Software zur Hochdurchsatzanalyse von Verteilungen und Distanzkorrelationen genomischer Elemente

Christian Naßl

27. März 2008

Diplomarbeit zur Erlangung des akademischen Grades Diplom-Ingenieur (FH) des Studiengangs Bioinformatik

Fachhochschule Weihenstephan

University of Applied Sciences

Fachbereich Biotechnologie und Bioinformatik

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass die vorliegende Arbeit von mir selbst und ohne fremde Hilfe verfasst und noch nicht anderweitig für Prüfungszwecke vorgelegt wurde. Es wurden keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt. Wörtliche und sinngemäße Zitate sind als solche gekennzeichnet.

Freising, den .....

.....
Unterschrift

Danksagung

Bedanken möchte ich mich bei Frau Kerstin Cartharius, Herrn Prof. Dr. Bernhard Haubold und Herrn Prof. Dr. Frank Leßke für die sehr gute Betreuung meiner Diplomarbeit. Ferner danke ich Herrn Dr. Matthias Scherf und Herrn Andreas Klingenhoff für das Korrekturlesen dieser Arbeit und wissenschaftliche Ratschläge.

Inhaltsverzeichnis

1 Zusammenfassung

Genomische Elemente sind Bereiche auf einem Genom mit einer definierten Funktion, wie beispielsweise Transkripte, Promotoren, Enhancer und microRNAs. Sie sind durch Position und Ausdehnung auf dem Genom definiert.
Um einen tieferen Einblick in die Struktur und die Organisation eines Genoms zu erhalten, ist es sinnvoll, zu untersuchen wie genomische Elemente über das Genom verteilt sind und ob es typische Abstände zwischen unterschiedlichen genomischen Elementen gibt. Besonders interessant sind dabei die Distanzkorrelationen zwischen Elementen, von denen bekannt ist, dass sie in einem Bezug zueinander stehen, zum Beispiel Transkriptionstart-Sites und Transkriptionsfaktorbindungsstellen in Promotoren.
Ziel meiner Diplomarbeit bei der Firma *Genomatix* war es, eine Software zu entwickeln, mit der Verteilungen und Distanzkorrelationen von genomischen Elementen untersucht werden können. Die Anzahl der Elemente in einem Datensatz kann dabei mehrere 100000 oder Millionen betragen.
Für die Analyse der Verteilung genomischer Elemente wurde ein auf r-Scans basierender Algorithmus entwickelt. Damit lassen sich Fragen bezüglich der Häufung oder des besonders gleichmäßigen Auftretens von Elementen im Genom beantworten. Anwendung fand der Algorithmus bei der Suche nach Clustern von mRNA-Fragmenten aus *NextGenerationSequencing*-Projekten [3], die auf das Humangenom gemappt worden waren.
Außerdem wurde ein Verfahren entwickelt, mit dem sich zwei Datensätze mit genomischen Elementen im Hinblick auf ihre Abstände untersuchen lassen. Durch eine graphische Darstellung der Distanzkorrelationen lässt sich so leicht feststellen, ob es Abstände zwischen genomischen Elementen gibt, die signifikant häufig auftreten.
Auf diese Weise wurde untersucht, ob Transkriptionsfaktorbindungsstellen häufiger in Bereichen mit bestimmten Abständen vom Transkriptionstart zu finden sind. Dies ist für eine Reihe von Transkriptionsfaktoren der Fall, konnte aber nicht für alle Transkriptionsfaktoren beobachtet werden. Außerdem wurde gezeigt, dass die Aussagekraft solcher Analysen maßgeblich von der Güte der Annotation der Promotoren abhängt. Wie er-

wartet stellte sich heraus, dass bei manchen Transkriptionsfaktoren diese Bereiche nur in solchen Promotoren deutlich zu sehen sind, von denen bekannt ist, dass der Transkriptionsfaktor dort regulierende Wirkung hat. So kann man bei Promotoren, die nur in einem spezifischen Gewebe die Transkription von Genen regulieren, deutliche Distanzkorrelationen beobachten.
Eine weitere Anwendung fand die Distanzkorrelationsanalyse bei der Untersuchung der Abstände von konservierten Regionen zu Transkripten, Exons und microRNAs. Es konnten Bereiche ermittelt werden, die besonders stark zwischen verschiedenen Spezies konserviert sind.

2 Einleitung

Die Entschlüsselung des Humangenoms und der Genome anderer höherer Vertebraten wie Maus, Ratte und Schimpanse haben in den letzten Jahren zu einem enormen Anwachsen des Bestands an Sequenzdaten geführt. Allerdings liefert das Sequenzieren alleine noch keine Hinweise auf die Organisation der Genome und die jeweilige Funktion von Teilbereichen. Um Gene und andere genomische Elemente annotieren zu können, gibt es diverse Algorithmen, die in einem Genom nach bestimmten Mustern suchen und so eine fundierte Genomannotation erstellen.
Auf der Basis einer Genomannotation lassen sich nun interessante Fragestellungen bezüglich der Relation von genomischen Elementen beantworten. Bisher gibt es jedoch nach meinem derzeitigen Wissensstand kaum Software, die im großen Maßstab die Verteilung von genomischen Elementen und ihre Abstände zueinander auswerten kann.
Im Rahmen meiner Diplomarbeit sollte nun eine Software zu entwickelt werden, die zur Hochdurchsatzanalyse von Verteilungen und Distanzkorrelationen genomischer Elemente eingesetzt werden kann.
Zum besseren Verständnis werde ich im Folgenden einen kurzen Überblick über einige genomische Elemente und ihre Darstellung in den Datenbanken von *Genomatix* geben, die für diese Diplomarbeit verwendet wurden.

