist

~> Betrachtung der Zeit zum Erreichen des Optimums

Vergleich zwischen Greedy, EA und EA*

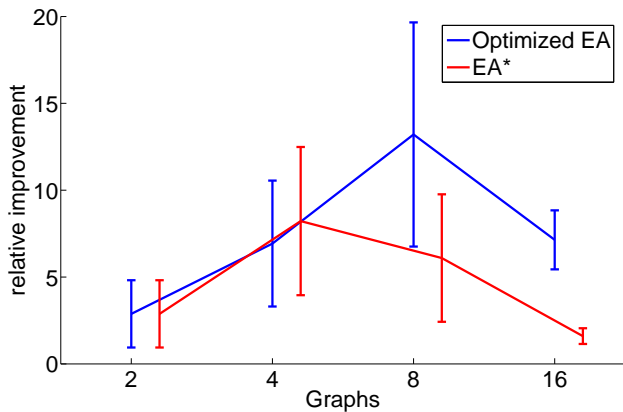
Zwei Kriterien:

- Laufzeit
- Verbesserung der Resultate
 - ▶ Geeignetes Maß: Relative improvement (ri)
 - ▶
$$ri = \frac{f(EA) - f(G)}{\min\{|f(EA)|, |f(G)|\}}$$

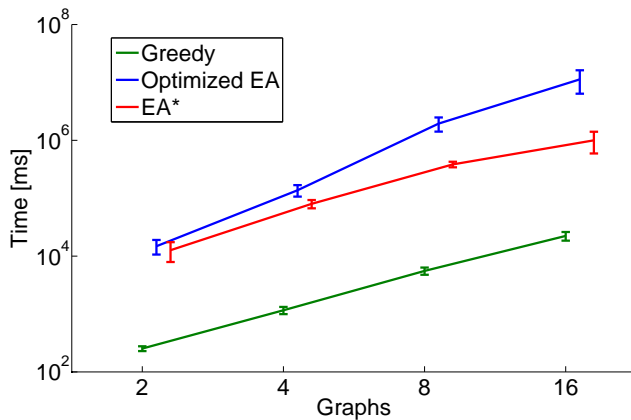
Verfahren:

- Vergleich für $m = 2^k, k = 1, \dots, 4$
- Für jedes m : Zufällige Teilmengen aus den Datensätzen
- Auf jeder der Teilmengen: 4 Versuche für die EA-Varianten
- Jeweils für Benzamidin und Thermolysin

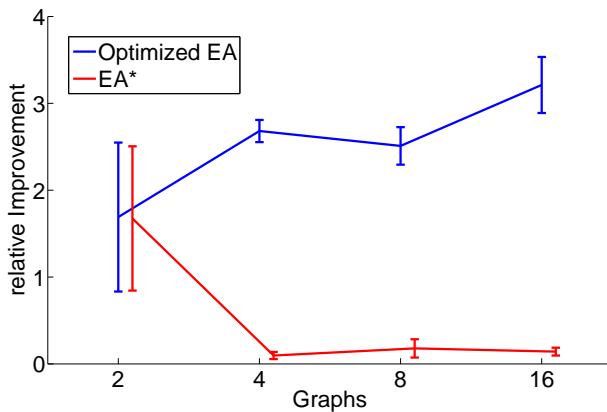
Benzamidin Datensatz: Relative Improvement



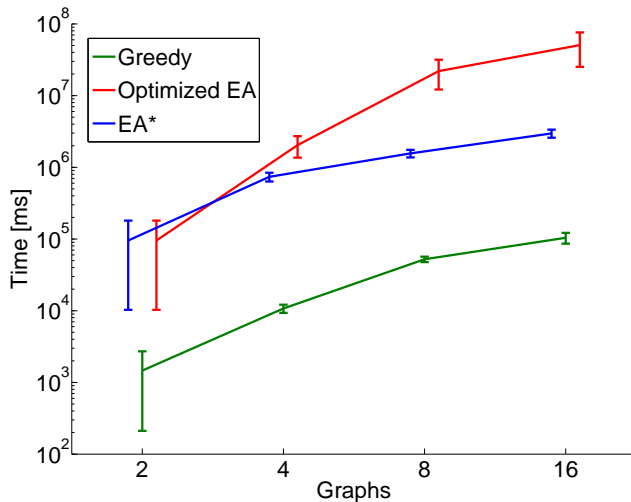
Benzamidin Datensatz: Laufzeit



Thermolysin Datensatz: Relative Improvement



Thermolysin Datensatz: Laufzeit



Übersicht

- 1 Modellieren mit Graphen
- 2 Multiples Graph Alignment
- 3 Algorithmen
- 4 Experimente
- 5 Zusammenfassung und Ausblick**

Zusammenfassung und Ausblick

Zusammenfassung

- Graphalignment: Graphbasiertes Pendant zum Sequenzalignment
- Neue Algorithmen führen zu signifikanter Verbesserung der Alignments auf Kosten der Laufzeit

Ausblick

- Verbesserte Aggregationsmethode
- Verbesserte (genetische) Operatoren
- Initialisierung basierend auf Vorwissen
- Allgemein: Verkleinerung des Suchraums