Seminar
Informatik für Biologen

Thema: REPuter

Von

Anja Braun und Harald Surowy

Inhaltsverzeichnis

1. Repetitive Sequenzen in Genomen 3
2. Die Suche nach repetitiven Sequenzen 3
3. REPuter 4
3.1. REPfind 5
3.1.1. Mathematische Grundlagen 5
3.1.2. Auffinden von exakten Repeats 6
3.1.3. Berechnung von degenerierten Repeats und Signifikanz 7
3.2. REPvis 7
4. Anwendungsgebiete 9
4.1. Assembly checking 9
4.2. Low copy repeats 10
4.3. Einzigartige Sequenzen 11
4.4. Vergleich von Genstrukturen 12
4.5. cDNA/EST mapping 13
5. Quellenangaben 14

1. Repetitive Sequenzen in Genomen

Das Genom von Säugetieren besteht aus wenig-, mittel- und hochrepetitiven Sequenzen, sowie aus einzigartigen Sequenzen. Einzigartige Sequenzen, zu denen die proteincodierenden Gene zählen, machen dabei den geringsten Teil aus. Den größten Anteil stellen die mittel- und hochrepetitiven Sequenzen dar mit über 70% des Gesamtgenoms. Zu den hochrepetitiven Sequenzen gehören Satelliten-DNA, die zumeist aus Tandemrepeats besteht, sowie LINEs und SINEs. Mittelrepetitive Sequenzen sind beispielsweise rRNA- und tRNA-Gene sowie Retrotransposons.

Repetitive Sequenzen sind im gesamten Genom verteilt und kommen in unterschiedlichen Formen vor. Sie können direkt sein oder auch palindromisch. Palindromische Repeats sind interessant, da sie Haarnadelstrukturen ausbilden können und so eine wichtige Funktion bei genetischen Vorgängen einnehmen. Solche Strukturen spielen beispielsweise eine wichtige Rolle bei der Termination der Transkription, aber auch bei anderen Prozessen. Andere interessante Repeatstrukturen sind auch Mikrosatelliten, die als Tandemrepeats vorkommen und die aufgrund ihres hohen Polymorphismus als Marker für genetische Untersuchungen eingesetzt werden. Außerdem spielen Repeats häufig eine wichtige Rolle bei Veränderungen in Genstrukturen, die zu Krankheiten führen. Als Transposons können sie beispielsweise innerhalb von codierenden Bereichen vorkommen.

Die Analyse solcher Repeatstrukturen kann demnach sehr interessante Einblicke bieten. Eine Betrachtung von abnormalen Verteilungen repetitiver Elemente und die Auswirkungen auf den Organismus bzw. der Vergleich von repetitiven Elementen unterschiedlicher Organsimen könnte außerdem helfen, die funktionelle und evolutionäre Rolle von Repeats aufzuklären.

Mittlerweile sind solche Untersuchungen mit Hilfe entsprechender Computerprogramme relativ leicht und schnell durchführbar.

2. Die Suche nach repetitiven Sequenzen

Ein Computerprogramm, das für die Analyse von repetitiven Sequenzen eingesetzt werden soll, muss gewisse Aufgaben erfüllen. Zunächst muss gesichert sein, dass alle Repeats eines Genoms, die in einem bestimmten Signifikanzbereich liegen, gefunden werden. Dies ist die Hauptaufgabe traditioneller Repeatanalysen, die sämtliche Programme erfüllen müssen. Darüber hinaus gibt es jedoch noch andere Aufgabenbereiche, die eine wichtige Rolle spielen für die Untersuchung von Genomen und Genstrukturen. Dazu zählt beispielsweise das

Auffinden von einzigartigen Sequenzen, die wichtig sind für die Entwicklung von Primern, oder die Möglichkeit, verschiedene Genome miteinander vergleichen zu können.
Außerdem muss solch ein Programm bestimmte Kriterien erfüllen, damit es sinnvoll genutzt werden kann.

Diese Kriterien sind:

1. Effizienz:

Um mit der Größe von gesamten Genomen fertig zu werden, sollten die Speicherkapazität und die Zeit des Algorithmus nur linear zur eingegebenen Sequenzlänge zunehmen.