2.1 Struktur und Darstellung von genomischen Elementen

Genomische Elemente sind über das Genom verteilte Bereiche mit einer definierten Funktion. Die bekanntesten dieser Elemente sind die Gene, das heißt Bereiche, die für funktionelle Produkte wie Proteine codieren. Ein Gen selbst ist in der Regel wieder in Unterelemente wie Exons und Introns unterteilt.

2.1.1 Transkripte

Der Bereich innerhalb eines Gens, der in mRNA übersetzt wird, heißt Transkript. Die Abschnitte am Anfang und Ende eines Transkripts werden normalerweise nicht translatiert. Diese nichttranslatierten Bereiche am Rand eines Transkripts werden als *untranslated regions* (UTR) bezeichnet. Die Transkripte bei *Genomatix* stammen aus verschiedenen Quellen wie *RefSeq* [6] und *GenBank* [5]. Sie werden je nach Güte der Annotation in Gold-, Silber- und Bronze-Transkripte eingeteilt. Bronze- und Silber-Transkripte sind Transkripte, die auf das Genom gemappt wurden, ohne dass experimentell bewiesen wurde, dass die 5'-UTR vollständig ist. Silber-Transkripte überlappen zusätzlich mit einer Region, die in silico mit dem Programm *PromoterInspector* als Promotor identifiziert wurde. Bei Gold-Transkripten wurde mit Experimenten bewiesen, dass sie in 5'-Richtung vollständig sind [1]. Je besser das Qualitätsmaß eines Transkripts ist, desto sicherer ist, dass es richtig und vollständig auf dem Genom lokalisiert wurde. Eine Region, von der bekannt ist, dass dort eine oder mehrere Transkriptionstart-Sites (TSS) liegen, wird als Transkriptionstart-Region (TSR) bezeichnet.

2.1.2 Promotoren

Als Promotoren werden Bereiche auf dem Genom bezeichnet, die vor (upstream) Transkripten liegen und regulierende Wirkung auf das Transkript haben. Bei *Genomatix* sind Promotoren so definiert, dass sie über 100 bp mit dem Transkript überlappen. Somit befindet sich die Transkriptionstart-Site 100 bp upstream vom Promotorende. Promotoren werden in der Regel mit einer Länge von 600 bp annotiert.

Abbildung 2.1 zeigt in einer schematischen Übersicht die typische Struktur eines Gens mit Transkript und vorgeschaltetem Promotor. Das Transkript selbst ist in Exons und Introns unterteilt. Ferner sind im Promotor Bindungsstellen für Transkriptionsfaktoren zu sehen.

2.1.3 Transkriptionsfaktorbindungsstellen

Transkriptionsfaktorbindungsstellen (TF-Sites) sind kurze Regionen (10-20 bps) im Promotor mit einer spezifischen Verteilung von Nukleotiden. Durch Bindung von speziellen Proteinen, den Transkriptionsfaktoren, an diese Bindungsstellen wird die Transkription reguliert. In der Regel ist es eine Kombination von Transkriptionsfaktoren, die die

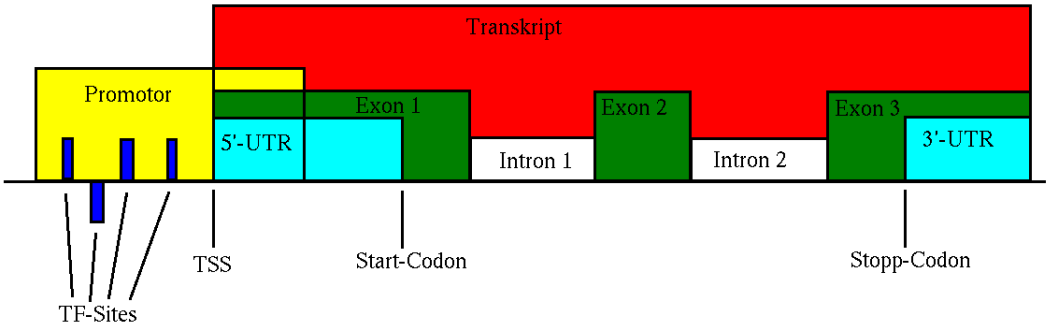

Abbildung 2.1: Struktur eines Gens
Leserichtung ist von links nach rechts (5' nach 3'). Alle genomischen Elemente außer einer TF-Site befinden sich auf dem Vorwärtsstrang (oben)

Transkription von einem Gen regulieren. Meist befinden sich innerhalb eines Promotors Bindungsstellen für verschiedene Transkriptionsfaktoren. Die Bindungsstellen können sowohl auf dem Strang liegen, auf dem der Promotor annotiert wurde, als auch auf dem Gegenstrang (siehe Abbildung 2.1). TF-Sites sind verhältnismäßig kurz, ihre Länge kann aber zwischen verschiedenen Familien von TF-Sites variieren.
TF-Sites werden bei *Genomatix* mit Hilfe des Programms *MatInspector* auf dem Genom lokalisiert [8, 21, 9]. Das Programm findet potentielle TF-Sites mit Hilfe einer Bibliothek von Positions-Gewicht-Matrizen. Es ist zu beachten, dass an die so gefundenen Sites nicht zwangsläufig tatsächlich ein Protein binden muss, da eine physikalische Site nicht notwendigerweise auch eine funktionelle Site ist [9]. In den Datenbanken von *Genomatix* werden die Positionen von TF-Sites relativ zum Beginn des zugehörigen Promotors angegeben.

