

Algorithmische Methoden zur Kartierung von DNA-Sequenzen

Andreas Fink, TU Braunschweig

Zusammenfassung

In der Molekularbiologie traten in den letzten Jahren vermehrt diskrete Optimierungs- und Zuordnungsprobleme im Bereich der Untersuchung des Erbguts und der Proteine auf. Dies gilt insbesondere für das *Human Genome Project*, dessen primäres Ziel die Erforschung der genetischen Information des Menschen ist. Hierzu gehört die vollständige Sequenzierung des Genoms, d. h. die Erforschung von Nucleotidsequenzen bzw. deren Funktion. Hier wird eine spezifische Problemstellung betrachtet, deren Ziel die Ermittlung der realen Ordnung von Klonen (Teilabschnitten von Chromosomen) mit Hilfe von experimentell ermittelten – in der Regel fehlerbehafteten – Daten ist (STS-Kartierungsstrategie). Die sich daraus ergebende kombinatorische Problemstellung wird diskutiert, und es werden verschiedene Methoden zur Generierung von Lösungen sowie zur Identifikation experimenteller Fehler entwickelt und analysiert.

Abstract

In molecular biology discrete optimization and assignment problems occur with increasing importance in the fields of the examination of the genome and of protein structures. This is especially valid for the *Human Genome Project*, where the primary goal is to discover the human genetic information. This involves the determination of the genome sequence, i. e. to explore the nucleotid sequence and their function, respectively. Here we consider a specific problem, where the aim is to determine the real ordering of clones (partial sequences of chromosomes) with the help of experimentally determined – generally faulty – data (STS sequencing strategy). The resulting combinatorial problem is discussed and several methods for the generation of solutions as well as for the identification of experimental faults are developed and analyzed.

1 Einleitung

In der Molekularbiologie traten in den letzten Jahren vermehrt diskrete Optimierungs- und Zuordnungsprobleme im Bereich der Untersuchung des Erbguts (Genoms) und der Proteine auf. Im Zusammenhang hiermit wurde der Begriff *Computational Molecular Biology* geprägt; vergleiche [5, 9, 10].

Diese Problemstellungen können teilweise mit Methoden des Operations Research, der Mathematik bzw. der Informatik angegangen werden. Die hier zusammengefaßte Arbeit [3] beschäftigt sich mit einer speziellen Problemstellung – bzw. mit der Entwicklung und Untersuchung von Algorithmen zu deren Lösung – im Rahmen der Analyse von Chromosomen.¹ Ein zukünftiges Ziel im Rahmen des *Human Genome Project* ist die vollständige Sequenzierung des Genoms, d. h. die Erforschung von Nucleotidsequenzen bzw. deren Funktion; als Zwischenschritt hierzu gilt die Konstruktion von sogenannten physischen Karten (*physical maps*) von Chromosomen. Dabei werden Positionsinformationen für verschiedene spezifische DNA-Teilsequenzen ermittelt. Bei der STS-Kartierungsstrategie, die im Rahmen des *Human Genome Project* u. a. angewendet wird, wird hierzu das Chromosom in überlappende Fragmente (Klone) aufgespalten, über die durch sogenannte Hybridisationsexperimente mit Sonden (spezifischen sehr kurzen DNA-Teilabschnitten) Informationen gewonnen werden,

¹Zu näheren Einzelheiten zur Molekulargenetik, die für das vollständige Verständnis der Problemstellung erforderlich sind, kann hier nur auf eines der grundlegenden Lehrbücher auf diesem Gebiet verwiesen werden (z. B. [7]).

um die relative Anordnung der Klone und Sonden auf dem Chromosom zu ermitteln. Die sich hierbei ergebende kombinatorische Problemstellung wird im Rahmen der Arbeit untersucht; Problemmodellierung, Verfahrensentwicklung, Implementierung von Algorithmen und analytische sowie experimentelle Bewertung bilden die groben Arbeitsschritte.

