Erkennung von Nachbarschaftsbeziehungen räumlicher Strukturen am Beispiel von DNA - Daten

(Diplomthema)

Name: Tim Schliebe

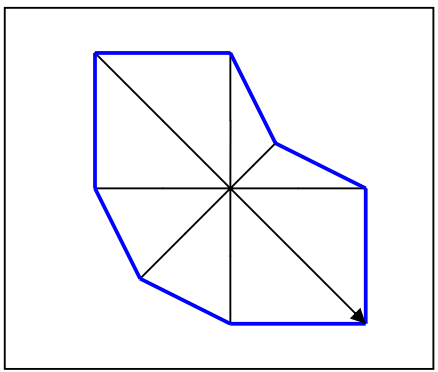
Betreuer: Dr. Sosna

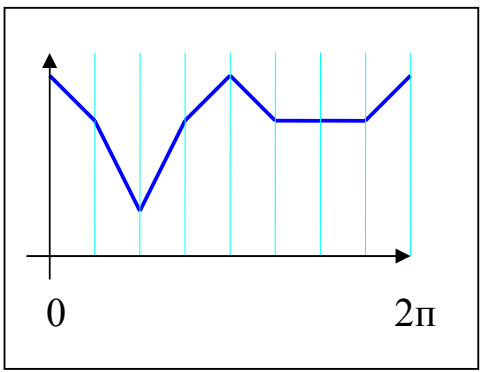
Leipzig, November 2005

Motivation

Idee von Professor Kriegel (UNI-München):

DNA's vergleichen, indem man die geometrische Gestalt vergleicht

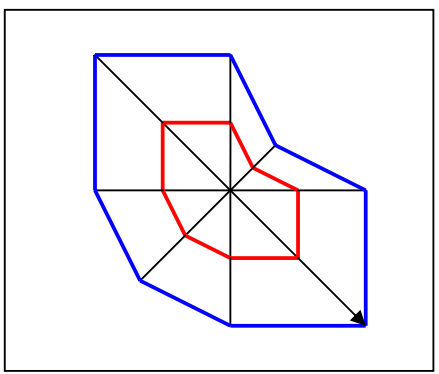


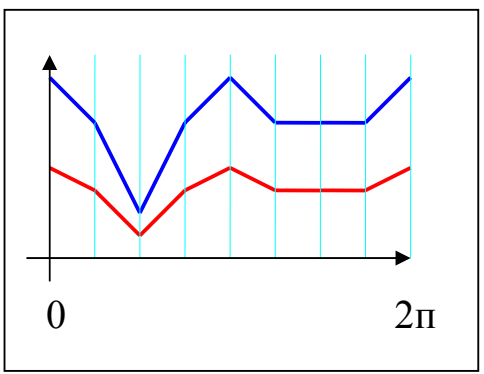


Motivation

Idee von Professor Kriegel (UNI-München):

DNA's vergleichen, indem man die geometrische Gestalt vergleicht



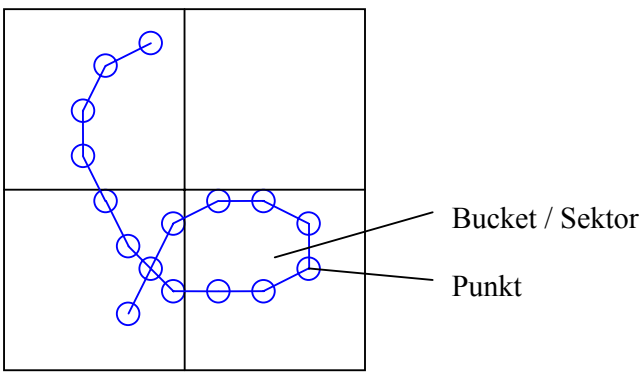


Vergleich der Ähnlichkeit der roten und blauen Figur

Motivation

- herausfinden, ob sich DNA - Abschnitte, die im Index weit auseinander liegen, wieder geometrisch zusammenkommen