2.1.4 CAGE-Tags

Die Methode der *cap analysis of gene expression* (CAGE) gibt Aufschluss darüber, wo sich Transkriptionstart-Sites auf dem Genom befinden [12]. Dazu werden sehr kurze Se-

quenzen in der Nähe der Cap-Site von mRNAs sequenziert und auf das Genom gemappt. Stellen auf dem Genom, auf denen mehrere der so erhaltenen CAGE-Tags in einem Cluster liegen, deuten stark darauf hin, dass dort Transkription beginnt. Zur Bestimmung von TSRs wird ein CAGE-Tag-Cluster als TSR annotiert. Anschließend wird die TSR iterativ erweitert, indem alle CAGE-Tags, die sich nicht mehr als 40 bp vom Rand der TSR entfernt befinden, dazugenommen werden [1]. In Abbildung 2.2 ist ein kurzer Ausschnitt des Humangenoms zu sehen, in dem sich CAGE-Tags und eine TSR befinden.

Abbildung 2.2: Darstellung von CAGE-Tags, einem CAGE-Tag-Cluster und einer TSR Die Abbildung zeigt einen Ausschnitt (bp 57780661-57781660) von Chromosom 14 des Humangenoms. In der Abbildung zu sehen sind mehrere CAGE-Tags (dunkelgrün), eine Transkriptionstart-Region (rot) und ein Cluster von CAGE-Tags (blau). Zusätzlich sind ein Promotor (gelb) und ein Ausschnitt aus einem Transkript (grau) dargestellt. Der Transkriptionsstart befindet sich innerhalb des CAGE-TAG-Clusters.

2.1.5 Konservierte Regionen

Als konservierte Regionen werden Bereiche auf dem Genom bezeichnet, die so auch in den Genomen anderer Spezies zu finden sind. Zur Lokalisierung solcher Regionen werden bei *Genomatix* Sequenzen von 15 bp Länge betrachtet, die in zwei Spezies absolut identisch vorkommen. Anschließend wird versucht, diese Kernsequenzen in beide Richtungen so weit wie möglich zu erweitern, wobei aber eine Ähnlichkeit von mindestens 80 % und eine Länge der Sequenz von mindestens 50 bp gefordert wird. Insertionen und Deletionen werden hier nicht berücksichtigt. Konservierte Regionen zwischen sehr nahe verwandten Spezies wie Mensch und Schimpanse sind in den Datenbanken von *Genomatix* nicht annotiert [1].
Konservierte Regionen können überlappen. Dies ist der Fall, wenn die Regionen zwi-

schen verschiedenen Spezies konserviert sind[1]. Für diese Arbeit wurden konservierte Regionen, die überlappen, zu einer konservierten Region zusammengefasst. Tabelle 2.1 zeigt die Anzahl der so erhaltenen konservierten Regionen mit ihrem Mindestgrad der Konservierung.

Konservierungsgrad	Anzahl konservierter Regionen
$\geq 80\%$	1550956
$\geq 85\%$	1263772
$\geq 90\%$	871189
$\geq 95\%$	399154
$\geq 100\%$	69610

Tabelle 2.1: Anzahl konservierter Regionen im Humangenom mit ihrem minimalen Konservierungsgrad

2.1.6 microRNAs

microRNAs sind kurze RNA-Moleküle, die posttranskriptionell Gene regulieren. Dazu bindet die microRNA an eine spezifische mRNA und verhindert, dass diese in ein Protein translatiert wird [7]. Als microRNA werden im Rahmen dieser Diplomarbeit Abschnitte auf dem Genom bezeichnet, von denen bekannt ist, dass sie in eine microRNA transkribiert werden. Die microRNAs in den Datenbanken von *Genomatix* basieren auf den Sequenzen, die in der miRBase [2] des Sanger Instituts hinterlegt sind.

2.2 Über diese Arbeit

Im Rahmen dieser Arbeit wurden die Abstände zwischen genomischen Elementen analysiert. Dazu wurden einerseits die Verteilungen von bestimmten Elementen auf einzelnen Chromosomen oder einem ganzen Genom betrachtet und andererseits die Distanzkorrelationen zwischen unterschiedlichen genomischen Elementen berechnet. Die Elemente werden durch ihre Position und Ausdehnung auf dem Genom repräsentiert. In einzelnen Analysen wurden Datensätze verarbeitet, die mehrere 100000 bis Millionen von Elementen enthalten konnten.

[1] Wenn beispielsweise eine Region zwischen Mensch und Hund konserviert ist und eine andere zwischen Mensch und Rind, so können diese beiden Regionen auf dem Humangenom überlappen.