Bei bisherigen Kartierungsprojekten dieser Art – vgl. [2, 4] – wurden oftmals einfache exakte oder heuristische Lösungsverfahren verwendet, die zumeist nicht näher dokumentiert sind. Die Zielsetzung ist es nun, aufbauend auf den ersten bisherigen Ergebnissen, Methoden des Operations Research (Modellierung, *Local Search*, graphentheoretische Verfahren, etc.) geeignet auf die gegebene Problemstellung anzupassen bzw. zu erweitern. Wichtig ist dabei insbesondere die Berücksichtigung von stark fehlerbehafteten Daten. Während in der Arbeit gezeigt wird, daß für fehlerfreie Daten verschiedene einfache Verfahren optimale Lösungen liefern, scheitern diese teilweise bei fehlerbehafteten Daten, so daß bei unscharfen Daten modifizierte Methoden angewendet werden müssen; diese müssen sowohl analytisch als auch experimentell untersucht werden. Die dabei gewonnenen Ergebnisse sollen auch für zukünftige Kartierungsprojekte Richtlinien für die Gestaltung von Experimenten ergeben: sowohl bezüglich qualitativer (wie groß darf die Fehlerrate sein?) als auch quantitativer (wieviele Experimente müssen durchgeführt werden?) Aspekte.

2 Problemstellung

In diesem Abschnitt wird die zugrunde liegende Problemstellung – zum Teil vereinfacht – beschrieben. Ziel ist die Untersuchung von Chromosomen, Strukturen innerhalb des Zellkerns, die jeweils ein stark verdichtetes DNA-Molekül enthalten. Die DNA (englische Abkürzung für Desoxyribonucleinsäure) stellt ein Polymermolekül dar, das aus vier verschiedenen Typen molekularer Bausteine besteht: Adenin, Thymin, Guanin und Cytosin (Nucleotide; abgekürzt A, T, G und C). Ein menschliches Chromosom enthält ca. 10^8 dieser Nucleotide. Die DNA speichert in der Sequenz ihrer Bausteine die genetische Information. Die Chromosomen können mit Hilfe verschiedener Enzyme (Restriktionsendonucleasen), die DNA-Moleküle an bestimmten spezifischen Nucleotidsequenzen aufspalten, in „handhabbare“ Fragmente zerlegt werden, die dann näher untersucht werden können. Diese Fragmente werden vervielfältigt; man erhält damit eine große Anzahl von Fragmentkopien, die als Klone (*clones*) bezeichnet werden.

Ein primäres Ziel ist es nun, Informationen über die einzelnen Klone zu gewinnen, und hiermit das vollständige DNA-Molekül (Chromosom) als eine geordnete Menge überlappender Klone zu rekonstruieren. Bei der STS-Kartierungsstrategie verwendet man hierzu sogenannte Sonden (*probes*). Diese stellen kurze DNA-Teilabschnitte (Oligonucleotide) des zu untersuchenden Chromosoms dar, deren Nucleotidsequenz bekannt ist, und die mit radioaktiven Atomen, fluoreszierenden Molekülen oder anderen leicht nachweisbaren chemischen Gruppen markiert sind. Hiermit ist es möglich, Klone, die Teilsequenzen enthalten, die zur Nucleotidsequenz der Sonde komplementär sind, über den Vorgang der Hybridisation, d. h. über das Anlagern der Sonde an komplementäre Teilabschnitte von Klonen, zu identifizieren. Jede Sonde besitze dabei eine genügende Länge, so daß sie eine eindeutige Position (*sequence tagged site*) auf dem DNA-Strang bestimmt.

In Experimenten wird nun untersucht, welche Sonden mit welchen Klonen hybridisieren (d.h., welche Sonden sich über komplementäre Sequenzen an welche Klone anlagern). Hieraus lassen sich Informationen über das Überlappen einzelner Klone ableiten, woraus eine überlappende Anordnung einzelner Klon-Teilmengen bzw. im Idealfall die relative Anordnung aller Klone bzw. Sonden ermittelt werden kann, die im weiteren auch als reale Ordnung oder Permutation bezeichnet wird.