2. Flexibilität und Signifikanz:

Das Programm sollte „flexibel“ sein, d.h., es sollte nicht nur exakte, sondern auch degenerierte und nicht nur direkte, sondern auch palindromische Repeats erkennen können. Außerdem sollte es die Signifikanz der gefundenen Repeats mit Hilfe statistischer Verfahren bewerten können.

3. Visualisierung:

Die erhaltenen Daten müssen visuell dargestellt werden. Dabei soll nicht nur ein Überblick über die Daten gewonnen werden, sondern es muss auch das Zoomen ins Detail möglich sein.

4. Zusammenwirken mit anderen Programmen:

Da die Untersuchung von Repeats nur einen kleinen Teil von Genstrukturanalysen darstellt, ist es notwendig, dass das Programm mit anderen entsprechenden Analyseprogrammen interagieren und Daten austauschen kann.

3. REPuter

REPuter ist ein Programm zur Berechnung von Repeatstrukturen in Genomen, das die oben genannten Kriterien erfüllt. Es besteht im Wesentlichen aus zwei Teilprogrammen, REPfind und REPvis. REPfind dient dabei zum Auffinden von Repeats und gibt die erhaltenen Daten im ASCII Format zurück, während REPvis die Visualisierung dieser Daten übernimmt.
Beide Teilprogramme werden im Folgenden genauer erläutert.

3.1. REPfind

Das Programm nutzt Suffixbäume, um exakte Repeats mit linearem Zeit- und Speicheraufwand zu lokalisieren. Die Laufzeit der Berechnung richtet sich dabei direkt nach der Länge der eingegebenen DNA-Sequenz und nach der Anzahl der gefundenen Repeats.
Die exakten Repeats werden dann als Vorlagen (Seeds) zur Konstruktion von signifikant degenerierten Repeats, welche Fehlpaarungen, Insertionen oder Deletionen erlauben, verwendet. Dabei sind die Signifikanzparameter vom Benutzer frei wählbar. Die Ausgabe ist nach dem Grad der Signifikanz sortiert. All dies ist für die Lokalisation von direkten (Vorwärts-) und palindromischen (revers-komplementären) Repeats möglich.

3.1.1. Mathematische Grundlagen

Die DNA Sequenz sei gegeben als String S der Länge $|S|$ über einem Alphabet $\sum$ mit den Elementen A, T, G, C. $S[i]$ bezeichnet das i-te Zeichen (Base) von S, für *alle* $i \in [1,n]$. S^{-1} bezeichnet die Umkehrung von S.
Für $i \leq j$ bezeichnet $S[i, j]$ einen Teilstring von S, der mit dem i-ten Zeichen beginnt und mit dem j-ten Zeichen endet und durch das Positionenpaar (i, j) beschrieben wird. Die Länge des Teilstrings ist $l(i, j) = j - i + 1$.
Da es nötig ist, die jeweils linken und rechten Zeichen eines jeden Zeichens in S zu referenzieren, werden als erstes Zeichen $S[0]$ und als letztes Zeichen $S[n+1]$ in S zwei verschiedene Zeichen verwendet, die nirgendwo sonst in S auftreten. Die Enden des Strings sind somit gekennzeichnet.

Abb. 1: Teilstring (i, j) einer Sequenz S.

Ein Positionenpaar (i_1, j_1), mit $i_1 \leq j_1$, enthält genau dann ein weiteres Positionenpaar (also einen Teilstring), wenn $(i_2 > i_1)$ und $(j_2 < j_1)$ ist.
Ein Paar Teilstrings (p_1, p_2), also ein Paar von Positionenpaaren enthält genau dann ein weiteres Paar Teilstrings (p_3, p_4), wenn p_3 in p_1 und p_4 in p_2 enthalten ist.

3.1.2. Auffinden von exakten Repeats

Bei einem Paar von Teilstrings $R = ((i_1, j_1), (i_2, j_2))$ handelt es sich genau dann um einen exakten Repeat, wenn $(i_1, j_1) \neq (i_2, j_2)$ und $S[i_1, j_1] = S[i_2, j_2]$. Die Teilstrings dürfen also bei identischem Inhalt keine identischen Positionen einnehmen, d.h. die Positionenpaare müssen unterschiedlich sein.

