UA RUHR

Algorithmen auf Sequenzen

Einführung

Sven Rahmann

Genominformatik
Universitätsklinikum Essen
Universität Duisburg-Essen
Universitätsallianz Ruhr

Autoren

UA RUHR

Folien:

- Dominik Kopczynski
- Sven Rahmann

Skript:

- Tobias Marschall
- Marcel Martin
- Dominik Kopczynski
- Sven Rahmann

Autoren

Folien:

- Dominik Kopczynski
- Sven Rahmann

Skript:

- Tobias Marschall
- Marcel Martin
- Dominik Kopczynski
- Sven Rahmann

Finanzierung:

- Mercator Research Center Ruhr:
  UA-Ruhr-Professur Sven Rahmann (Pe-2013-12)

- Qualitätsverbesserungskommission (Quest)
  Fakultät für Informatik, TU Dortmund

Warum Algorithmen auf Sequenzen?

UA RUHR

- Sequenzen kommen überall vor:
  Text,
  Bild- / Audio- / Videodaten,
  Biologische Sequenzen (DNA, Proteine)
- Als Informatiker/in hat man entsprechende Algorithmen ständig um sich:
  Code-Parser, Syntax-Highlighting, Korrektur- bzw. Codevorschläge in IDEs.
- Algorithmen sind elementar und elegant
- Wenig Vorkenntnisse nötig
- Spaß!

Einführung

Beispiele für Sequenzen sind:

- Biosequenzen (DNA, RNA, Proteine)
- Texte (Literatur, wissenschaftliche Texte)
- Quelltexte von Programmen
- Dateien, Datenströme
- Zeitreihen, Spektren (Audiosignale, Massenspektren, ...)
- Bilder, Videos

Einführung

UA RUHR

Fragestellungen der Sequenzanalyse:

- Mustersuche: exakt oder fehlertolerant
- Sequenzvergleich:
  Ermitteln von Gemeinsamkeiten und Unterschieden
- Kompression:
  Erkennung und Nutzung von Wiederholungen
- Muster- und Signalentdeckung:
  Suche nach unbekannten auffälligen Mustern

Begriffe und Definitionen

Definitionen:

- Ein **Alphabet** Σ ist eine (endliche oder unendliche) Menge.
- Eine **Indexmenge** $\mathcal{I}$ ist eine endliche oder abzählbar unendliche linear geordnete Menge. Beispiele hierfür sind $\mathbb{N}, \mathbb{Z}$ und die Ordnung $\leq$.
- Eine **Sequenz** ist eine Funktion $s : \mathcal{I} \to \Sigma$ oder äquivalent, ein Tupel $s \in \Sigma^{\mathcal{I}}$.

Wir befassen uns ausschließlich mit endlichen Sequenzen,
also $\mathcal{I} = \{0, \ldots, n-1\}$ für $n \in \mathbb{N}$ (Indizierung beginnt bei 0).
Zur Vereinfachung schreiben wir Σ^n.

Begriffe und Definitionen

Beispiele für Sequenzen über verschiedenen Alphabeten:

Sequenztyp	**Alphabet Σ**
DNA-Sequenz	$\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$
Protein-Sequenz	20 Standard-Aminosäuren
C-Programme	ASCII-Zeichen (7-bit)
Java-Programme	Unicode-Zeichen
Audiosignal (16-bit samples)	$\{0,\ldots,2^{16}-1\}$
Massenspektrum	Intervall $[0,1]$ (unendlich) oder double

Begriffe und Definitionen

UA RUHR

Definitionen:

- Die Elemente von Σ^n nennt man **Wörter**, **Tupel**, **Strings** oder **Sequenzen** der Länge n, sowie n-**Mere** oder n-**Gramme**.
- Sei $\Sigma^+ := \bigcup_{n \geq 1} \Sigma^n$ die Menge aller endlichen Strings der Länge $n \geq 1$.
- Sei $\Sigma^* := \bigcup_{n \geq 0} \Sigma^n$ die Menge aller endlichen Strings über Σ, wobei $\Sigma^0 = \{\epsilon\}$ und ϵ der **leere String** ist.

Begriffe und Definitionen

UA RUHR

Buchstabe, Teilstring, Teilsequenz

Sei $s \in \Sigma^*$ ein String.

- Sei $s_i = s[i]$ der **Buchstabe**, der in s an Position i steht ($i \in \mathcal{I}$)
- $s[i:j]$ ist der **Teilstring** von i (einschließlich) bis j (ausschließlich). Falls $i \geq j$, sei $s[i:j] = \epsilon$.
- $s[i..j]$ ist der **Teilstring** von i bis j (jeweils einschließlich) oder ϵ falls $i > j$.
- **Teilsequenzen** von s sind $(s_i)_{i \in I}$ für beliebige $I \subset \mathcal{I}$.

Eine Teilsequenz ist im Gegensatz zum Teilstring also nicht notwendigerweise zusammenhängend. Die Begriffe Teilstring und Teilsequenz sind daher auseinanderzuhalten.

Begriffe und Definitionen

UA RUHR

Präfix, Suffix

- Sei $s[:i] := s[0:i]$ das **Präfix** der Länge i von s.
- Sei $s[i:] := s[i:|s|]$ das **Suffix** der Länge $|s| - i$ von s.

Begriffe und Definitionen

UA RUHR

Präfix, Suffix

- Sei $s[:i] := s[0:i]$ das **Präfix** der Länge i von s.
- Sei $s[i:] := s[i:|s|]$ das **Suffix** der Länge $|s| - i$ von s.
- Analog: $s[..i]$ ($s[i..]$) ist das Präfix bis (Suffix ab) Position i

Begriffe und Definitionen

Präfix, Suffix

- Sei $s[:i] := s[0:i]$ das **Präfix** der Länge i von s.
- Sei $s[i:] := s[i:|s|]$ das **Suffix** der Länge $|s| - i$ von s.
- Analog: $s[..i]$ ($s[i..]$) ist das Präfix bis (Suffix ab) Position i

Ein Präfix (Suffix) t von s mit $t \neq \epsilon$ und $t \neq s$ heißt **echtes Präfix (Suffix)**.

Anregungen

UA RUHR

- Vorlesung nicht nur anhören, sondern damit arbeiten:
  Varianten ausdenken, kleine Lemmas beweisen, ...
- Algorithmen implementieren oder mit Stift und Papier durchführen
- Das Skript durcharbeiten
- Zusätzliche Literatur suchen (Bibliotheken, WWW)
- Originalveröffentlichungen mit Algorithmen lesen

Literaturvorschläge

- Gonzalo Navarro, Mathieu Raffinot
  Flexible Pattern Matching in Strings, Cambridge University Press
- Dan Gusfield
  Algorithms on Strings, Trees and Sequences, Cambridge University Press
- David Sankoff und Joseph P. Kruskal
  Time Warps, String Edits, and Macromolecules, University of Chicago Press