(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
28. Februar 2002 (28.02.2002)

PCT

(10) Internationale Veröffentlichungsnummer
WO 02/17127 A2

(51) Internationale Patentklassifikation[7]: G06F 17/20

(21) Internationales Aktenzeichen: PCT/EP01/09710

(22) Internationales Anmeldedatum:
22. August 2001 (22.08.2001)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
100 41 325.0 23. August 2000 (23.08.2000) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: **GENPROFILE AG** [DE/DE]; Robert-Rössle-Strasse 10, 13125 Berlin (DE).

(72) Erfinder; und
(75) Erfinder/Anmelder *(nur für US)*: **TERHALLE, Werner** [DE/DE]; c/o GenProfile AG, Robert-Rössle-Strasse 10, 13125 Berlin (DE).

(74) Anwalt: **V. BEZOLD & SOZIEN**; Akademiestrasse 7, 80799 München (DE).

(81) Bestimmungsstaaten *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Bestimmungsstaaten *(regional)*: ARIPO-Patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), eurasisches Patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI-Patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Veröffentlicht:
— *ohne internationalen Recherchenbericht und erneut zu veröffentlichen nach Erhalt des Berichts*

Zur Erklärung der Zweibuchstaben-Codes und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

(54) **Title:** METHOD AND DEVICE FOR THE CORRELATION ANALYSIS OF DATA SERIES

(54) **Bezeichnung:** VERFAHREN UND VORRICHTUNG ZUR KORRELATIONSANALYSE VON DATENFOLGEN

(57) **Abstract:** A method for processing data series, comprising a number of data points in a predetermined sequence of positions is disclosed, comprising the following steps: determination of correlation values for all doublets, triplets, or n-tuplets of positions in a set of data series on the basis of a pre-determined correlation scale, determination of position weightings from the correlation values for each position of the data series, determination of groups of positions correlated together in the data series, the position weightings of which are not zero and are different from a given threshold value and preparation of derived data series, formed from data in the correlated positions. A correlation device for the conversion method is also disclosed.

(57) **Zusammenfassung:** Es wird ein Verfahren zur Bearbeitung von Datenfolgen, die jeweils eine Anzahl von Daten in einer vorbestimmten Reihenfolge von Positionen umfassen, beschrieben, das die Schritte aufweist: Ermittlung von Korrelationswerten für alle Paare, Tripel oder n-Tupel von Positionen in einem Satz von Datenfolgen auf der Grundlage eines vorbestimmten Korrelationsmasses, Ermittlung von Positionsgewichtungen aus den Korrelationswerten für jede Position der Datenfolgen, Erfassung von Gruppen zueinander korrelierter Positionen in den Datenfolgen, deren Positionsgewichtungen ungleich Null sind und von einem vorbestimmten Schwellwert abweichen, und Bereitstellung von abgeleiteten Datenfolgen, die durch Daten an den korrelierten Positionen gebildet werden. Es wird auch eine Korrelatorvorrichtung zur Umsetzung des Verfahrens beschrieben.

WO 02/17127 A2

Verfahren und Vorrichtung zur Korrelationsanalyse von Datenfolgen

Die Erfindung betrifft Verfahren zur Bearbeitung von Datenfolgen, insbesondere zur Korrelationsanalyse von Datenfolgen, um Positionen von miteinander korrelierten Daten in verschiedenen Datenfolgen zu erfassen, wie z. B. Verfahren zur Kompression von Datenfolgen, zur Identifikation von bedeutungstragenden Positionen in Datenfolgen und/oder zur Klassifikation von Datenfolgen mittels Korrelationsanalysen, Vorrichtungen zur Durchführung der Verfahren und Anwendungen der Verfahren.

In allen Bereichen von Forschung und Technik fallen Daten an, die in Form von Symbolen mit technischem Bedeutungsinhalt (z. B. Alphabete aus Zahlen, Buchstaben, Benennungen von Substanzen oder Systemzuständen, oder dgl.) Informationen über einen technischen Aufbau, eine chemische Reaktion, ein biologisches System, einen physikalischen Zustand oder dgl. gegeben sind. Die Daten fallen in der Regel in einer bestimmten Reihenfolge an, die sich beispielsweise aus einer zeitlichen Reihenfolge, einer geometrischen Anordnung oder auch einem zahlenmäßigen Systemparameter ergibt. Datenfolgen können eindimensional (z. B. Zeitreihen von Messwerten, biologische Substanzfrequenzen) sein. Sie können aber auch mehrdimensional sein: dies ist offensichtlich bei Grauwertmatrizen in der Bildverarbeitung, aber auch beispielsweise bei DNA-Sequenzen gegeben. Letztere werden zu mehrdimensionalen Datenfolgen, wenn man zu jeder Nukleinsäure ihre Strukturparameter abspeichert. Die zur Verfügung stehenden Datenmengen wachsen durch sich erweiternde Mess- und Speichermöglichkeiten ständig. Beispielsweise liegen in der Gentechnik umfangreiche biologisch relevante Informati-

onen in Form von Datenfolgen, z. B. als DNA-Sequenzen, Proteinsequenzen, kodierte Umweltdaten, kodierte Phänotypen, Bandenmuster einer gelelektrophoretischen Analyse, Haplotypen, oder Kombinationen aus diesen, vor. Es besteht ein Interesse an Verfahren, um die anwendungsabhängig wichtigeren von den weniger wichtigen Daten zu trennen oder die Daten nach vorgegebenen Gesichtspunkten zu klassifizieren. Dies ist sowohl für eine effektive Handhabung der Daten in Datenverarbeitungsanlagen (Speicherbedarf, Rechenzeiten und dgl.) als auch für die Auswertung der Daten (Mustererkennung, Gewinnung neuer Systemparameter oder dgl.) von Bedeutung. Speziell in der Bioinformatik sollen in Datenfolgen biologisch bedeutungstragender Symbole, die relevanten Positionen und/oder Gruppen von Positionen und deren Assoziation zu äußeren Ausprägungen oder Umweltbedingungen des betrachteten biologischen Systems erkannt werden. Es besteht ein besonderes Interesse an der Charakterisierung des Verhaltens von komplexen Systemen, zu denen mehrere Datenfolgen, z. B. in Bezug auf innere Systemzustände und äußere Systembedingungen, vorliegen. Bisher sind keine effektiven Verfahren zur Verarbeitung von Datenfolgen komplexer Systeme, insbesondere zur Erfassung von Korrelationen zwischen bedeutungstragenden Positionen in den Datenfolgen, verfügbar.