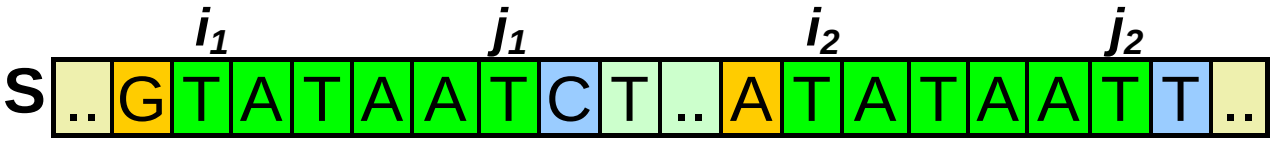

Abb. 2: Maximaler exakter Repeat

Die Länge des Repeats ist $l(R) = j_1 - j_1 + 1 = j_1 - i_2 + 1$.

Ein exakter Repeat $R = ((i_1, j_1), (i_2, j_2))$ wird genau dann als 'maximal' bezeichnet, wenn die linken und rechten Ränder nicht übereinstimmen, also $S[i_1 - 1] \neq S[i_2 - 1]$ und $S[j_1 + 1] \neq S[j_2 + 1]$. Nicht-maximale Repeats sind in maximalen Repeats enthalten.

Um auch palindromische Repeats aufzufinden, muss der String auf das Vorhandensein des reversen Komplements eines Teilstrings hin durchsucht werden. Dies ist der Fall, wenn $S[i_1, j_1] = \overline{S[i_2, j_2]}$.

Die Berechnung von maximalen Repeats basiert auf Hashing-Methoden. Dabei wird eine gegebene DNA-Sequenz w mit einer festen Länge r und ihre Positionen $P(w)$ in S in eine Tabelle eingetragen, wenn sie im String vorgefunden wird.

Für jedes w und i,j ϵ P(w) mit i < j ist ((i, i + r – 1),(j, j + r – 1)) ein exakter Repeat. In Worten: Jeder Tabelleneintrag, in dem Länge und Inhalt des Strings übereinstimmen ist ein exakter Repeat des vorgegebenen Teilstrings.

Um nun maximale Repeats mit der Mindestlänge l zu finden, muss jeder gefundene exakte Repeat nach links und rechts erweitert werden, um festzustellen ob er Teil eines maximalen Repeats der gewünschten Länge ist. D.h. die Teilstrings werden um jeweils ein Zeichen nach links oder rechts erweitert und die gefundenen Zeichen paarweise miteinander verglichen.

In der Praxis geben die verwendeten Hashing-Techniken eine Sequenzlänge r von 10 bis 13 Zeichen (Basen) vor, während die gewünschte Repeat-Länge normalerweise bei mindestens 20 liegt. Im Genom werden jedoch äußerst viele kürzere Repeats gefunden, wie sie durch die Hashing-Techniken vorgegeben sind. Im Vergleich ist die Zahl der gesuchten längeren (maximalen) Repeats sehr gering. Um die kleine Zahl an maximalen Repeats nicht aus den

vielen Kanditaten-Repeats herausfiltern zu müssen, wird für die interne Darstellung von S ein Suffix-Baum verwendet.
Des Weiteren wird der Gusfield-Algorithmus verwendet, um die maximalen exakten Repeats direkt zu berechnen. Dieser benötigt einen linearen Zeitaufwand, der sich aus der Länge des Strings (der DNA-Sequenz) und der gefundenen Repeats zusammensetzt: $O(n + z)$.
Somit ist der Rechenaufwand unabhängig von der Repeatlänge, und es kann auch nach sehr langen Repeats, beispielsweise nach einer Duplikation ganzer Chromosomensegmente, gesucht werden.

3.1.3. Berechnung von degenerierten Repeats und Signifikanz

Degenerierte Repeats können mit unterschiedlichen Methoden beschrieben werden.
Die Hamming – Distanz zwischen zwei gleichlangen Strings, $d_H(S_1, S_2)$, ist die Anzahl an Positionen, wo sich die beiden Strings unterscheiden.
Die Edit – Distanz (Levenshtein – Distanz) zwischen zwei gleich langen Strings, $d_E(S_1, S_2)$, ist die minimale Anzahl von Editierungsoperationen die nötig sind, um S_1 in S_2 zu transformieren. Operationen sind hierbei Deletionen, Insertionen und Fehlpaarungen von einzelnen Zeichen.
In REPfind ist für jede Methode ein Algorithmus implementiert, mit dessen Hilfe aus den gefundenen exakten Repeats jeder, nach den Benutzerangaben mögliche, degenerierte Repeat berechnet und die Sequenz auf dessen Vorhandensein überprüft wird.
Des Weiteren ist es möglich die Signifikanz der gefundenen Repeats zu berechnen. Dabei wird die Anzahl der gefundenen exakten oder degenerierten Repeats mit der statistisch erwarteten Anzahl in einer zufälligen DNA-Sequenz verglichen.