Indem die Verteilung von einem genomischen Element eines bestimmten Typs untersucht wird, kann festgestellt werden, ob die Elemente an bestimmten Stellen gehäuft auftreten (in Clustern), ob sie gleichmäßig verteilt sind oder weit verstreut liegen. Anwendung fand diese Art der Analyse in dieser Arbeit in Form von Cluster-Analysen. Hier wurde berechnet, an welchen Stellen eines Genoms die betrachteten Elemente gehäuft auftraten, um codierende Bereiche und Transkriptionstart-Regionen zu identifizieren. Bei den Elementen handelte es sich in diesem Fall um kurze Fragmente auf dem Genom, von denen bekannt ist, dass sie transkribiert werden.
Für die Bestimmung von Distanzkorrelationen werden die Abstände, die genomische Elemente eines Typs zu Elementen eines anderen Typs haben, berechnet und gezählt. Zur Interpretation wurden die Distanzhäufigkeiten graphisch dargestellt. Auf diese Weise lässt sich feststellen, ob es typische Abstände oder Intervalle von Abständen gibt, innerhalb derer die Elemente voneinander entfernt auftreten. Von besonderem Interesse ist es dabei herauszufinden, ob es signifikante Abstände zwischen Elementen gibt, die in einem Bezug zueinander stehen. Ein Schwerpunkt dieser Arbeit war es deshalb zu untersuchen, ob es Transkriptionsfaktoren gibt, die in bestimmten Abständen von der TSS entfernt an die DNA binden. Aber auch die Abstände von diversen genomischen Elementen zu konservierten Regionen wurden analysiert, um Elemente zu identifizieren, die überdurchschnittlich häufig stark konserviert sind.
Zudem wurde für diese Diplomarbeit ein Programm mit einer Benutzeroberfläche entwickelt, mit dem man Distanzkorrelationen berechnen und graphisch darstellen kann. Die Software kann dazu verwendet werden, Listen von Positionen genomischer Elemente innerhalb eines Genoms im Hinblick auf statistisch signifikante Abstände zu untersuchen. So können nun allgemeine Fragestellungen wie etwa „Gibt es Bereiche in Promotoren, in denen sich die Bindungsstellen eines bestimmten Transkriptionsfaktors häufen?“ oder auch „Sind bestimmte Abstände zwischen zwei verschiedenen TF-Sites überrepräsentiert?“ beantwortet werden. Außerdem können mit Hilfe der Software Bereiche auf dem Genom identifiziert werden, in denen Cluster von genomischen Elementen liegen.

3 Methoden

3.1 Clusteranalyse mit r-Scans

Für die Untersuchung der Verteilung genomischer Elemente kam die bewährte Methode der r-scans [17, 19, 18, 16, 15, 23, 22] zum Einsatz. Mit dieser statistischen Methode, die auf asymptotischen Formeln basiert, kann untersucht werden, ob die Verteilung von bestimmten genomischen Elementen zufällig ist. Es lässt sich feststellen, ob Elemente statistisch signifikant dicht aneinander auftreten, ob sie übermäßig weit verstreut liegen oder gleich verteilt sind.
Für die r-scan-Statistik werden zuerst die Abstände U_i zwischen aufeinander folgenden Elementen (U_i = Abstand zwischen dem i-ten und $(i+1)$-ten Element) berechnet. Dann werden jeweils r ($r \in \mathbf{N}$) aufeinander folgende Distanzen addiert, wie in Gleichung (3.1) beschrieben, so dass man insgesamt $n - r + 1$ solcher r-Fragmente R_i erhält, wobei n die Anzahl der Elemente ist.

$$R_i = \sum_{j=i}^{i+r-1} U_j \qquad i = 1, 2, ..., n - r + 1 \tag{3.1}$$

Danach müssen die r-Fragmente noch skaliert werden, indem jedes durch die Länge N der betrachteten Sequenz geteilt wird, so dass sich Werte zwischen 0 und 1 ergeben. Für die Clusteranalyse sind nun die k kleinsten (z. B. $k = 3 \Rightarrow$ das 3.-kleinste r-Fragment) dieser r-Fragmente von Bedeutung. Um zu untersuchen, ob das r-Fragment signifikant klein ist, also ob an dieser Stelle ein Cluster liegt, wird nun ein Schwellenwert berechnet, dessen Größe von k, r, der Irrtumswahrscheinlichkeit α und der Anzahl n der Elemente auf dem betrachteten Chromosom abhängt.

$$Pr\left[m_k^{(r)} < \frac{x}{n^{1+\frac{1}{r}}}\right] \approx 1 - exp(-\lambda) * \left(\sum_{i=0}^{k-1} \frac{\lambda^i}{i!}\right) \qquad \lambda = \frac{x^r}{r!} \tag{3.2}$$

Gleichung (3.2) zeigt, wie sich die Wahrscheinlichkeit berechnen lässt, dass das k.-kleinste r-Fragment $m_k^{(r)}$ kleiner als der Schwellenwert $\frac{x}{n^{1+\frac{1}{r}}}$ ist [18]. Der Formel liegt die Annahme zu Grunde, dass die n Elemente zufällig und somit unabhängig voneinander gleichmäßig verteilt sind [19]. Außerdem wird angenommen, dass $n \to \infty$ gilt, weshalb die Formel nur eine Annäherung ist und nur für große n ($n > 10000$) brauchbare Ergebnisse liefert. Mit Gleichung (3.2) wird also die theoretische Wahrscheinlichkeit berechnet, dass $m_k^{(r)}$ kleiner ist als das k.-kleinste r-Fragment, wenn die Elemente zufällig verteilt wären [17]. Ersetzt man nun die linke Seite von Gleichung (3.2) durch die gewünschte Irrtumswahrscheinlichkeit α, können λ, x und damit auch der Schwellenwert berechnet werden, den ein r-Fragment unterschreiten muss, damit behauptet werden kann, dass an dieser Stelle ein Cluster liegt.