Herkömmliche Verfahren zur Analyse und Klassifizierung von Datenfolgen basieren auf einer nur positionsweisen Untersuchung der Daten und einer darauf additiv aufbauenden Berechnung. Solche herkömmlichen Techniken sind beispielsweise in von M. J. Bishop et al. in "DNA and Protein Sequence Analysis" Oxford 1997, dargestellt. Sie sind jedoch nicht in der Lage, die Bedeutung von Positionen in den Datenfolgen zu erkennen, wenn diese sich erst aus dem Kontext einer oder mehrerer anderer, unter Umständen in der Datenfolge weit auseinander liegender Positionen ergibt, und führen deshalb durch die Vernachlässigung oder gar Unterschlagung solcher Positionen bei jeder auf der Unterscheidung wichtiger bzw. unwichtiger Positionen beru-

henden Datenkompression und Klassifikation zu fehlerhaften Ergebnissen.

Die Aufgabe der Erfindung ist es, verbesserte Verfahren zur Untersuchung von Datenfolgen anzugeben, die sich insbesondere dadurch auszeichnen, dass die Daten nicht nur mit hoher Effektivität, sondern derart verarbeitet und gegebenenfalls reduziert werden können, dass Fehler vermieden werden, die auf einer Nichtberücksichtigung von bestehenden Abhängigkeiten zwischen den Positionen in den Datenfolgen beruhen. Das verbesserte Verfahren soll insbesondere auch eine zuverlässige Klassifikation von Daten ermöglichen. Die Aufgabe der Erfindung ist es auch, Vorrichtungen zur Umsetzung der Verfahren und neue Anwendungen anzugeben.

Diese Aufgaben werden mit Verfahren, Computerprogrammprodukten und Vorrichtungen mit den Merkmalen gemäß den Patentansprüchen 1, 14 bzw. 15 gelöst. Vorteilhafte Ausführungsformen und Anwendungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Die Grundidee der Erfindung ist es, Zusammenhänge oder Wechselwirkungen (Interdependenzen) zwischen einzelnen Positionen verschiedener Datenfolgen durch eine Korrelationsanalyse mit den folgenden Schritten zu erfassen. Zunächst wird in der Gesamtheit aller Datenfolgen für alle Paare von Positionen mit einem vorgegebenem Korrelationsmaß jeweils ein Korrelationswert ermittelt. Die Datenfolgen können als Vektoren aufgefasst werden, deren Komponenten durch die Daten gebildet werden. Auf alle Komponentenpaare wird das Korrelationsmaß zur Ermittlung des jeweiligen Korrelationswertes angewendet. Um die ermittelten Korrelationswerte in Bezug auf ihre Signifikanz beurteilen zu können, werden zum Vergleich systembezogene Referenzwerte oder ggf. Simulationskorrelationswerte bzw. aus diesen gewonnene repräsentative Referenzwerte herangezogen. Die Ermittlung

von Simulationskorrelations- bzw. Referenzwerten erfolgt anwendungsabhängig ein- oder mehrmalig vor oder nach der Ermittlung der paarweisen Korrelationswerte. Durch Vergleich der Korrelationswerte insbesondere mit den zu den entsprechenden Positionspaaren gehörenden Referenzwerten kann im Rahmen eines einfachen Schwellwertverfahrens festgestellt werden, ob der jeweilige Korrelationswert oder ein davon abgeleiteter Positionsgewichtungswert so hoch ist, dass die zugehörigen Daten bzw. Positionen einer Gruppe von korrelierten Daten bzw. Positionen zugeordnet werden oder nicht. Die genannten Schritte können analog auch auf Tripel oder höhere n-Tupel von Positionen angewendet werden.

Je nach dem Ergebnis des Schwellwertverfahrens wird zu jeder Datenfolge (mindestens) eine abgeleitete Datenfolge erzeugt, die durch die korrelierten Positionen der Ausgangsdatenfolgen gebildet wird. Auf der Basis des Vergleichs der Korrelationswerte mit den Simulationskorrelationswerten oder den repräsentativen Referenzwerten können auch differenziertere Klassifikationen innerhalb der Gruppen der korrelierten bzw. nichtkorrelierten Daten vorgenommen werden.

Die Ermittlung und Bewertung paarweiser Korrelationswerte besitzt den Vorteil, dass die weitere Verarbeitung der abgeleiteten Datenfolgen sowie die oft zeit- und kostenaufwendige Erzeugung eventuell weiterer zum betrachteten Datensatz gehörender Datenfolgen je nach dem interessierenden Gesichtspunkt auf den relevanten Teil der Datenfolge beschränkt werden kann. Das erfindungsgemäße Verfahren ergibt eine Datenkompression, die Speicher- und Rechenzeiten sowie Arbeitszeit und -kosten spart. Des Weiteren ergibt sich als besonderer Vorteil, dass zwischen Datenfolgen, die zu einem System gehören, jedoch ganz verschiedene Datentypen enthalten, Assoziationen zwischen verschiedenen Positionen bestimmt werden können. Beispielsweise können die Datenfolgen jeweils DNA-Sequenzen, relevante Um-

weltdaten und auch die zugehörigen Phänotypen in geeignet kodierter Form enthalten. Die erfindungsgemäß ermittelten Assoziationen liefern Zusammenhänge zwischen Gruppen von DNA-Positionen, Umwelteinflüssen und Phänotypen und damit wiederum neue Informationen als Ausgangspunkt für eine Bewertung oder Veränderung des betrachteten biologischen Systems.