Die der Problemstellung zugrunde liegende Struktur kann somit bei einer Sondenmenge $\mathcal{P} = \{P_1, \dots, P_m\}$ mit $m = |\mathcal{P}|$ und einer Klonmenge $\mathcal{C} = \{C_1, \dots, C_n\}$ mit $n = |\mathcal{C}|$ durch eine $m \times n$ -Matrix H beschrieben werden:

$$H = (h_{ij})_{m \times n}, \quad h_{ij} = \begin{cases} 1 & \text{falls die Nucleotidsequenz von Sonde } P_i \text{ komplementär} \\ & \text{zu einer Teilsequenz von Klon } C_j \text{ ist,} \\ 0 & \text{sonst.} \end{cases}$$

Bei den Experimenten werden jedoch in der Regel fehlerbehaftete Daten gewonnen. Insbesondere ergeben sich irrtümliche Hybridisationen (*false positives*), obwohl keine komplementären Sequenzen vorliegen, sowie fehlende Hybridisationen (*false negatives*), obwohl komplementäre Sequenzen vorhanden sind. D. h., man erhält eine fehlerbehaftete Hybridisationsmatrix \tilde{H} ; Ziel ist nun die Ermittlung einer Sonden-Permutation $\Pi^P = (\pi_1, \dots, \pi_m)$ bzw. einer Klon-Permutation $\Pi^C = (\pi_1, \dots, \pi_n)$, die der realen Ordnung der Sonden bzw. Klone entspricht.