Wenn k, r und die Irrtumswahrscheinlichkeit α zunehmen, wird auch der Schwellenwert größer, während er mit zunehmender Anzahl der Elemente abnimmt. Somit muss r mit Bedacht gewählt werden. Ist r zu klein (z. B. $r = 1$), wird der Schwellenwert teilweise kleiner als 1bp, so dass man nur noch Cluster findet, wenn die Elemente überlappen oder direkt aufeinander folgen. Bei einem zu großen r ($r > 10$) nimmt auch die Wahrscheinlichkeit zu, Cluster zu entdecken, bei denen es dann allerdings fraglich ist, ob sie signifikant sind. Außerdem werden kleinere Cluster übersehen. Aus diesen Gründen wurden in dieser Diplomarbeit die Werte für r üblicherweise zwischen 3 und 10 gewählt. Um vernünftige Ergebnisse zu erhalten, sollte immer gelten: $r \ll n$ [19].

Übermäßige Streuung kann mit einer ähnlichen Formel wie Gleichung (3.2) berechnet werden. Hier muss dann betrachtet werden, ob das k.-größte r-Fragment größer als der zugehörige Schwellenwert ist. Setzt man in Gleichung (3.2) für die Wahrscheinlichkeit $1 - \alpha$ ein, kann auch getestet werden, ob die Elemente signifikant gleich verteilt sind, indem man prüft, ob der dann berechnete Schwellenwert kleiner als das k.-kleinste r-Fragment ist [19].

Da man mit den r-scans nur Cluster der Länge $r+1$ lokalisieren, nicht aber ihre vollständige Ausdehnung feststellen kann und auch nur die k signifikantesten Cluster entdeckt werden können, wobei diese auch noch überlappen können, wurde die Methode in dieser Diplomarbeit erweitert.

Um möglichst viele Cluster in ihrer vollen Ausdehnung zu lokalisieren, wurde ein iterativer Algorithmus entwickelt. In einer bestimmten Anzahl von Schritten werden in jedem Schritt zuerst die Distanzen zwischen den Elementen und, ausgehend von diesen Abständen, die r-Fragmente berechnet. Wenn Elemente überlappen, wird die Distanz

zwischen ihnen auf 0 gesetzt. Anschließend wird verglichen, ob die k kleinsten dieser Fragmente kleiner als ihre zugehörigen Schwellenwerte sind. Ist dies der Fall, so wurde ein signifikantes Cluster gefunden, das jetzt vervollständigt werden soll. Dazu werden alle Elemente zu dem Cluster dazugenommen, die davon nicht mehr als einen bestimmten Abstand entfernt sind. Dieser Abstand ist der vorher berechnete Schwellenwert geteilt durch r, da dies auch der mittlere Abstand der Elemente innerhalb eines r-Fragments ist, das gerade noch als Cluster gilt. Das Cluster wird allerdings nur dann vervollständigt, wenn es nicht schon vorher als Teil eines größeren Clusters identifiziert wurde, da r-Fragmente auch überlappen können. Wurden die k kleinsten Fragmente alle betrachtet, wird versucht in der nächsten Iteration weitere signifikante Anhäufungen von Elementen zu finden. Die Elemente, die in der aktuellen Iteration als Teil eines Clusters identifiziert wurden, werden für die weitere Analyse nicht mehr betrachtet.

3.2 Distanzkorrelationen genomischer Elemente

Zur Berechnung von Distanzkorrelationen kamen unterschiedliche Verfahren zum Einsatz, abhängig davon in welcher Form die zu analysierenden Daten vorlagen.

3.2.1 Distanzkorrelationen von beliebigen genomischen Elementen

In der Regel liegen von genomischen Elementen die absoluten Positionen auf den Chromosomen vor. Für die Korrelationsanalyse ist es zuerst notwendig, die Elemente beider Typen, die verglichen werden sollen, getrennt voneinander nach Chromosomen zu ordnen. Ferner muss bei einem Typ ein Bezugspunkt gewählt werden, zu dem die Abstände zu den Elementen vom zweiten Typ berechnet werden. Dieser Bezugspunkt kann der Anfang, das Ende oder die Mitte des Elements sein. Die Elemente von Typ 2 müssen außerdem nach Positionen sortiert werden.
Danach wird die Liste mit den Elementen vom ersten Typ durchlaufen. Für jedes Element von Typ 1 wird mittels binärer Suche ein Element von Typ 2 gesucht, von dem zumindest eine Position einen Abstand zum Bezugspunkt des aktuellen Typ-1-Elements nicht größer als ein bestimmter Maximalabstand m hat. Wird ein solches Element gefunden, muss unterschieden werden, ob das Element auf einen Punkt reduziert oder in seiner vollen Ausdehnung betrachtet wird. Kurze Elemente ($\leq$ 20 bp) werden üblicher-

weise auf einen Punkt reduziert, indem von der Anfangs- und Endposition des Elements der Mittelwert berechnet wird. Wenn die Distanz von dem Typ-2-Element nun immer noch kleiner als die maximale Distanz ist, wird der Abstand gezählt. Bei längeren Elementen wird von jeder einzelnen Position des Elements der Abstand zum Bezugspunkt gezählt, sofern das Distanz-Kriterium erfüllt wird. Nun wird in 3'- und 5'-Richtung nach weiteren Typ-2-Elementen gesucht, von denen mindestens eine Position vom Bezugspunkt nicht weiter als erlaubt entfernt ist. Auch für diese Elemente wird wieder analog zu dem oben beschriebenen Fall die Distanz gezählt, sofern das Distanz-Kriterium erfüllt ist. In Abbildung 3.1 ist der genaue Ablauf des Algorithmus in einem Ablaufdiagramm graphisch dargestellt.