- effizientere Suche nach benachbarten räumlichen DNA-Abschnitten

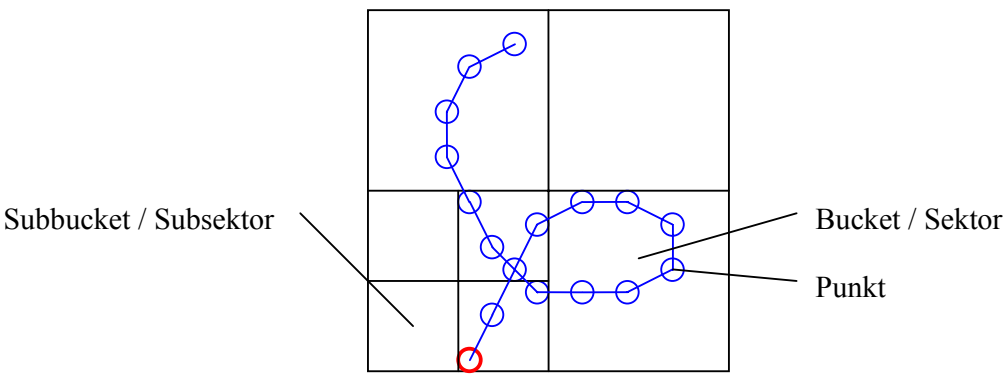


Zustand vor Einfügen eines neuen Punktes bei einer Bucketgröße von 6

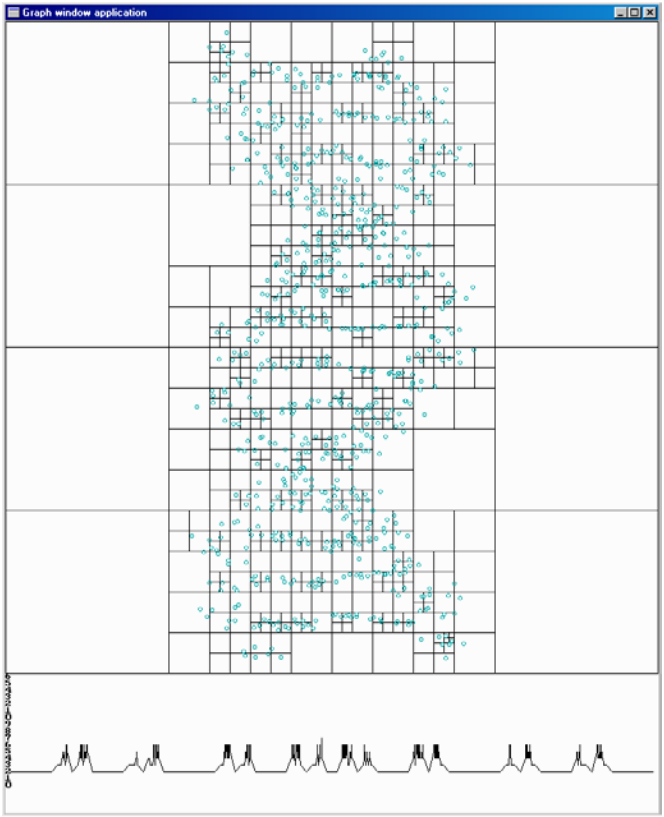
Motivation

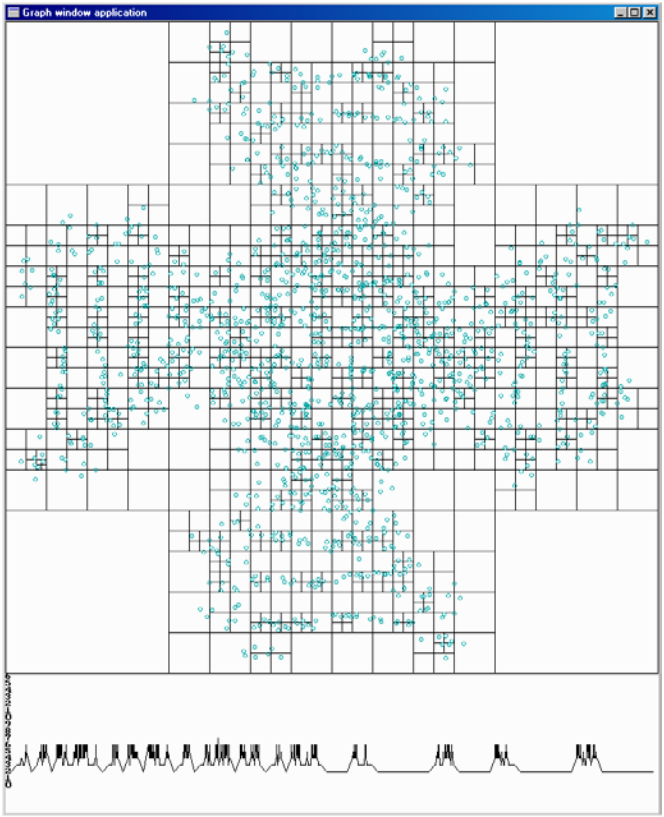
- herausfinden, ob sich DNA - Abschnitte, die im Index weit auseinander liegen, wieder geometrisch zusammenkommen