3.2. REPvis

Dieses Programm ist speziell dafür entworfen worden, um es Biologen zu ermöglichen, die anfallenden Datenmengen schnell und effizient auszuwerten.
Die angefallenen Daten werden von REPfind in Form einer ASCII – Tabelle ausgegeben, wobei jeweils Länge und Position der beiden Repeat-Teile sowie seine Ausrichtung, die Zahl von Abweichungen im Repeat und sein E-Value angegeben werden. Der folgende Tabelleneintrag, 8 103 F 8 159 0 6.83e-01, bedeutet also, dass ab den Positionen 103 und 159 eine Sequenz mit 8 Zeichen (Basen) Länge existiert, die direkt (forward, F) und exakt wiederholt ist (keine Abweichungen), und einen bestimmten Signifikanzwert aufweist.
REPvis stellt nun diese Daten in Form eines Visualisierungs-Fensters dar.

Abb.3: Visualisierungsfenster von REPvis

Das Fenster des Repeat-Graphen wird nach oben und unten von zwei horizontalen Balken begrenzt. Der obere repräsentiert die eingegebene DNA-Sequenz, der untere ihre Kopie. Die Repeats sind über Linien miteinander verbunden, wobei die Farbe der Linien die Länge der jeweiligen Repeats kennzeichnet. Über einen Schieber ist die minimale Größe der dargestellten Repeats regelbar. Dieses Fenster gibt einen ersten Überblick über die Anzahl und die Verteilung der gefundenen Repeats.

Mit einem Mausklick ist das heranzoomen von auffälligen Regionen möglich. Dabei öffnet sich das Inspector-Fenster, das den vergrößerten Ausschnitt wiedergibt.

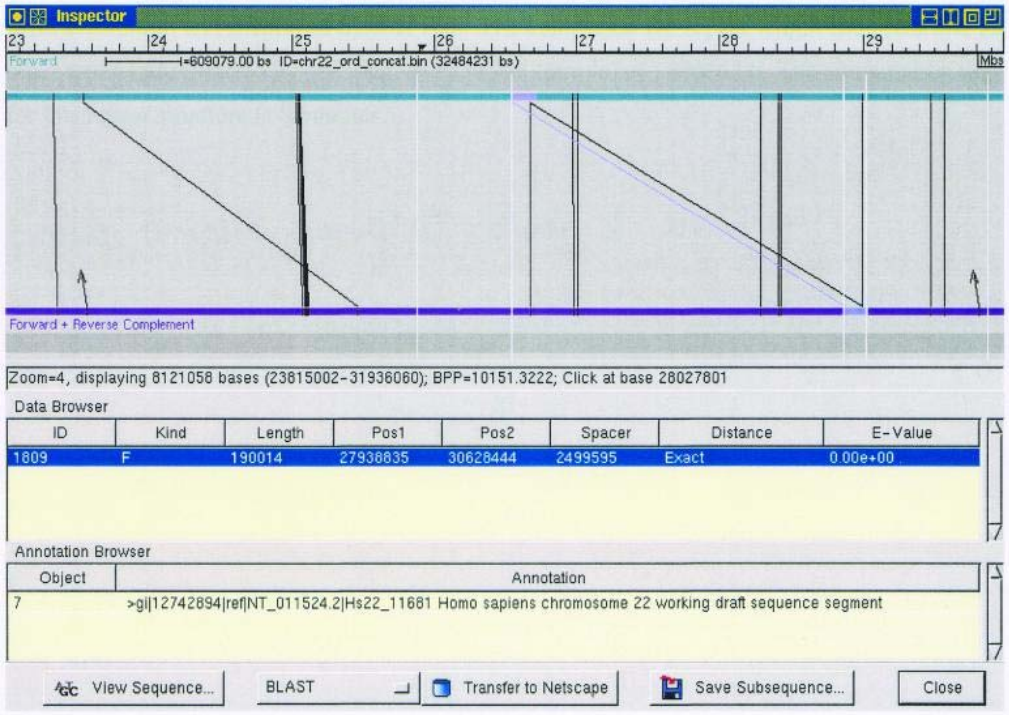

Abb.4: Das Inspectorfenster von REPvis zeigt einen Ausschnitt der obigen Abbildung.