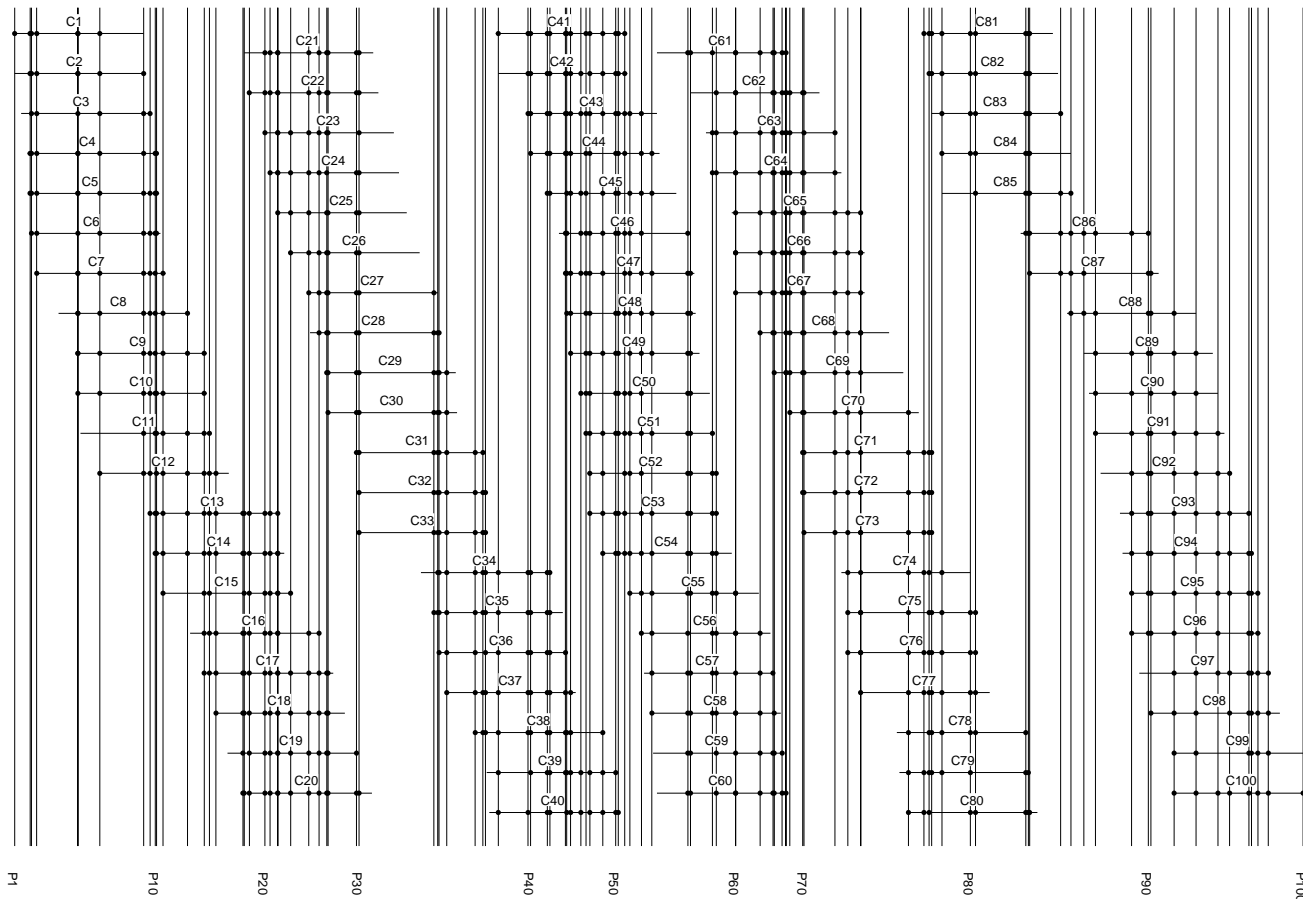


Abbildung 1: Beispiel für die Darstellung einer Problemstruktur.

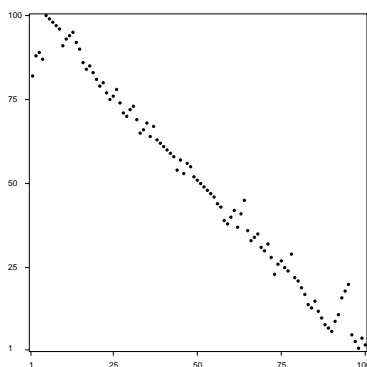


Abbildung 2: Beispiel für die Darstellung einer generierten Permutation.

Zur Visualisierung einer Probleminstanz vergleiche Abbildung 1. Klone (Sonden) sind über horizontale (vertikale) Linien dargestellt; fehlende Punkte in den Schnittpunkten von Klonen und Sonden kennzeichnen *false negatives*, *false positives* werden in der Abbildung nicht dargestellt. In Abbildung 2 ist eine mögliche Lösung dargestellt, die durch eine der im folgenden beschriebenen Methoden erreicht werden kann. Ein Punkt an Position (u, v) in der Lösungsdarstellung heißt, daß Objekt X_v in der Lösung an Position u eingeordnet ist (wobei X_1, \dots, X_n die reale Ordnung sei).

3 Vorgehensweise und Methodenbeschreibung

3.1 Grundlagen

Ausgangspunkt der Verfahren sind Klon-Klon- oder Sonde-Sonde-Korrelationsmatrizen $C = (c_{jk})_{n \times n}$, die geeignet zu definieren bzw. aus der Hybridisationsmatrix \tilde{H} abzuleiten sind. Die Matrixelemente c_{jk} sollen dabei ein Maß für die „Nähe“ der betrachteten Objekte (Klone oder Sonden) darstellen. Ziel ist nun die Ermittlung einer Ordnung (Permutation) der Objekte, bei der Objekte mit hoher Korrelation möglichst „nahe“ zusammen angeordnet werden. Hierzu wurden verschiedene Korrelationsmaßstäbe entwickelt.

Für jeden Klon C_j sei H_j^C als die experimentell ermittelte Menge der Sonden definiert, die mit dem Klon C_j hybridisieren.² Es wurden sechs alternative Möglichkeiten einer Korrelation $\gamma_r^C : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ definiert; als Beispiel sei die Definition von $\gamma_1^C(C_j, C_k)$ angegeben:

$$\gamma_1^C(C_j, C_k) := \frac{|H_j^C \cap H_k^C|}{|H_j^C \cup H_k^C|}.$$

Man kann zeigen, daß bei fehlerfreien Daten gewisse dieser Korrelationsfunktionstypen die Eigenschaft erfüllen, daß für alle Objekte die Korrelationen mit anderen Objekten entsprechend der realen Ordnung auf beiden Seiten monoton fallen.