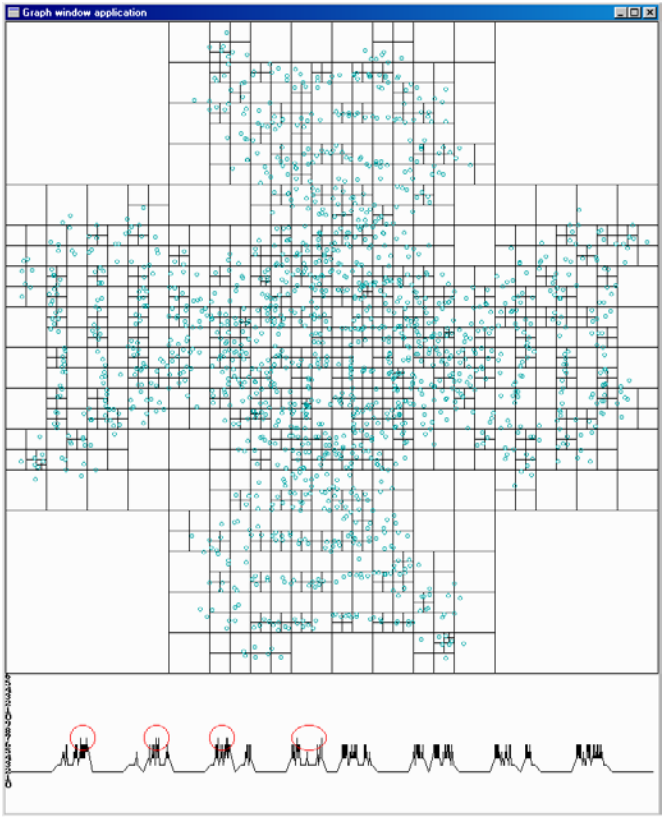
- effizientere Suche nach benachbarten räumlichen DNA-Abschnitten



Zustand nach Einfügen eines neuen Punktes bei einer Bucketgröße von 6


Graph window application

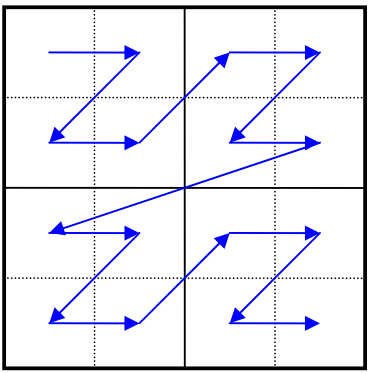

Graph window application


Graph window application

Anfragemöglichkeiten

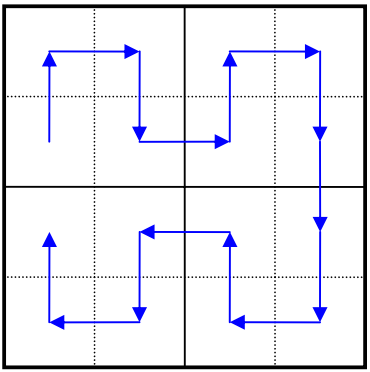
- welche DNA - Abschnitte befinden sich im gleichen Sektor
- welche Basen sind in der Nähe eines Sektors

