Publ. I.R.M.A. Strasbourg, 1987, 340/S–15
Actes 15^{e} Séminaire Lotharingien, p. 77–78

Zugriffsstrategien für Gendatenbanken

R. Laue, Lehrstuhl II für Mathematik,

Universität Bayreuth, 8580 Bayreuth

Kurzfassung *

Um zu einer vorgelegten kurzen Testsequenz alle gleichen oder ähnlichen Vorkommen in einer großen Datenbank von bekannten langen Sequenzen zu finden, wird ein zweistufiges Verfahren vorgeschlagen. Zuerst werden den langen Sequenzen lokale Eigenschaftsvektoren zugeordnet. Diese Eigenschaftsvektoren vergröbern die ursprüngliche Struktur zu Punkten im Eigenschaftsraum. Diese Punkte werden vom Grid-File-System [1,2] verwaltet, so daß auf benachbarte Punkte leicht zugegriffen werden kann. Ein Zugriff erfolgt nun durch Errechnung der lokalen Eigenschaftsvektoren der Testsequenz und Aufsuchen der entsprechenden Punkte oder ihrer Nachbarn im Eigenschaftsraum. Jeder gefundene Punkt zeigt nun auf eine kleine Datei mit den Adressen der Orte in der Datenbank, an denen der lokale Eigenschaftsvektor gerade durch diesen Punkt dargestellt wird. An den gefundenen Orten stehen nun Kandidaten für den konventionellen exakten Vergleich mit der Testsequenz. Da die konventionellen Methoden mit mindestens linearem Aufwand (proportional zur Länge der Datenbank) arbeiten, kann durch ein solches Vorgehen die Suchzeit erheblich verkürzt werden. Entscheidend ist dabei die Selektionswirkung der Eigenschaftsvektoren.

Um eine gute Selektionswirkung zu erlangen, wird vorgeschlagen, Häufigkeiten von Teilwörtern als Eigenschaften zu verwenden. Es werden zunächst die Möglichkeiten untersucht, vier Teilwörter zu wählen. Jede Kombination der Teilwörter läßt sich durch einen Graphen darstellen, so daß alle Umbenennungen des zugrunde liegenden Alphabets $\{A, C, G, T\}$ gerade zu den isomorphen Graphen führen. Da bei gleichverteilten Daten die Güte der Selektionswirkung einer Wahl von Teilwörtern invariant ist gegenüber Umbenennung des Alphabets, reicht es, nur für jeden Isomorphietyp von Graphen diese Selektionswirkung zu ermitteln.

Eine Computerauswertung zu lokalen Eigenschaftsvektoren für Abschnitte der Länge 10 ermittelt hier die günstigste Strategie. Es wird ebenfalls untersucht, wie sich diese Strategie durch Betrachtung von fünf Teilwörtern

* Eine ausführliche Version wird an anderer Stelle erscheinen

verbessern läßt. Bei Gleichverteilung entspricht bei der besten Strategie dann ein Eigenschaftsvektor im Mittel nur 0,049 % und maximal 0,52 % der Daten. Für ein Testwort der Länge 19 werden 9 Vergleiche von Teilwörtern nötig, so daß im Mittel nicht mehr als ein halbes Prozent der Daten noch dem exakten Vergleich unterzogen werden müssen.

Literatur

[1] H. Hinterberger, J. Nievergelt, K.C. Sevcik.
The grid file: an adaptable, symmetric multikey file structure
ACM TODS 9 (1984), 38-71.

[2] R. Laue
Abbildungen und Algorithmen.
Séminaire Lotharingien de Combinatoire, Schney, 1986.