3.2 Varianz-analytisches Verfahren

Ziel dieses Verfahrens (siehe [8]) ist es, ausgehend von einer Korrelationsmatrix C , n Objekte auf einer linearen Skala so zu positionieren – hieraus ergibt sich implizit auch eine Ordnung –, daß hochkorrelierte Objekte relativ nahe zusammen positioniert werden. Bei Bezeichnung der Position von Objekt i mit x_i , $1 \leq i \leq n$, sowie $\mathbf{x} = (x_1, \dots, x_n)^T$, wird von folgender Problemstellung ausgegangen: Minimiere $S(\mathbf{x}) := 1/2 \sum_{i=1}^n \sum_{j=1}^n c_{ij} (x_i - x_j)^2$ unter der Nebenbedingung $\sum_{i=1}^n x_i^2 = 1$.

Diese Problemstellung stellt ein Extremalproblem mit Nebenbedingungen dar, das mittels der Lagrangeschen Multiplikatormethode gelöst werden kann. Die Lösung ergibt sich aus dem dem betragskleinsten Eigenwert ungleich Null entsprechenden Eigenvektor, der als Positionsvektor interpretiert wird (bei Betrachtung einer Matrix D , die sich über die Korrelationsmatrix ergibt). Zur Lösung des Eigenwertproblems für die Matrix D wird ein zyklisches Jacobi-Verfahren in Verbindung mit einer Schwellenwertmethode angewendet. Dabei handelt es sich um eine iterative, approximative Methode, die für relativ dünn besetzte symmetrische Matrizen ohne Bandstruktur gut geeignet und weiterhin numerisch relativ stabil ist.

Das varianz-analytische Verfahren eignet sich jedoch in der Regel nur bei Daten, die im wesentlichen keine auf *false positives* zurückgehenden fehlerhaften Hybridisationen enthalten (vgl. [3]).

3.3 Heuristische Verfahren

Ziel ist die Ermittlung einer Permutation Π der betrachteten Objekte (Klone oder Sonden) zur Maximierung einer zu definierenden Bindungsenergiefunktion $\Gamma(\Pi)$, die ein Maß für kumulierte Korrelationen „nahe zusammen“ angeordneter Objekte darstellen soll. Hierzu wurden verschiedene Bindungsenergiefunktionen entwickelt; diese lassen sich primär in additive und multiplikative Typen untergliedern. Eine typische additive Funktion wurde folgendermaßen definiert:

$$\Gamma_d^+(\Pi) := \sum_{i=1}^{n-1} \sum_{j=i+1}^{\min\{i+d, n\}} \frac{d-j+i+1}{d} c_{\pi_i, \pi_j}.$$

²Die weitere Beschreibung beschränkt sich auf diese Sichtweise mit Klonen als primär zu sortierenden Objekten; in [3] wurden sämtliche Untersuchungen und Ergebnisse auch für Sonden als primäre Objekte durchgeführt.

Bei einem Parameter $d > 1$ werden dabei Skalierungen vorgenommen, um den Beitrag einer Korrelation bei größerem „Abstand“ niedriger zu bewerten. Für den Wert d erscheint als Größenordnung die Anzahl der Objekte, mit denen ein Objekt bei fehlerfreien Daten im Mittel auf beiden Seiten Korrelationen größer als Null besitzt, sinnvoll. Man kann zeigen, daß gewisse dieser Funktionen bei fehlerfreien Daten ihr Maximum bei der realen Ordnung der Objekte annehmen.

Zur Maximierung dieser Funktionen wurde zum einen ein einfaches Greedy-Eröffnungsverfahren untersucht und implementiert. Dieses Verfahren generiert bei einer fehlerfreien, zusammenhängenden Datenbasis bei gewissen Zielfunktionen der realen Objektordnung äquivalente Permutationen. Weiterhin wurde ein 3-opt-Verbesserungsverfahren entwickelt; siehe [3] für nähere Einzelheiten. Wichtig sowohl beim Eröffnungs- als auch insbesondere beim Verbesserungsverfahren ist die effiziente Ausführung der jeweiligen Operationen über eine adaptive Neuberechnung der Zielfunktion mit minimalem Aufwand.