Die genannten Vorteile spielen nicht nur in der Auswertung biologisch relevanter Daten eine Rolle. Es ergeben sich allgemein eine Vereinfachung und Beschleunigung von Arbeiten wie z. B. der Laboranalyse biologischer Sequenzen, der automatisierten Bilderkennung oder der Überwachung technischer Anlagen, und der anwendungsrelevanten Interpretation der Datenfolgen. In komplexen technischen Anlagen können Korrelationen zwischen Systemzuständen zuverlässig erfasst und in Bezug auf die Steuerung von Prozessparametern oder die Abgabe von Warnsignalen verwendet werden. Bevorzugte Anwendungen der Erfindung ergeben sich somit neben der Informationsverarbeitung an technischen Anlagen vor allem in der Molekularbiologie, der Medizin, der Biologie, der Veterinärmedizin, der Agrarwirtschaft und der Ökobiologie.

Gegenstand der Erfindung ist auch ein Computerprogrammprodukt, das zur Kompression von Datenfolgen, Erfassung von Mustern in Datenfolgen und/oder Erfassung von Klassen in Datenfolgen nach dem erfindungsgemäßen Verfahren eingerichtet ist.

Gegenstand der Erfindung ist ferner eine Korrelatorvorrichtung zur Verarbeitung von Datenfolgen nach dem erfindungsgemäßen Verfahren. Eine Korrelatorvorrichtung umfasst insbesondere eine Speichereinrichtung zur Speicherung der zu bearbeitenden Datenfolgen, eine Recheneinrichtung zur Ermittlung von Korrelationswerten, Simulationskorrelationswerten und Referenzwerten, und eine Vergleichereinrichtung zur Bewertung der Korre-

lationswerte und zur Erfassung der Positionen von korrelierten bzw. nicht-korrelierten Daten.

Weitere Einzelheiten und Vorteile der Erfindung werden im Folgenden anhand einer Darstellung des erfindungsgemäßen Grundkonzepts der Korrelationsanalyse, einer Verfahrensdarstellung und eines Beispiels verdeutlicht. Die Erläuterung bezieht sich auf die Verarbeitung biologisch relevanter Informationen. Die Erfindung ist jedoch nicht auf diese Anwendung beschränkt, sondern auch in allen anderen technischen Gebieten zur Verarbeitung von Datenfolgen anwendbar.

Prinzipien der erfindungsgemäßen Korrelationsanalyse

Dem erfindungsgemäßen Verfahren liegen die folgenden Erkenntnisse der Erfinder zu Grunde. Die einzelnen Positionen der betrachteten Menge von Datenfolgen sind mehr oder weniger "verrauscht". Einige Positionen sind in (nahezu) allen Datenfolgen identisch besetzt, während andere Positionen hochvariabel sind. Zum Zwecke der Klassifikation oder Zuordnung unterschiedlicher Funktionsausprägungen zu den Datenfolgen sind die konstanten Positionen unbrauchbarer. Es sind vielmehr die variablen Positionen, an denen die zu klassifizierenden Datenfolgen nicht übereinstimmen, zu betrachten. Unter Funktionsausprägung wird hier und im folgenden allgemein ein Zusammenhang zwischen Datenfolgen und Systembedingungen verstanden, der in der Regel in der einen oder anderen Richtung kausal interpretiert wird. Eine Änderung der Systembedingungen kann eine Änderung der in der Datenfolge festgehaltenen Messwerte verursachen. Andererseits kann eine Änderung z. B. in einer Gensequenz zu einer Änderung des Phänotypen führen. Dabei kann die Funktionsausprägung in geeignet kodierter Form selbst Bestandteil der Datenfolge sein.

Es sind zwei prinzipiell verschiedene Qualitäten der Variabilität einer Position in einer Datenfolge unterscheidbar. Einerseits kann eine Position hochvariabel sein, weil eine Änderung der Besetzung keine Auswirkung auf die Ausprägung der Funktion hat. Andererseits kann eine hohe Variabilität gegeben sein, weil die jeweilige Position mit unterschiedlichen Funktionsausprägungen assoziiert ist. Da die Funktionsausprägung einer Datenfolge durch spezifische Besetzung einer Kombination mehrerer, im allgemeinen nicht benachbarter Positionen bestimmt wird, ist davon auszugehen, dass die in Zusammenhang mit der betrachteten Funktion bedeutungstragenden Positionen voneinander abhängig besetzt sind und korreliert veränderlich sind ("synchron rauschen"), während die zufällig rauschenden Positionen eher unabhängig von jeder anderen Position besetzt sind.

Die Erfinder haben ferner festgestellt, dass das synchrone Rauschen der bedeutungstragenden Positionen nicht nur auf Datenpaare beschränkt ist, sondern auch größere Gruppen von Daten an bestimmten Positionen betreffen. Das erfindungsgemäße Verfahren ist nun darauf gerichtet, die im Zusammenhang mit einer betrachteten Funktion stehende Bedeutung der einzelnen Positionen in einer Menge von Datenfolgen zu quantifizieren und auf dieser Grundlage die Datenfolgen Kompressions-, Klassifizierungs- und/oder Vorhersageprozeduren zu unterziehen. Datenkompression bedeutet, dass in der weiteren Verarbeitung der Datenfolgen nur die relevanten Positionen oder Positionsgruppen in Betracht gezogen werden.

Die durch die erfindungsgemäße Korrelationsanalyse gewonnene Information kann auch unmittelbar zur Klassifikation benutzt werden. Die Datenfolgen, die an den Positionen einer Gruppe stark voneinander abhängiger, verrauschter Positionen (zumindest nahezu) dieselben Besetzungen besitzen, werden zu einer Teilklasse zusammengefasst. Von den vielen theoretisch mögli-

chen Besetzungen an diesen Positionen kommen wegen der gegenseitigen Abhängigkeiten nur wenige, die jeweilige Teilklasse charakterisierenden Muster vor.