Raumfüllende Kurven / Fraktale



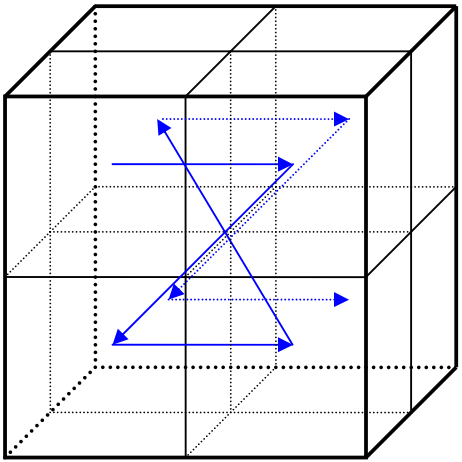
Z – Kurve der Stufe 2 in der Ebene

Raumfüllende Kurven / Fraktale



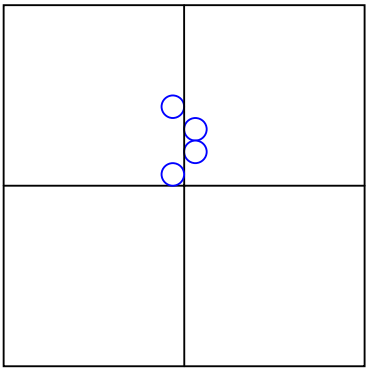
Hilbert – Kurve der Stufe 2 in der Ebene