Zur leichteren Interpretation wird die Häufigkeitsverteilung der Abstände graphisch dargestellt (z. B. Abbildung 3.2). Um die Kurven zu glätten habe ich ein „Sliding-Window"-Verfahren angewendet: Es wurden für die Distanzen Intervalle einer festen Größe[1] gebildet, so dass die Anzahl für ein solches Intervall ebenfalls erhöht wird, wenn eine Distanz in dem Intervall enthalten ist und gezählt wird. Der erhaltene Wert für ein Intervall muss zur Normalisierung dann noch durch die Größe des Intervalls geteilt werden.

Des weiteren wird bei jeder Analyse die durchschnittliche Häufigkeit der Distanzen und die Standardabweichung berechnet. Auch wird angezeigt, welche Distanz am häufigsten auftritt. Ferner wird die Anzahl aller Elemente von Typ 1 und 2 angegeben, die nicht mehr als m bps voneinander entfernt sind, also die Anzahl der Elemente, deren Distanzkorrelationen zueinander tatsächlich in den Graphen dargestellt werden.

Zusätzlich wurde bei einigen Analysen der durchschnittliche GC-Gehalt sowie der durchschnittliche Gehalt an einzelnen Nukleotiden in den betrachteten Sequenzen berechnet. Die Sequenzen reichen von der Position, die sich m bps upstream vom Bezugspunkt befindet, zu der Position, die sich m bps downstream vom Bezugspunkt befindet. Die Länge der Sequenzen ist also der doppelte Maximalabstand. Zur Berechnung des GC-Gehalts wurde für jede Position dieser Sequenzen die Anzahl an Gs und Cs gezählt und die so erhaltene Anzahl normalisiert. Die durchschnittliche Häufigkeit einzelner Nukleotide, also A, C, G und T, wurde analog berechnet und graphisch dargestellt.

In Abbildung 3.2 ist eine Kurve von Distanzhäufigkeiten exemplarisch dargestellt. In der Mitte der horizontalen Achse befindet sich die Position, die in den Elementen vom ersten