4. Anwendungsgebiete

Im Wesentlichen gibt es fünf Anwendungsbereiche, in denen REPuter eingesetzt werden kann. Diese werden nachfolgend genauer erläutert.

4.1. Assembly checking

Es gibt bestimmte Computerprogramme, die DNA-Sequenzen so zusammensetzen, dass komplette BACs oder Contigs entstehen. Dabei muss jedoch gesichert sein, dass die Sequenzen richtig herum und in der richtigen Reihenfolge eingesetzt werden. Dies kann mit Hilfe von REPuter überprüft werden. So können beispielsweise sehr lange Repeats entstehen, wenn sich einzelne Sequenzen überlappen, was durch REPfind erkannt werden kann. Palindromische Repeats weisen darauf hin, dass Sequenzen möglicherweise falsch herum eingesetzt wurden.

So wurde beispielsweise der Zusammenbau von 11 Contigs des menschlichen Chromosoms 22 überprüft. Dabei ist ein außergewöhnlich langer Repeat entdeckt worden. Beim genaueren Betrachten der betroffenen Region wurde festgestellt, dass das Contig 8 schon zu Beginn des Contig 7 eingebaut war, was daraufhin korrigiert werden konnte.

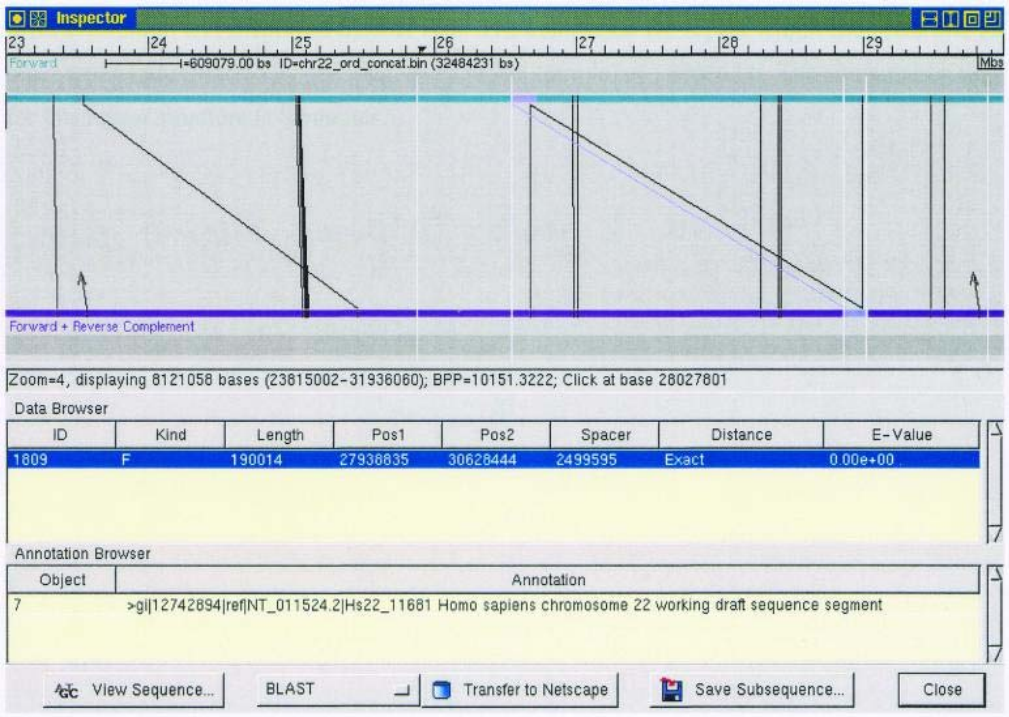

Abb.5: Diese Abbildung zeigt die durch REPvis realisierte graphische Darstellung der Repeats der Contigs 5 – 9 vom Chromosom 22, die jeweils durch einen weißen Strich getrennt sind. Auffällig ist dabei der sehr lange Repeat (lila gekennzeichnet) in Contig 7 und 8, der darauf hinweist, dass das Contig 8 bereits im Contig 7 vorkommt.