Besitzt nun die so konstruierte Klassifikation die Eigenschaft, dass jeweils in einer Teilklasse zusammengefasste Positionsfolgen sich in ihrer Funktionsausprägung nicht oder nur unwesentlich unterscheiden, so hat man eine Korrelation mit der betrachteten Funktion gefunden, die im Hinblick auf die betrachtete Funktion auch Vorhersagen zukünftiger Systemzustände möglich macht. Sind zusätzlich zu den ursprünglich betrachteten Datenfolgen weitere Datenfolgen gegeben und besitzen diese an den ausgezeichneten Positionskombinationen bekannte, d. h. im Rahmen der Klassifikation ermittelte Besetzungen, so können diese Positionskombinationen mit der entsprechenden Funktionsausprägung in Beziehung gebracht werden. Anwendungsabhängig kann vorgesehen sein, dass derartige Vorhersagen durch zusätzliche Verfahren oder Informationen validiert werden.

Die technische Anwendung der erfindungsgemäßen Korrelationsanalyse ergibt sich aus der Datenkompression, bei der in Bezug auf eine bestimmte Funktion die wichtigen Datenpositionen erkannt und weiter verarbeitet werden, der Mustererkennung bzw. Klassifikation, bei der Kombinationen von Positionsbesetzungen an den erkannten wichtigen Positionen ermittelt werden, die relevante Teilklassen der betrachteten Datenfolgen beschreiben, der Assoziation von Mustern in den Positionsfolgen zu Ausprägungen der betrachteten Funktionen und der Vorhersage von Funktionsausprägungen in neuen Datenfolgen.

Durchführung der erfindungsgemäßen Korrelationsanalyse

1. Schritt: Bereitstellung der Daten

In einem ersten Schritt werden die interessierenden Daten für die erfindungsgemäße Korrelationsanalyse bereitgestellt, z. B. auf eine Korrelatorvorrichtung übertragen. Anwendungsabhängig werden zunächst die Daten gemessen oder erfasst, über eine Schnittstelle in die Korrelatorvorrichtung eingegeben, zwischengespeichert und zu Datenfolgen zusammengestellt. Dieser Teilschritt ist nicht zwingend notwendig, die Datenfolgen können bereits bspw. als Messwertfolgen vorliegen. Anschließend werden die Datenfolgen zur Bildung einer Menge von Folgen, die einander entsprechende Daten an jeweils derselben Position besitzen und die alle die gleiche Länge besitzen, formatiert. Falls die zunächst bereitgestellten Daten zu Datenfolgen mit verschiedenen Längen führen, wie dies beispielsweise bei Datenfolgen zur Beschreibung eines Phänotyps der Fall sein kann, entstehen in der entsprechenden Datenfolge Lücken. Zur Formatierung werden die Lücken aufgefüllt oder die entsprechenden Positionen in den übrigen Datenfolgen (z.B. Gensequenzen) gestrichen. Das Auffüllen erfolgt beispielsweise mit einem gesonderten "Lücke"- oder "gap"-Symbol, mit dem an dieser Position häufigsten Wert oder - bei numerischen Daten - mit einem Durchschnittswert.

Die Datenfolgen basieren gegebenenfalls auf jeweils verschiedenen Symbolvorräten oder "Alphabeten" und liegen beispielsweise in gespeicherter Form vor.

2. Schritt: Ermittlung von Korrelationswerten und Positionsgewichtungen

Je nach der Aufgabenstellung wird eine problemrelevante Methode zur Berechnung der Abhängigkeiten zwischen je zwei Positio-

nen verschiedener Datenfolgen verwendet. Die paarweisen gegenseitigen Abhängigkeiten (Korrelationswerte) werden in einem ersten Teilschritt durch ein Korrelationsmaß entsprechend der gewählten Methode ermittelt. Im Folgenden werden beispielhaft zwei Korrelationsmaße, nämlich die Transinformation und die Vorhersagbarkeit, illustriert. Die Erfindung ist jedoch nicht auf diese Maße beschränkt, sondern mit allen Methoden umsetzbar, die allgemein geeignet sind, Assoziationen oder Korrelationen zwischen Positionen durch Angabe von quantitativen Korrelationswerten zu charakterisieren. Verschiedene solche Methoden sind an sich bekannt und basieren beispielsweise auf χ^2-Tests oder lehrbuchbekannten Algorithmen.

(a) Transinformation

Die Transinformation ist ein auf der Shannon'schen Entropie basierendes Korrelationsmaß, das aus der Informationstheorie zur Charakterisierung der Kombination zweier Signale an sich bekannt ist (siehe z. B. H. Rohling "Einführung in die Informations- und Codierungstheorie", Stuttgart, 1995). Der Korrelationswert Transinformation wird wie folgt gebildet. Sind A_i das Alphabet für die Position i und A_j das Alphabet für die Position j, p_i bzw. p_j die zugehörigen Häufigkeitsverteilungen und p_{ij} die gemeinsame Häufigkeitsverteilung der beiden Positionen, so ist die Transinformation T (ij) der Positionen i und j gemäß der folgenden Gleichung gegeben.

$$T(i,j) = \sum_{a \in A_i} p_i(a) \log \frac{1}{p_i(a)} + \sum_{b \in A_j} p_j(b) \log \frac{1}{p_j(b)} - \sum_{a \in A_i, b \in A_j} p_{i,j}(a,b) \log \frac{1}{p_{i,j}(a,b)},$$

Die Transinformation T ergibt sich als Summe der Entropien für die einzelnen Positionen, vermindert um die Entropie des Positionenpaares. Die Transinformation ist in der Informationstheorie ein gebräuchliches Maß für die Beschreibung der gegenseitigen Beeinflussung zweier Signale. Sie ist minimal, wenn be-

trachtete Positionen statistisch unabhängig sind, und maximal, wenn beide Positionen gleichverteilt und sich gegenseitig in eindeutiger Weise bestimmend sind.

Das Korrelationsmaß Transinformation liefert für jedes Positionenpaar eine Zahl, die die Korrelation beschreibt. Aus dem quantitativen Wert allein ist die Korrelation ohne Zusatzinformationen nicht bewertbar, da die Größe von T auch von der Zahl der Symbole in den Datenfolgen abhängt. Je mehr Symbole die Alphabete umfassen, desto größere T-Werte treten auf. Die Bewertung erfolgt im dritten Schritt (siehe unten).