[1] Sofern nicht anders angegeben, beträgt die Größe eines Intervalls 10 bp

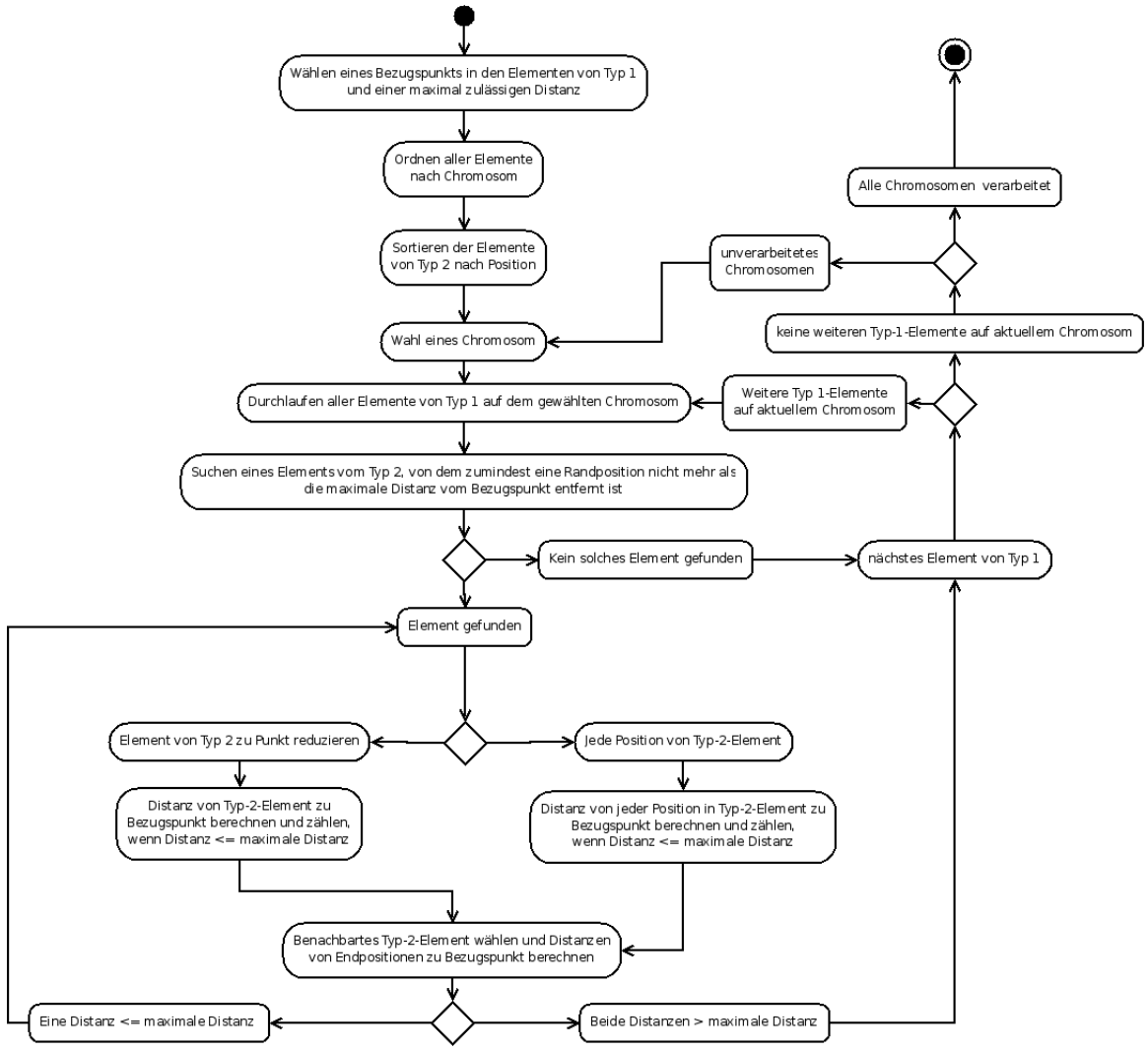

Abbildung 3.1: Ablaufdiagramm des Algorithmus zur Analyse von Distanzkorrelationen

Typ als Bezugspunkt gewählt wurde, weshalb diese Position den Nullpunkt darstellt. Das linke Ende der Abszisse ist die maximal zulässige Distanz zu Elementen, die upstream vom Bezugspunkt liegen, während das rechte Ende den maximal zulässigen Abstand zu Elementen downstream vom Bezugspunkt bildet. Auf der vertikalen Achse am linken Bildrand werden die absoluten Distanzhäufigkeiten aufgetragen, auf der rechten Ordinate der prozentuale Anteil von Nukleotiden an den einzelnen Positionen. In dunkelgrün ist die Verteilung der Distanzhäufigkeiten dargestellt, in braun die geglättete Verteilung. In blau wird der arithmetische Mittelwert der Distanzhäufigkeiten angezeigt (dicke Linie) sowie deren Standardabweichung (dünne Linie). Die weiteren Kurven repräsentieren den Gehalt der einzelnen Nukleotide sowie den GC-Gehalt an den einzelnen Positionen.

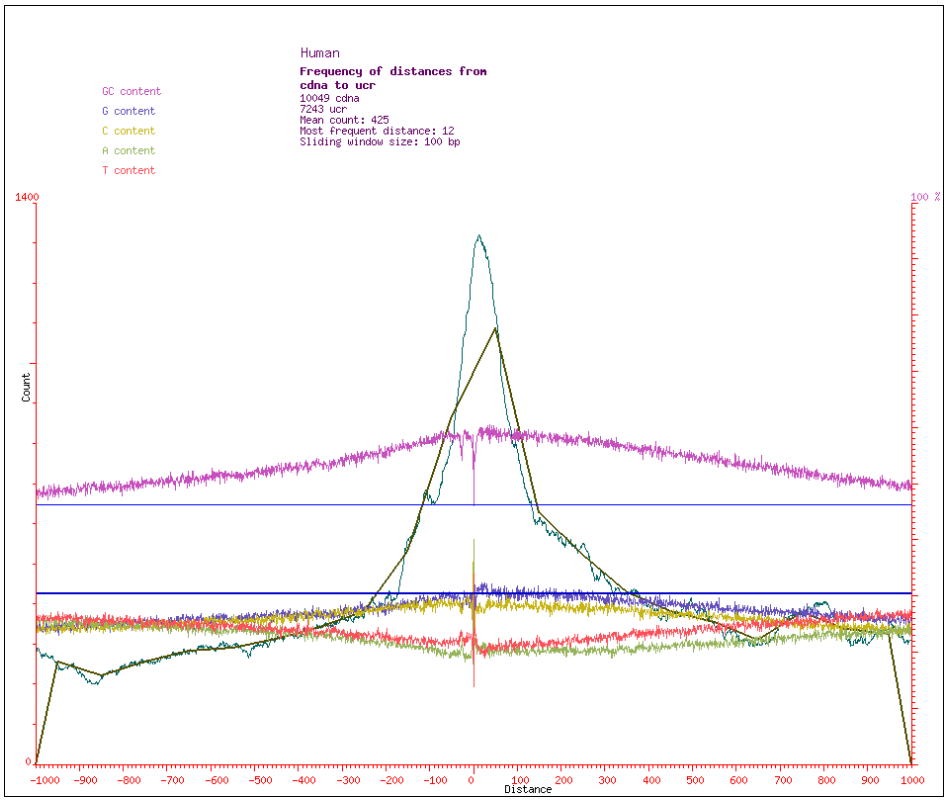

Abbildung 3.2: Distanzkorrelationen zwischen dem 5'-Ende von Transkripten und konservierten Regionen
Die konservierten Regionen wurden in ihrer vollen Ausdehnung betrachtet. Der Grad der Konservierung beträgt 100%, die Größe des Sliding-Windows 100 bp.