4.2. Low copy repeats

Viele genetisch bedingte Krankheiten sind Folgen von Deletionen oder Duplikationen innerhalb von Chromosomen. Entstehen diese Deletionen bzw. Duplikationen innerhalb von Genen, kann dies schwerwiegende Folgen für die Entwicklung oder den Stoffwechsel haben.

Häufig sind betroffene Regionen von Repeats flankiert, die als Bruchstellen fungieren. So sind beispielsweise Transposons bzw. Retrotransposons, die durch ihren Einbau zu solchen Deletionen führen können, immer von direkten Repeats flankiert, die eine wichtige Rolle spielen beim Ausschneiden bzw. beim Einbau dieser Elemente. Solche Strukturen können mit Hilfe von REPfind gefunden werden, was wiederum Aufschlüsse gibt über Regionen, die möglicherweise deletiert sind.

4.3. Einzigartige Sequenzen

Viele biologische Untersuchungen werden heutzutage mit Hilfe von Hybridisierungsmethoden bewerkstelligt. Diese Methoden beruhen darauf, dass eine markierte DNA-Probe mit einer bestimmten Zielsequenz hybridisiert. Somit können diese Sequenzen sichtbar gemacht werden, was beispielsweise bei Chromosomenuntersuchungen anhand von FISH-Analysen zunutze gemacht wird oder bei der DNA-Chip-Technologie. Dabei ist es jedoch wichtig, dass die eingesetzten Proben aus einzigartigen Sequenzen bestehen, da sie nur an eine bestimmte Zielsequenz binden sollten. Enthalten die Proben repetitive Sequenzen, die an mehreren Stellen im Genom vorkommen, entstehen falsch-positive Ergebnisse.

Dies kann mit Hilfe von REPuter verhindert werden, da REPfind sämtliche repetitive Sequenzen herausfiltert. Die übrigen nicht markierten Bereiche sind damit einzigartig und können als Hybridisierungsproben eingesetzt werden, aber auch zur Herstellung von Primern, die ebenfalls einzigartig sein sollten und nur an bestimmte Stellen binden sollten.

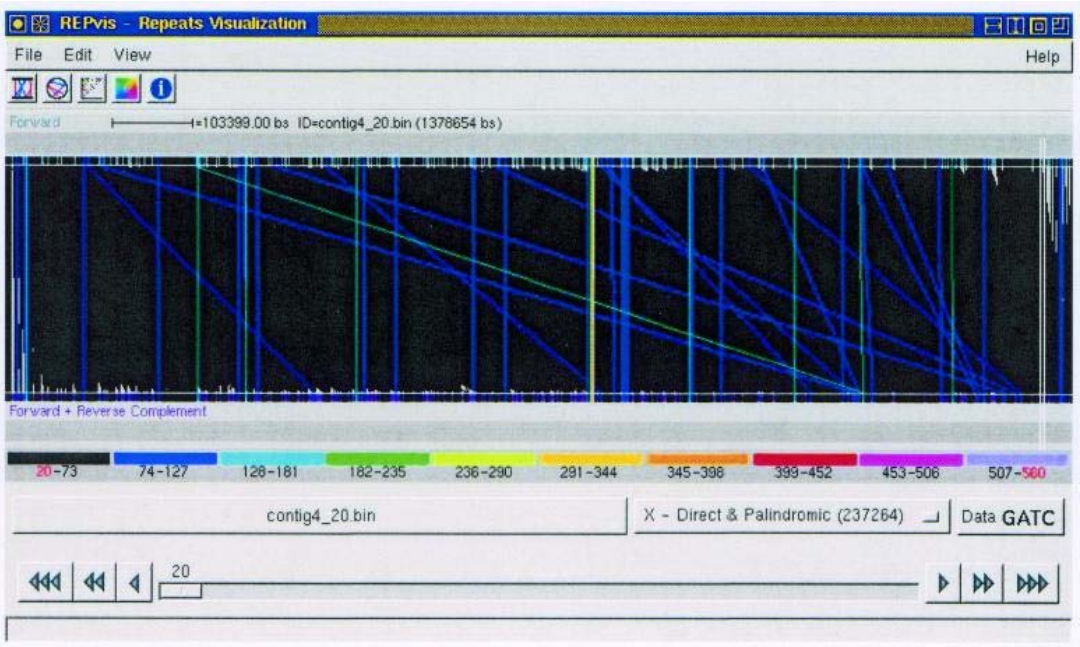

Abb.6: In dieser Abbildung sind die direkten und palindromischen Repeats des Contigs 4 vom Chromosom 21 mit einer minimalen Länge von 20bp dargestellt. Dabei sieht es aufgrund des schwarzen Hintergrundes so aus, als gäbe es keine einzigartigen Sequenzen in diesem Contig.