(b) Vorhersagbarkeit

Die Vorhersagbarkeit ist ein neu entwickeltes, gerichtetes Maß für Korrelationen zwischen verschiedenen Positionen, das davon abhängt, ob bei zwei betrachteten Positionen die eine aus der anderen ableitbar oder vorhersagbar ist. Der Korrelationswert Vorhersagbarkeit ist ein quantitatives Maß für die Aussage "falls an Position i ein a, dann an Position j ein b". Das Maß Vorhersagbarkeit ergibt sich aus den folgenden Überlegungen. Für jedes $a \in A_i$ sei $f_{ij}(a) \in A_j$ der am häufigsten mit einem a an Position i einhergehende "Buchstabe" an Position j. Falls es mehrere häufigste Buchstaben gibt, so wird einer von ihnen beliebig ausgewählt, da das Ergebnis der Ermittlung der Vorhersagbarkeit nicht von dieser Auswahl unter den häufigsten Buchstaben abhängt. Ist N die Anzahl aller Datenfolgen und $n_{ij}(a)$ die Anzahl derjenigen Datenfolgen unter ihnen, die an Position i ein a und an Position j ein $f_{ij}(a)$ besitzen, so ist die Vorhersagbarkeit V(i,j) der Position j durch Position i durch die folgende Gleichung gegeben.

$$V(i,j) = \sum_{a \in A_i} \frac{n_{i,j}(a)}{N} \cdot H(j)$$

Dabei ist H(j) die Entropie $H(j) = -\sum_{b \in A_j} p_j(b) \log p_j(b)$. Die Vorhersagbarkeit ist die mit der Entropie der vorherzusagenden Position gewichtete Anteil derjenigen Datenfolgen, bei denen die Vorhersage der Position j richtig ist, falls man aus der Kenntnis der Besetzung von Position i auf die jeweils am häufigsten damit einhergehende Besetzung von Position j schließt.

Schließlich werden in einem weiteren Teilschritt aus den paarweise für alle Positionen der Datenfolgen ermittelten Korrelationswerten Positionsgewichtungen bestimmt. Für jede Position der Datenfolgen werden alle zugehörigen Korrelationswerte einer Summation (gleichbedeutend einer Mittelwertbildung) oder einer Maximumsbildung unterzogen, so dass sich jeweils als quantitativer Parameter die Positionsgewichtung ergibt, die zusätzlich zu den Korrelationswerten als eine Form der Informationsverdichtung ausgegeben bzw. gespeichert wird. Hierdurch werden diejenigen Positionen stark gewichtet, die - im Falle der Summation - im Mittel zu allen anderen Positionen eine starke Abhängigkeit besitzen bzw. - im Falle der Maximumsbildung - zu mindestens einer anderen Position.

Bereits nach diesem Schritt kann anwendungsabhängig eine erste Reduzierung der Datenfolge durch Streichung aller Positionen erfolgen, deren Wert der Positionsgewichtung Null beträgt oder so niedrig ist, dass eine Korrelation mit anderen Positionen ausscheidet. Hierzu erfolgt beispielsweise ein Vergleich mit vorbestimmten systembezogenen Referenzwerten.

3. Schritt: Ermittlung von Referenzwerten für die statistische Bewertung der Positionsgewichtungen

Die mit dem Korrelationsmaß gelieferten quantitativen Werte zur Charakterisierung der gegenseitigen Abhängigkeit zwischen Positionen können in Bezug auf ihre statistische Signifikanz

durch ein Simulationsverfahren bewertet werden. Die Durchführung des Simulationsverfahrens ist kein zwingendes Merkmal der Erfindung. Anwendungsabhängig kann darauf verzichtet werden, falls beispielsweise Zusatzinformationen über das betrachtete System vorliegen oder wenn die ermittelten Korrelationen ohne weiteres dahingehend beurteilt werden können, ob sie im System technisch oder biologisch sinnvoll sind.

Das Simulationsverfahren umfasst die Erzeugung einer großen Anzahl von randomisierten Referenzdatensätzen (sogenannte "Shuffles"). Die Referenzdatensätze bestehen jeweils aus derselben Anzahl an Datenfolgen wie der betrachtete Datensatz, besitzen alle dieselbe Länge wie die gegebenen Datenfolgen und gehen auf folgende Weise aus diesen hervor: Stellt man sich die einzelnen Datenfolgen des gegebenen Datensatzes zeilenweise untereinander geschrieben vor, so werden die Daten innerhalb der Spalten, also die jeweils an derselben Position stehenden Daten untereinander zufällig vertauscht. Derartige positionsinterne Vertauschungen verändern das Rauschen der Positionen nicht, brechen jedoch gegebene Abhängigkeiten auf und schaffen möglicherweise neue Abhängigkeiten. Für jeden Referenzdatensatz wird wie bei Schritt 2 das Korrelationsmaß zur quantitativen Bewertung gegenseitiger Abhängigkeiten angewendet. Es ergeben sich eine Vielzahl von Simulationskorrelationswerten für alle Paare von Positionen jedes betrachteten „Shuffles".

Es wird für jeden Referenzdatensatz des Simulationsverfahrens die jeweils maximale auftretende Abhängigkeit zwischen zwei Positionen bestimmt. Ferner wird für jeden Referenzdatensatz die maximale Positionsgewichtung entsprechend dem für den gegebenen Datensatz gewählten Verfahren bestimmt. Jeweils Mittelwert und Varianz dieser beiden Werte, über alle Referenzdatensätze ermittelt, werden als repräsentative Referenzwerte für den späteren Vergleich mit den für die betrachteten Daten

folgen berechneten Korrelationswerten und Positionsgewichtungen ausgegeben oder gespeichert.

4. Schritt: Erfassung der Positionen von miteinander korrelierten Daten

In einem ersten Teilschritt werden Abhängigkeitsgruppen von Positionen ermittelt. Hierzu werden die paarweisen Abhängigkeiten der Positionen mit einem vorbestimmten Schwellwert verglichen. Der Schwellwert ist beispielsweise (wie bei Entscheidungen über statistische Signifikanz üblich) die Summe aus Mittelwert und Varianz der in Schritt 3 bestimmten maximalen Abhängigkeit in den Referenzdatensätzen. Alternativ kann als Schwellwert eine anwendungsabhängig eingestellte Größe verwendet werden, die auf Zusatzinformationen, Erfahrungswerten oder dgl. basiert. Die Bestimmung von korrelierten Positionen erfolgt vorzugsweise durch Bildung von Abhängigkeitsgruppen der Positionen nach dem folgenden Schema.