3.4 Screening

Es wurden verschiedene Methoden entwickelt, die versuchen, irrtümliche Hybridisationen soweit wie möglich im vorhinein zu identifizieren. Zum einen bezweckt dies, den oben beschriebenen Methoden „bessere“ Daten zur Verfügung zu stellen; zum anderen ist es in der praktischen Anwendung möglich, zweifelhafte Hybridisationen erneut experimentell zu testen.

Bei dem Verfahren *Screening durch Bestimmung kürzester Umwege* wird ein Graph definiert, der für alle Sonde-Klon-Paare (P_i, C_j) genau dann eine Kante besitzt, falls diese im Experiment hybridisierten. Die Idee besteht nun darin, jede Kante $(P_i, C_j) \in E$ auf eine irrtümliche Hybridisation zu testen, indem man einen kürzesten Weg – d. h. einen Weg mit minimaler Kantenanzahl – von C_j nach P_i in dem durch Löschen dieser Kante reduzierten Graphen bestimmt. Handelt es sich bei der jeweils betrachteten Kante um eine irrtümliche Hybridisation, so wird dieser *Umweg* in der Regel relativ lang sein, woraufhin die Kante bzw. die Hybridisation als irrtümlich eingestuft wird.

Beim *Screening durch Bewertung der Klon-Nachbarschaft* wird eine Klon-Klon-Nachbarschaftsmatrix $A_{n \times n}$ für alle $j, k = 1, \dots, n$ durch $a_{jk} := |H_j^C \cap H_k^C|$ definiert. Zwei Klone C_j und C_k werden genau dann als benachbart bezeichnet, wenn a_{jk} echt positiv ist. Irrtümliche Hybridisationen eines Klons C_j führen nun in der Regel dazu, daß sich eine gewisse Anzahl bei korrekten Daten nicht benachbarter Klone C_k mit relativ kleinen positiven Einträgen a_{jk} ergibt; relativ große a_{jk} sind dagegen ein Anzeichen für eine tatsächliche Überlappung der entsprechenden Klone. Hieraus ergibt sich die Idee, für jede Hybridisation eines Klons C_j die Veränderung relativ kleiner Einträge in A bei einer Exklusion dieser Hybridisation zu bewerten. Übersteigt eine geeignete Bewertung solcher Veränderungen eine vorzugebende Schranke, so wird die Hybridisation als irrtümlich angesehen.

Die Verfahren wurden analytisch untersucht, um Anhaltspunkte für deren Effektivität zu gewinnen. Hierzu wurden Abschätzungen für die Wahrscheinlichkeiten hergeleitet, daß durch das Screening zum einen korrekte Hybridisationen als irrtümlich eingestuft werden, sowie zum anderen irrtümliche Hybridisationen nicht identifiziert werden.

4 Ergebnisse

Die Implementierung der Verfahren sowie von Routinen zur Visualisierung von Probleminstanzen bzw. Lösungen wurde in C++ portabel durchgeführt. Um eine systematische Untersuchung zu gewährleisten, wurden möglichst realitätsnahe Probleminstanzen über einen Problemgenerator erzeugt.

Die in der Arbeit dargestellten Algorithmen zur Lösung von Anordnungsproblemen im Rahmen der physischen Kartierung von Chromosomen erbrachten sehr gute Ergebnisse bei der Anwendung auf Probleminstanzen, bei denen genug „Redundanz“ in der Problemstruktur vorhanden war, so daß auch relativ hohe Fehlerraten möglich waren. Abhängig von der Problemstruktur und den Fehlerwahrscheinlichkeiten erbrachte bei schwierigen Problemen das 3-opt-Verbesserungsverfahren zur Maximierung gewisser Bindungsenergiefunktionstypen die besten Ergebnisse (vgl. [3]). Die hierbei er-

reichten Lösungen erscheinen für gut konditionierte Probleme sehr gut, sowie für „problematischere“ Probleminstanzen noch so gut – wenige Bruchstellen bei relativ langen korrekt geordneten Teilsequenzen –, daß eine vollständige Lösung durch darauf aufbauende Experimente erreicht werden kann. Im Zusammenhang hiermit ist darauf hinzuweisen, daß für eine vollständig korrekte Lösungsermittlung ein bestimmter Informationsgehalt der experimentellen Daten notwendig ist. Die theoretische Analyse solcher Bedingungen stellt ein offenes Problem dar (vgl. [6]). Die in [3] durchgeführten experimentellen Untersuchungen ergaben hierfür konkrete Ansatzpunkte.