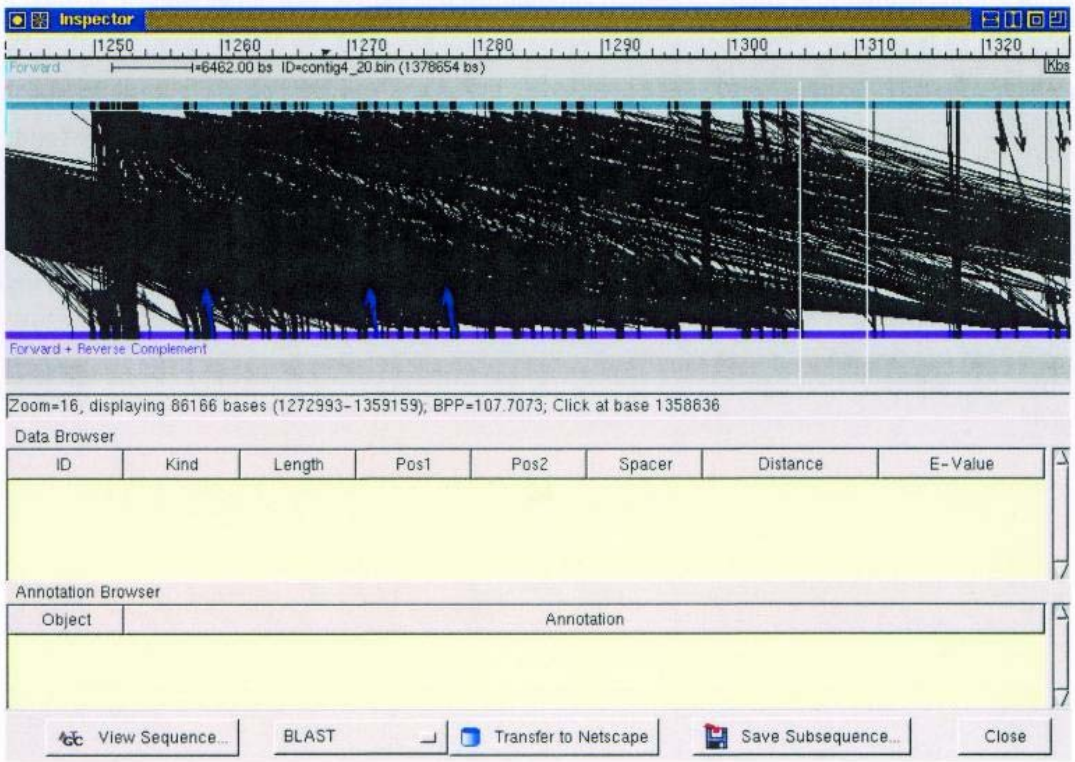

Abb.7: Diese Abbildung zeigt nun einen vergrößerten Ausschnitt der oberen Abbildung. Hier ist nun sehr deutlich zu erkennen, dass sehr wohl einzigartige Sequenzen existieren (weiße Abschnitte).

4.4. Vergleich von Genstrukturen

Die Maus dient für viele biologische Untersuchungen als Modellorganismus, da das Mausgenom am besten untersucht ist und da es die größte Ähnlichkeit mit dem menschlichen Genom aufweist. So können Regulationsprozesse anhand der Maus relativ einfach untersucht werden, wobei die Ergebnisse aufgrund der vielen gemeinsamen Strukturen und Mechanismen meist auch auf den Menschen zurückzuführen sind.

REPuter erleichtert nun die Bewertung der Ergebnisse, da aufgrund der Repeatstrukturanalyse die Gensequenzen verschiedener Organismen miteinander verglichen werden können und der tatsächliche Ähnlichkeitsgrad, beispielsweise zwischen Mensch und Maus, ermittelt werden kann.