Gruppen von Positionen, deren paarweise Abhängigkeiten voneinander sämtlich über dem Schwellwert liegen, werden als sogenannte Cliquen zusammengefasst. Falls die Mehrzahl der Korrelationswerte über dem Schwellwert liegen, eine kleine Anzahl von Positionspaaren jedoch geringere Korrelationswerte ergeben, so werden die zugehörigen Positionen in Gruppen zusammengefasst, die als "Beinahe-Cliquen" bezeichnet werden. Bei der Definition einer "Beinahe-Clique" kann ein zweiter, niedrigerer Schwellwert als Mindestgröße für diejenigen Korrelationswerte berücksichtigt werden, die den Schwellwert für eine Clique nicht erreichen. Als schwächste Form einer Abhängigkeitsgruppe werden Positionen, die lediglich mittelbar voneinander stark abhängig sind, als "Komponenten" zusammengefasst. Dabei ist eine mittelbare Abhängigkeit der Positionen i und q dann gegeben, wenn es Positionen j, k, ... q derart gibt, dass die Positionenpaare (i, j), (j, k), ... , (p, q) jeweils über dem

Schwellwert liegende Korrelationswerte besitzen. Ein hoher Korrelationswert für das Positionenpaar (i, q) muss jedoch nicht notwendigerweise vorliegen.

Zum Zwecke der Verkürzung der Datenfolgen und damit der Datenkompression können alle außerhalb der Abhängigkeitsgruppen liegenden Positionen gestrichen (gelöscht) werden. Es bleiben dann nur die relevanten für die weitere Verarbeitung gewünschten Daten bestehen.

In einem weiteren Teilschritt werden die Abhängigkeitsgruppen ausgegeben bzw. gespeichert. Den Positionen der Datenfolgen wird eine Information zugeordnet, wonach sie zu einer der genannten Abhängigkeitsgruppen gehören oder nicht. Es werden abgeleitete Datenfolgen gebildet, die ausschließlich die korrelierten Positionen umfassen. Die abgeleiteten Datenfolgen werden anwendungsabhängig an eine Schnittstelle zu einem weiteren Auswertungs- oder Diagnosegerät gegeben, gespeichert, angezeigt oder anderweitig dargestellt.

5. Schritt: Bestimmung von Teilklassen der Datenfolgen

Auf der Grundlage der bei Schritt 4 ermittelten Abhängigkeitsgruppen werden anschließend Teilklassen der gegebenen Menge von Datenfolgen ermittelt. Die Abhängigkeitsgruppen bilden bestimmte Muster, d. h. Kombinationen von Positionsbesetzungen. Die Teilklassen und die sie charakterisierenden Muster innerhalb der Datenfolgen werden ausgegeben bzw. gespeichert.

Im Ergebnis sind die für die weitere Bearbeitung, Anzeige oder Auswertung relevanten Datenfolgen in ihrer Anzahl durch Auswahl jeweils einer repräsentativen Datenfolge je Teilklasse reduziert worden.

6. Schritt: Vorhersage

Die Vorhersage umfasst die Bearbeitung einer oder mehrerer neuer Datenfolgen entsprechend den Schritten 1 bis 5 und den Vergleich der bei Schritt 5 für die neuen Datenfolgen ermittelten Muster mit den Mustern der vorher verarbeiteten Datenfolgen. Wenn Übereinstimmungen charakteristischer Muster gegeben sind, so wird den jeweiligen Positionen der neuen Datenfolgen die entsprechend für die zuerst verarbeiteten Datenfolgen ermittelte Teilklasse zugeordnet bzw. die entsprechende Zugehörigkeit zu dieser Teilklasse vorhergesagt.

Beispiel

1. Schritt:

Das erfindungsgemäße Verfahren wird an einem konstruierten Beispiel erläutert Es werden 16 Positionsfolgen der Länge 9 betrachtet, die in Position 8 über dem Alphabet "1,2,3...", in Position 9 über dem Alphabet "+,-", sonst über dem Alphabet "A,C,G,T" gebildet sind. Es handelt sich bspw. um DNA-Sequenzen der Länge 7 mit einem in der angehängten Position 8 codierten Umwelteinfluss und einem in Position 9 vermerkten Vorhandensein einer phänotypischen Eigenschaft.

Position		1	2	3	4	5	6	7	8	9
Folge 1	:	G	A	A	A	A	A	A	3	+
Folge 2	:	T	C	A	T	C	C	A	3	+
Folge 3	:	A	T	A	C	T	C	G	2	-
Folge 4	:	A	A	A	C	A	A	G	2	+
Folge 5	:	C	C	A	G	C	C	T	1	-
Folge 6	:	G	G	A	A	G	G	A	3	+
Folge 7	:	C	G	A	G	G	A	T	1	-
Folge 8	:	G	T	A	A	T	C	A	3	+
Folge 9	:	G	T	A	A	T	C	A	3	+
Folge 10	:	A	G	A	C	G	G	G	2	+
Folge 11	:	C	T	A	G	T	T	T	1	-
Folge 12	:	T	T	A	T	T	T	A	3	-
Folge 13	:	C	A	A	G	A	G	T	1	-
Folge 14	:	G	C	A	A	C	T	A	3	+
Folge 15	:	T	G	A	T	G	A	A	3	-
Folge 16	:	A	C	A	C	C	T	G	2	+

2. Schritt:

Die paarweisen Abhängigkeiten zwischen den Positionen werden als Korrelationswert Transinformation berechnet:

Pos.i	Pos.j	T(i,j)	Pos.i	Pos.j	T(i,j)
1	2	0,0551	3	7	0,0000
1	3	0,0000	3	8	0,0000
1	4	1,3705	3	9	0,0000
1	5	0,0551	4	5	0,0551
1	6	0,0551	4	6	0,0551
1	7	1,0397	4	7	1,0397
1	8	1,0397	4	8	1,0397
1	9	0,4254	4	9	0,4254
2	3	0,0000	5	6	0,6943
2	4	0,0551	5	7	0,0169
2	5	1,3705	5	8	0,0169
2	6	0,6943	5	9	0,0418
2	7	0,0169	6	7	0,0169
2	8	0,0169	6	8	0,0169
2	9	0,0418	6	9	0,0091
3	4	0,0000	7	8	1,0397
3	5	0,0000	7	9	0,2636
3	6	0,0000	8	9	0,2636

Ein Wert der Transinformation von 0 bedeutet stochastische Unabhängigkeit im üblichen Sinne. Diese liegt insbesondere vor, wenn eine der betrachteten Positionen konstant ist, wie hier die Daten in Position 3. Die stärksten Abhängigkeiten in dem Beispiel bestehen zwischen den Positionen 1 und 4 bzw. zwischen den Positionen 2 und 5: Während die Positionen 2 und 5 identisch besetzt sind, also offensichtlich im höchsten Maße voneinander abhängig sind, so bestimmen sich auch die Positionen 1 und 4 gegenseitig eindeutig - ein "G" an Position 1 ist stets mit einem "A" an Position 4 verbunden, ein "T" mit einem "T", ein "A" mit "C" und ein "C" mit einem "G".

Anschließend folgt die durch Summenbildung bestimmte Positionsgewichtung:

Pos.	Gewicht
1	4,0406
2	2,2506
3	0,0000
4	4,0406
5	2,2506
6	1,5417
7	3,4334
8	3,4334
9	1,4707

Die Positionen 1 und 4 sind im Sinne dieser Gewichtung von größter Bedeutung, da alle anderen Positionen von ihnen durchschnittlich am stärksten abhängig sind.

3. Schritt:
Die Überprüfung der statistischen Relevanz mittels Simulation ergibt: 100 „Shuffles" besitzen durchschnittlich eine maximale Abhängigkeit zweier Positionen voneinander von 0,5941 bei einer Varianz von 0,0870; für die Positionspaare mit einer stärkeren Abhängigkeit als 0,5941 + 0,0870 = 0,6811 ist die statistische Relevanz gegeben.

4. Schritt:
Wählt man als Schwellwert 0,5941 + 2 · 0,0870 = 0,7681, betrachtet also nur diejenigen Positionspaare mit einer Transinformation, die um mindestens zwei Varianzen größer als der zu erwartenden maximalen ist, so findet man zwei Cliquen: die Gruppe der Positionen 1,4,7,8 (je zwei dieser vier Positionen besitzen eine über der gewählten Schwelle liegende Transinformation) und die Gruppe der Positionen 2,5.

5. An den Positionen 1,4,7,8 kommen folgende Muster innerhalb der Menge von Positionsfolgen vor:

Position		1	4	7	8
Folge 1	:	G	A	A	3
Folge 2	:	T	T	A	3
Folge 3	:	A	C	G	2
Folge 4	:	A	C	G	2
Folge 5	:	C	G	T	1
Folge 6	:	G	A	A	3
Folge 7	:	C	G	T	1
Folge 8	:	G	A	A	3
Folge 9	:	G	A	A	3
Folge 10	:	A	C	G	2
Folge 11	:	C	G	T	1
Folge 12	:	T	T	A	3
Folge 13	:	C	G	T	1
Folge 14	:	G	A	A	3
Folge 15	:	T	T	A	3
Folge 16	:	A	C	G	2

Dies führt zur Einteilung der Menge in vier Teilklassen:

Teilklasse 1 (zum Muster "GAA3"):	Folgen 1,6,8,9,14
Teilklasse 2 (zum Muster "TTA3"):	Folgen 2,12,15
Teilklasse 3 (zum Muster "ACG2"):	Folgen 3,4,10,16
Teilklasse 4 (zum Muster "CGT1"):	Folgen 5,7,11,13

Hier ist zu bemerken, dass die Klassifizierung nach den an den Positionen 2,5 vorkommenden Mustern zu einer anderen Einteilung geführt hätte:

Teilklasse 1 (zum Muster "AA"):	Folgen 1,4,13
Teilklasse 1 (zum Muster "CC"):	Folgen 2,5,14,16
Teilklasse 1 (zum Muster "TT"):	Folgen 3,8,9,11,12
Teilklasse 1 (zum Muster "GG"):	Folgen 6,7,10,15

Wahlweise können zu allen in Schritt 4 gefundenen Positionsgruppen die jeweils implizierte Klassifizierung ausgegeben werden, um dann unter Ausnutzung zusätzlicher Informationen zu entscheiden, welche auf das Problem bezogen am geeignetsten ist. Es ist auch möglich, eine gemeinsame Partitionierung zu konstruieren – je nach Zielsetzung etwa die gröbste Partitionierung, die feiner als alle gefundenen ist, oder die feinste unter den gröberen.

6. Schritt:
Schließlich wird für die nicht zu den ursprünglichen Positionsfolgen gehörende Folge "GGAATTC3" ein "+" für die in Position 9 codierte Funktion, also das Vorhandensein der betrachteten phänotypischen Eigenschaft, vorhergesagt, da ihr Muster "GAA3" an den Positionen 1,4,7,8 mit dem die Teilklasse 1 charakterisierenden Muster übereinstimmt und jede Positionsfolge aus dieser Teilklasse ein "+" an Position 9 besitzt.

Vorrichtung zur Korrelationsanalyse

Eine erfindungsgemäße Korrelatorvorrichtung umfasst eine Formatierungseinrichtung zur Bereitstellung einer Vielzahl von Datenfolgen gleicher Länge, eine Recheneinrichtung zur Bestimmung der Korrelationswerte zwischen allen Positionspaaren der Datenfolgen und der daraus abgeleiteten Positionsgewichtungen, eine Vergleichereinrichtung zum Vergleich der Positionsgewichtungen mit vorbestimmten Referenzwerten und zur Ermittlung von korrelierten Positionen, und eine Einrichtung zur Anzeige, Ausgabe oder Speicherung von abgeleiteten Datenfolgen, die durch die korrelierten Positionen gebildet werden. Die verschiedenen Komponenten der Korrelatorvorrichtung werden vorzugsweise durch eine Datenverarbeitungsanlage, z. B. einen Computer, implementiert.