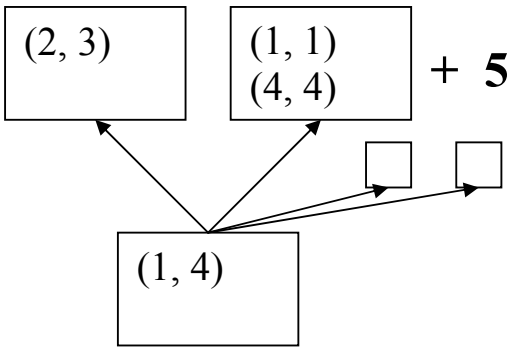
Raumfüllende Kurven / Fraktale

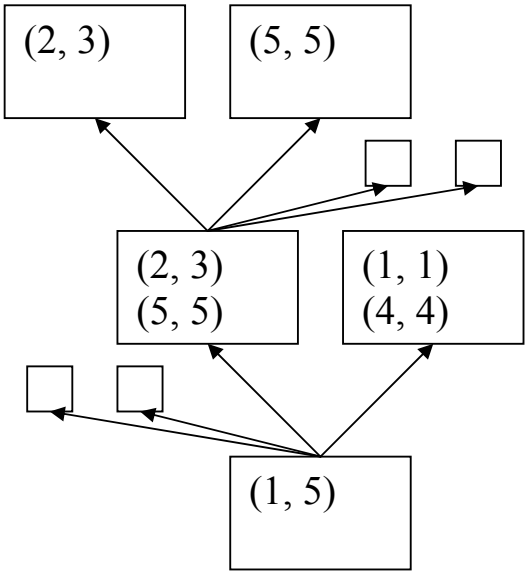


Z – Kurve der Stufe 1 im Raum

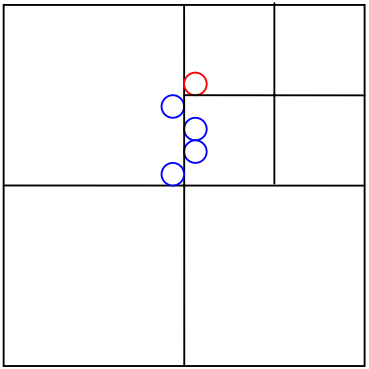
Einfügen eines Punktes in den Baum

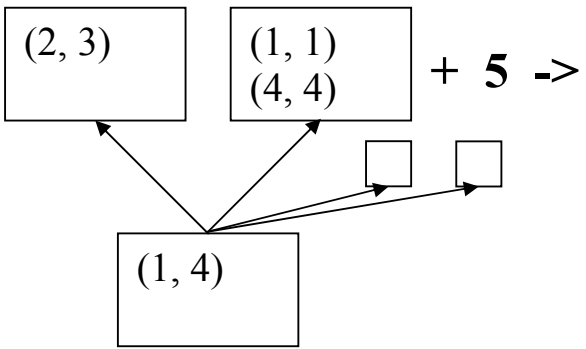


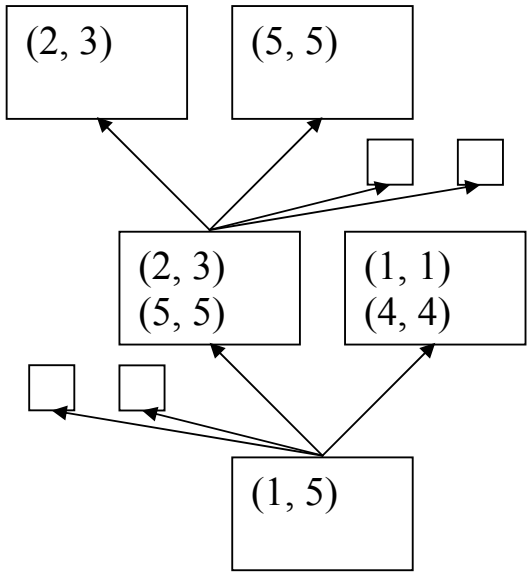




Einfügen eines Punktes in den Baum







Verarbeitung der Gesamtdatenmenge

Zelle(Punkteliste, x, y, z):

Neuen Knoten im Baum erstellen

Punkteliste im Blatt eintragen

Anzahl der Punkte in Zelle > Maximalwert

Ja

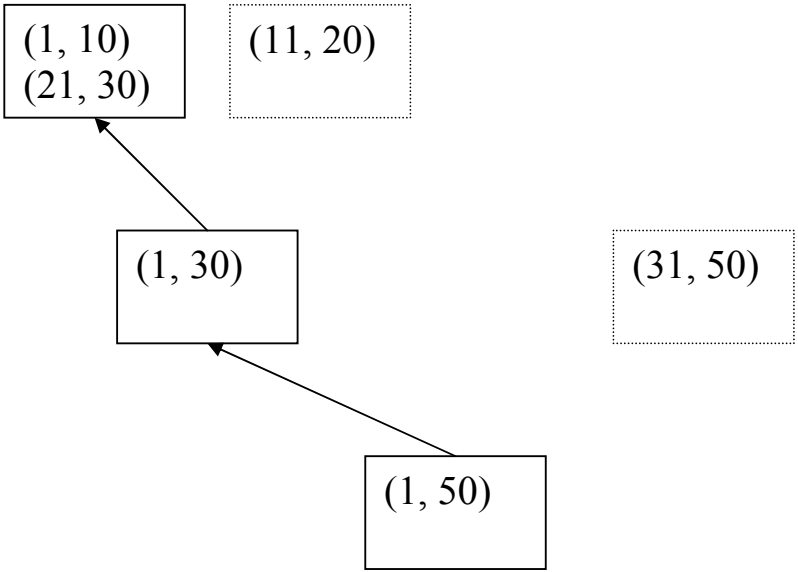
Nein

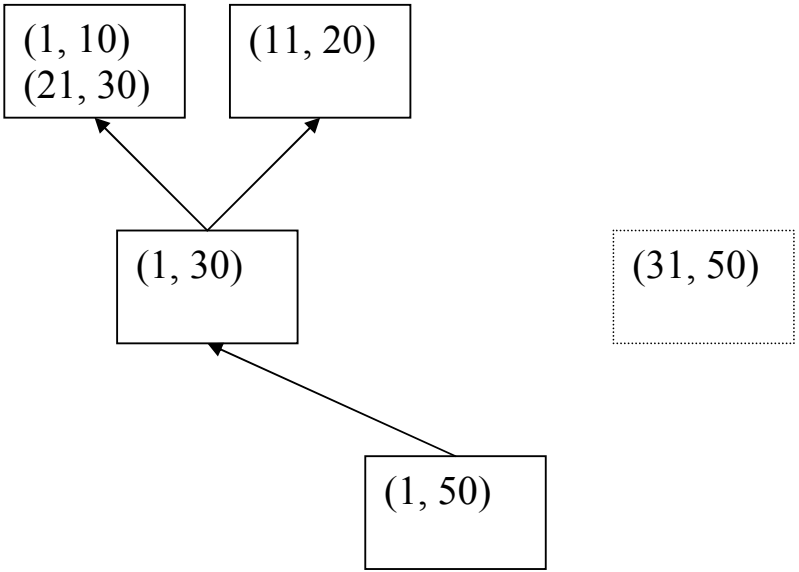
Ermittle für jeden Sektor den x, y und z Bereich