Die Verbindung von Screening und Lösungsermittlung erbrachte keine Vorteile – anders als in [1], wo ein Screening irrtümlicher Hybridisationen eine Voraussetzung für die Ermittlung guter Lösungen darstellt. Die hier verwendeten Local-Search Verfahren erscheinen als relativ robust gegenüber einem gewissen Anteil fehlerhafter Daten; dies läßt sich durch die Verwendung einer Zielfunktion mit $d > 1$ begründen, wodurch einzelne irrtümliche Hybridisationen nur eine relativ geringe Auswirkung besitzen, da in die zu maximierende Zielfunktion die Korrelationen mit zweimal d anderen Objekten eingehen. Allerdings war der Informationsgehalt einiger untersuchter Probleminstanzen mit niedrigen Überdeckungsgraden teilweise so niedrig, daß eine vollständig korrekte Lösungsgenerierung für bereits mäßige Fehlerwahrscheinlichkeiten aus informationstheoretischen Gesichtspunkten unwahrscheinlich bzw. unmöglich erscheint (vgl. [3, S. 73f.]).

Literatur

- [1] Farid Alizadeh, Richard M. Karp, Deborah K. Weisser und Geoffrey Zweig. *Physical Mapping of Chromosomes Using Unique Probes*. Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms (SoDA), 489–500, 1994.
- [2] Ilya Chumakov, Philippe Rigault, Sophie Guillou, Pierre Ougen, Alain Billaut, Ghislaine Guasconi, Patricia Gervy, Isabelle LeGall, Pascal Soularue, Laurent Grinas, Lydie Bougueleret, Christine Bellané-Chantelot, Bruno Lacroix, Emmanuel Barillot, Philippe Gesnouin, Stuart Pook, Guy Vaysseix, Gerard Frelat, Annette Schmitz, Jean-Luc Sambucy, Assumpcio Bosch, Xavier Estivill, Jean Weissenbach, Alain Vignal, Harold Riethman, David Cox, David Patterson, Kathleen Gardiner, Masahira Hattori, Yoshiyuki Sakaki, Hitoshi Ichikawa, Misao Ohki, Denis Le Paslier, Roland Heilig, Stylianos Antonarakis und Daniel Cohen. *Continuum of overlapping clones spanning the entire human chromosome 21q*. Nature 359, 380–387, 1992.
- [3] Andreas Fink. *Algorithmische Methoden zur Kartierung von DNA-Sequenzen*. Diplomarbeit, Technische Hochschule Darmstadt, 1995.
- [4] Simon Foote, Douglas Vollrath, Adrienne Hilton und David Page. *The Human Y Chromosome: Overlapping DNA Clones Spanning the Euchromatic Region*. Science 258, 60–66, 1992.
- [5] Richard M. Karp. *Mapping the Genome: Some Combinatorial Problems Arising in Molecular Biology*. Proceedings of the 25th Annual ACM Symposium on Theory of Computing (SToC), 278–285, 1993.
- [6] Eric S. Lander und Michael S. Waterman. *Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis*. Genomics 2, 231–239, 1988.
- [7] Benjamin Lewin. *Genes V*. Oxford University Press, 1994.
- [8] Philip M. Morse. *Optimal Linear Ordering of Information Items*. Operations Research 20, 741–751, 1972.
- [9] Eugene W. Myers. *Guest Editor’s Foreword*, Algorithmica 13, Special Issue: Computational Molecular Biology, 1–6, 1995.
- [10] Michael S. Waterman. *Introduction to Computational Biology: Maps, sequences and genomes*. Chapman & Hall, 1995.