Patentansprüche

1. Verfahren zur Bearbeitung von Datenfolgen, die jeweils eine Anzahl von Daten in einer vorbestimmten Reihenfolge von Positionen umfassen, mit den Schritten:
- Ermittlung von Korrelationswerten für alle Paare, Tripel oder n-Tupel von Positionen in einem Satz von Datenfolgen auf der Grundlage eines vorbestimmten Korrelationsmaßes,
- Ermittlung von Positionsgewichtungen aus den Korrelationswerten für jede Position der Datenfolgen,
- Erfassung von Gruppen zueinander korrelierter Positionen in den Datenfolgen, deren Positionsgewichtungen ungleich Null sind und von einem vorbestimmten Schwellwert abweichen, und
- Bereitstellung von abgeleiteten Datenfolgen, die durch Daten an den korrelierten Positionen gebildet werden.

2. Verfahren gemäß Anspruch 1, bei dem die Erfassung der korrelierten Positionen die folgenden Schritte umfasst:
- Ermittlung von Simulationskorrelationswerten für alle Paare von Positionen in einer Vielzahl randomisierter Referenzdatensätze,
- Ermittlung von repräsentativen Referenzwerten aus den simulierten Referenzdatensätzen für die Korrelationswerte und Positionsgewichtungen,
- Ermittlung von Schwellwerten aus den Referenzwerten,
- Zuordnung der Positionen, für die die Positionsgewichtung der zu bearbeitenden Datenfolgen größer oder kleiner als der entsprechende Schwellwert ist, zu einer Gruppe der korrelierten Positionen oder einer Gruppe von nicht-korrelierten Positionen.

3. Verfahren gemäß Anspruch 1 oder 2, bei dem als Korrelationswerte Transinformations- oder Vorhersage-Werte ermittelt werden.

4. Verfahren gemäß Anspruch 2 oder 3, bei dem die repräsentativen Referenzwerte durch Berechnung statistischer Momente (Erwartungswert, Varianz, höhere Momente) und Kombinationen aus ihnen oder anderer mathematischer Funktionen aus allen oder den jeweils maximalen Korrelationswerten und Positionsgewichtungen über alle Referenzdatensätze ermittelt werden.

5. Verfahren gemäß Anspruch 4, bei dem die Gruppe der korrelierten Positionen in Untergruppen unterteilt wird, bei denen alle paarweisen Korrelationen oder die Mehrzahl aller paarweisen Korrelationen oder alle paarweisen mittelbaren Korrelationen der Positionen die Schwellwerte überschreiten.

6. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem die Datenfolgen und/oder die gemäß einem der vorhergehenden Ansprüche von ihnen abgeleiteten Datenfolgen einer Klassifizierung und/oder einer Mustererkennung unterzogen wird.

7. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem die Bereitstellung der Datenfolgen und/oder ihrer Ableitungen ein Speichern, Anzeigen oder Senden an eine Schnittstelle einer Datenverarbeitungseinrichtung umfasst.

8. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem eine Formatierung der zu bearbeitenden Datenfolgen derart vorgesehen ist, dass in jeder Datenfolge die gleiche Anzahl von Positionen gegeben ist.

9. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem die Datenfolgen über demselben Alphabet oder verschiedenen Alphabeten gebildet sind.

10. Verfahren gemäß Anspruch 9, bei dem die Alphabete der einzelnen Positionen der Datenfolgen biologische Substanzen, Eigenschaften biologischer Substanzen, Nukleinsäuren, Aminosäuren, Strukturparameter, Ausprägungen phänotypischer Merkmale und/oder Ausprägungen von Umweltmerkmalen kodieren.

11. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem die Datenfolgen Gensequenzen, Nukleinsäuresequenzen, Aminosäuresquenzen, Bandenmuster gelelektrophoretischer Analysen, Haplotypen, kodierte Phänotypen, kodierte Umweltdaten oder Kombinationen aus diesen umfassen.

12. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem die Alphabete Gruppen von Systemparametern eines Regelsystems, Messwerten und/oder Bildwerten umfassen.

13. Verfahren gemäß einem der vorhergehenden Ansprüche, bei dem die abgeleiteten Datenfolgen als Eingangsgröße für ein Vorhersage- oder Diagnoseverfahren bereitgestellt werden.

14. Computerprogrammprodukt, das zur Kompression von Datenfolgen, Erfassung von Mustern in Datenfolgen und/oder Erfassung von Klassen in Datenfolgen nach einem Verfahren gemäß einem der vorhergehenden Ansprüche eingerichtet ist.

15. Korrelatorvorrichtung, die umfasst:
- eine Formatierungseinrichtung zur Bereitstellung einer Vielzahl von Datenfolgen gleicher Länge,
- eine Speichereinrichtung zur Zwischenspeicherung der zu bearbeitenden Datenfolgen,
- eine Recheneinrichtung zur Bestimmung der Korrelationswerte zwischen allen Positionspaaren der Datenfolgen und der daraus abgeleiteten Positionsgewichtungen,

- eine Vergleichereinrichtung zum Vergleich der Positionsgewichtungen mit vorbestimmten Schwellwerten und zur Ermittlung von korrelierten Positionen, und
- eine Ausgabe- und/oder Speichereinrichtung zur Ausgabe oder Speicherung von abgeleiteten Datenfolgen, die durch die korrelierten Positionen gebildet werden.

16. Korrelatorvorrichtung gemäß Anspruch 15, die zur Durchführung der Schritte eines Verfahrens gemäß einem der Ansprüche 1 bis 14 eingerichtet ist.

17. Korrelatorvorrichtung gemäß Anspruch 15 oder 16, die durch eine Datenverarbeitungsanlage gebildet wird.

18. Verwendung eines Verfahrens, eines Computerprogrammprodukts oder einer Vorrichtung gemäß einem der vorhergehenden Ansprüche zur Erfassung von Korrelationen zwischen Positionen in Datenfolgen.