Außerdem kann der Vergleich von Repeatstrukturen verschiedener Organismen helfen, die funktionelle und evolutionäre Rolle von Repeats aufzuklären, worüber bis jetzt noch sehr wenig bekannt ist.

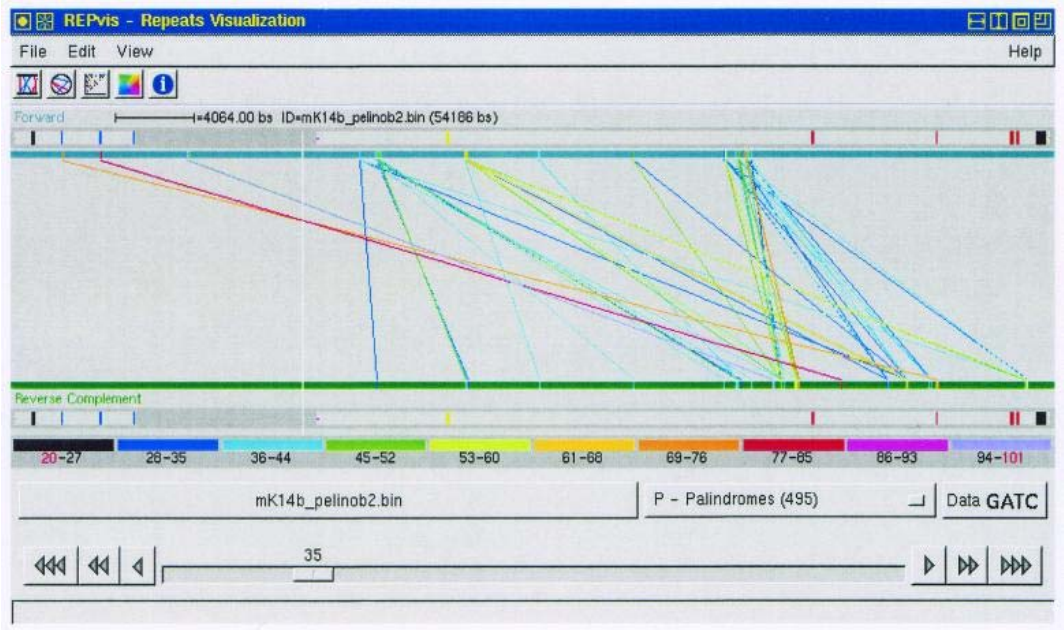

Abb.8: Gezeigt ist ein Concatemer der Pellino1-Gensequenzen der Maus und des Menschen, getrennt durch die weiße Linie. Da das Gen in Maus und Mensch auf den entgegen gesetzten Strängen exprimiert wird, sind nur palindromische Repeats dargestellt. Die vom GeneScan-Programm vorhergesagten Genstrukturen sind über bzw. unter dem Repeat-Fenster dargestellt, die Exons sind als vertikale Balken eingetragen. An der roten Linie ist deutlich zu erkennen, dass eine Exon-Sequenz des Maus-Gens auch im Menschen vorhanden ist, aber von GeneScan nicht als Exon identifiziert wurde.

4.5. cDNA/EST mapping

Als Expressed Sequence Tags (ESTs) werden kurze cDNA-Fragmente bezeichnet, die kleine Abschnitte einer proteincodierenden mRNA wiedergeben, die zur eindeutigen Identifizierung dieser mRNA ausreichen. In vielen Datenbanken sind hunderttausende ESTs gespeichert.

REPuter ermöglicht es nun, eine DNA-Sequenz auf das Vorhandensein bestimmter ESTs hin zu überprüfen, indem die Sequenz nach Repeats der ESTs hin durchsucht wird. Somit können noch wenig charakterisierte DNA-Sequenzen dahingehend untersucht werden, ob sie bestimmte proteincodierende Bereiche besitzen.

5. Quellenangaben

1.: REPuter: the manifold applications of repeat analysis on a genomic scale; Stefan Kurtz, Jomuna V. Choudhuri, Enno Ohlebusch, Chris Schleiermacher, Jens Stoye, and Robert Giegerich

2.: REPuter: fast computation of maximal repeats in complete genomes; Stefan Kurtz and Chris Schleiermacher

3.: Hauptseminar Computational Comparative Genomics: REPuter; Thorsten Gräßle

4.: http://www.genomes.de

5.: http://bibiserv.techfak.uni-bielefeld.de/reputer/