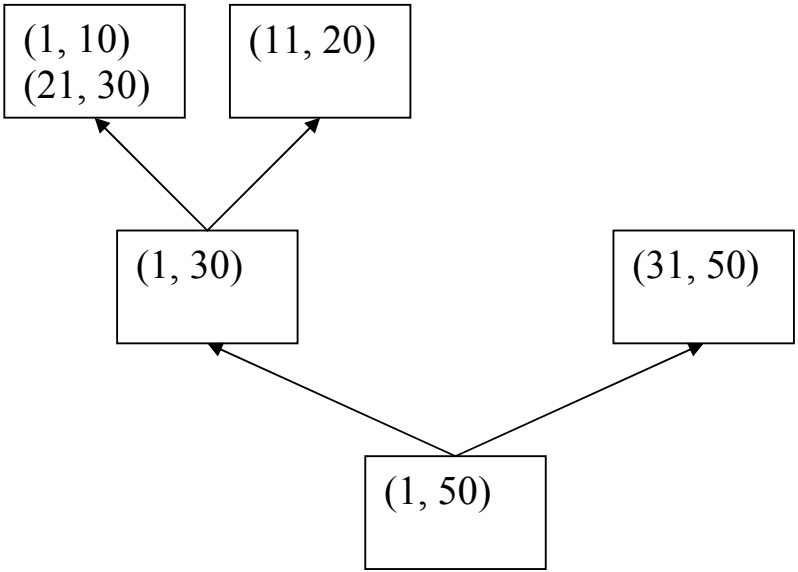
Wiederhole für alle 8 Sektoren

- Ermittle die Punkte, die in diesem Sektor sind
- Zelle(Punkteliste, x, y, z)

Ausgabe des Blättergraph


(1, 10)
(21, 30)
(11, 20)
(1, 30)
(31, 50)
(1, 50)


(1, 10)
(21, 30)
(11, 20)
(1, 30)
(31, 50)
(1, 50)


(1, 10)
(21, 30)
(11, 20)
(1, 30)
(31, 50)
(1, 50)

Laufzeitabschätzung

$$\sum_{x=0}^{\log(n)} (8\hat{}\, x) * n / (8\hat{}\, x)$$

$$= \sum_{x=0}^{\log(n)} n = \mathrm{n} * \sum_{x=0}^{\log(n)} 1$$

$$= O(n \log(n))$$

Vorraussetzung: Baum etwa ausbalanciert

Sequentielles Einfügen

Für alle Punkte

    Füge ein

Füge ein

Suche Blatt	O(log n)
Update entlang des Weges	O(log n)

Anzahl der Punkte in Blatt > Maximalwert

Ja | Nein

Splitte das Blatt in 8 Blätter
Einfügen Update Liste

Laufzeitabschätzung

Das Update der Knoten geschieht in konstanter Zeit, da der aktuelle Punkt nur mit dem letzten Listenelement, auf das ein Zeiger zeigt, verglichen werden muß.

Vorraussetzung: Baum etwa ausbalanciert

$$\sum_{x=1}^{n} \log(x) = \log(\prod_{x=1}^{n} x) = \log(n!)$$

Stirlingsche Formel: n! = (n/e)^n * $\sqrt{2\pi * n}$ * e^(1/12n)

log(n!) = log((n/e)^n * $\sqrt{2\pi * n}$ * e^(1/12n))

= n log(n) + ½ log(2π*n) + 1/12n

= O(n log(n)) → noch besser? → sortieren?

Sortierverfahren

- Einfügen = Umsortieren der Daten

- Deutung als Sortierverfahren, bei dem die Punkte nach den Sektoren und Subsektoren sortiert werden.

- Um aus dem Baum eine sortierte Liste zu machen, muß man nur sequentiell die Blätter des Baumes auslesen und in einer Liste speichern.

→ Da bewiesen ist, daß es kein schnelleres allgemeines Sortierverfahren als mit O(n log(n)) gibt, folgt:

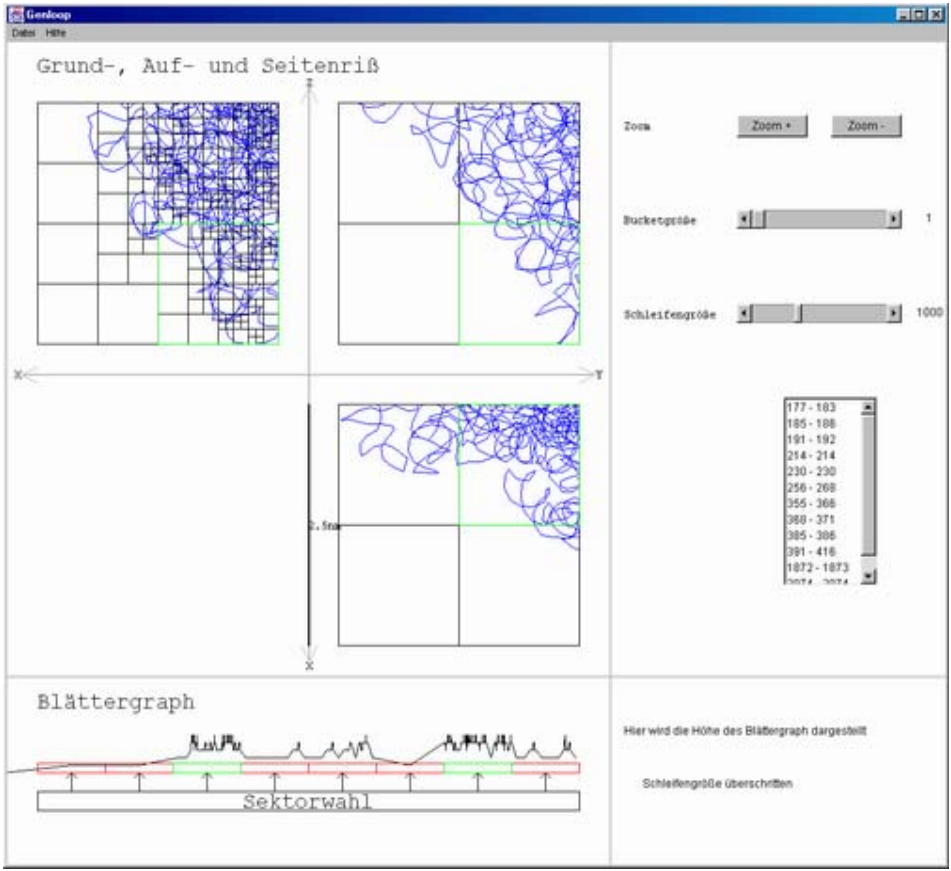
Die minimale Komplexität für die Erstellung des Baumes beträgt O(n log(n)). Somit gibt es für dieses Problem auch keinen besseren allgemeinen Algorithmus.

Resultate

- Näherungen von Teilen der DNA sind im Blättergraph als Spitzen sichtbar
- Laufzeit zur Ermittlung von Nachbarschaftsbeziehungen beträgt: n log(n)
- Basen im gleichen Sektor können in logarithmischer Zeit gefunden werden

Ausblick

- Programm zur Visualisierung
    - Möglichkeit zum Einstellen der Bucket-Größe
    - Anzeigefunktion
- Test an realen Daten
- Optimierung der Parameter


Genloop
Datei Hilfe
Grund-, Auf- und Seitenriß
Z
X
Y
Zoom
Zoom +
Zoom -
Bucketgröße
1
Schleifengröße
1000
177 - 183
185 - 188
191 - 192
214 - 214
230 - 230
256 - 268
355 - 366
368 - 371
385 - 386
391 - 416
1872 - 1873
Blättergraph
Sektorwahl
Hier wird die Höhe des Blättergraph dargestellt
Schleifengröße überschritten