3.2.2 Distanzkorrelationen von TF-Sites und Promotoren

Um die Verteilung von Transkriptionsfaktorbindungsstellen in Promotoren zu untersuchen, kam ein Verfahren zum Einsatz, das sich von dem oben beschriebenen in einigen Punkten unterscheidet.
Da die Länge von Promotoren variieren kann, die Transkriptionstart-Site sich definitionsgemäß allerdings immer 100 bp upstream vom Promotorende befindet, macht es Sinn für die Abstandsmessung zwischen einer Transkriptionsfaktorbindungsstelle und einem Promotor als Bezugspunkt stets das 3'-Ende des Promotors zu wählen. Von TF-Sites liegen die Positionen relativ zum 5'-Ende des Promotors vor, in dem sie sich befinden. Aufgrund ihrer geringen Länge wurden TF-Sites für diese Analyse auf einen Punkt reduziert. Für die Berechnung des Abstands einer TF-Site zum 3'-Ende des zugehörigen Promotors, wurde die absolute Position der TF-Site berechnet und diese von der Position des 3'-Endes des Promotors subtrahiert. Da der Großteil der Promotoren in den Datenbanken von *Genomatix* mit einer Länge von 600 bp annotiert ist, wurde als maximal zulässige Distanz 600 bp gewählt.
Downstream vom 3'-Ende eines Promotors sind keine TF-Sites annotiert, weshalb dieser Bereich auch nicht graphisch dargestellt wird. Somit befindet sich der Bezugspunkt am rechten äußeren Rand der horizontalen Achse, während das linke Ende der Abszisse die Position 600 bp upstream vom Bezugspunkt darstellt. Zusätzlich zu dem Gehalt von Nukleotiden im Promotor, wurde außerdem der prozentuale Anteil des Startcodons ATG an den einzelnen Positionen im Promotor graphisch dargestellt. Abbildung 3.3 zeigt exemplarisch die Verteilung von Bindungsstellen eines Transkriptionsfaktors innerhalb von einem Promotor.

3.2.3 Distanzkorrelationen von TF-Sites und CAGE-Tag-Clustern

CAGE-Tags [12] deuten darauf hin, dass in der Nähe dieser Stellen auf dem Genom Transkription beginnt. Für die Analysen in dieser Diplomarbeit wurde der Mittelpunkt von CAGE-Tag-Clustern als hypothetische Transkriptionstart-Site verwendet. Um diesen Punkt wird ein hypothetischer Promotor definiert, der 500 bp upstream beginnt und 100 bp downstream endet (siehe Abbildung 3.4). Diese Sequenz von 600 bp Länge wird dann aus der Datenbank extrahiert und in eine FASTA-Datei geschrieben. Diese Datei dient dem Programm *MatInspector* [9, 8, 21] als Input, das alle theoretischen TF-Sites in der Sequenz lokalisiert.

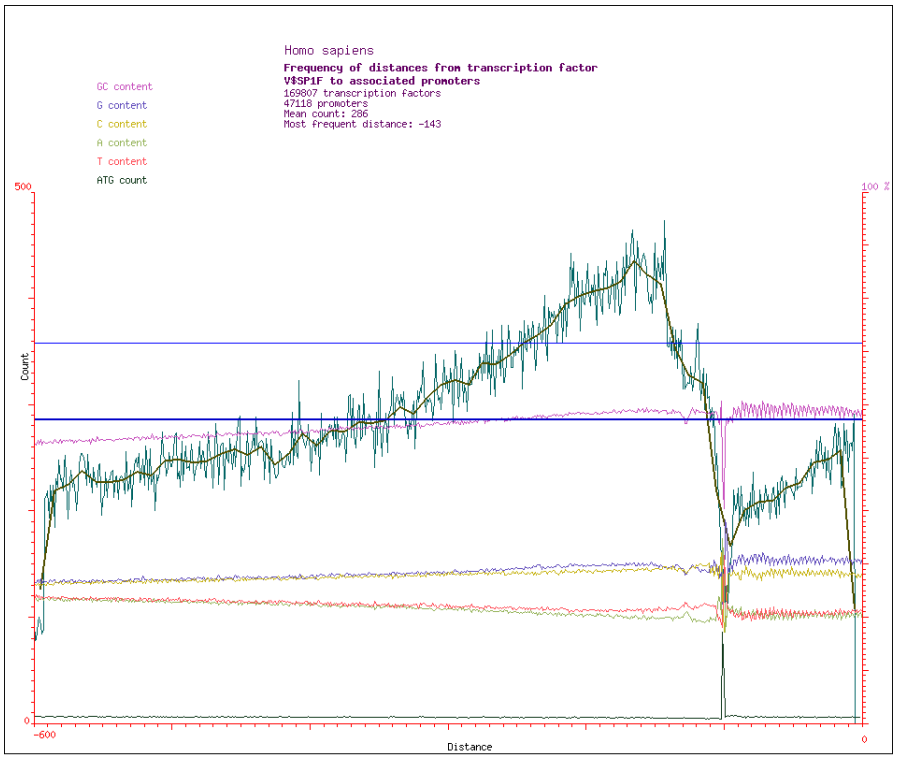

Abbildung 3.3: Verteilung der Bindungsstellen des Transkriptionsfaktors SP1F innerhalb von Promotoren

Die Positionen der TF-Sites werden relativ zum 5'-Ende der Input-Sequenz angege-

Abbildung 3.4: Hypothetischer Promotor:
In orange ist der hypothetische Promotor von 600 bp Länge dargestellt, in grün CAGE-Tags, in blau CAGE-Tag-Cluster, die sich aus mehreren CAGE-Tags zusammensetzen. Der rote Punkt markiert den Mittelpunkt des Clusters und stellt die hypothetische Transkriptionstart-Site dar. Er befindet sich 100 bp upstream vom 3'-Ende des Promotors.

ben. Für die Berechnung der Distanz von TF-Site zu Promotorende wird die TF-Site wieder auf einen Punkt reduziert. Die Berechnung und graphische Darstellung der Distanzhäufigkeiten von TF-Sites erfolgt nach der oben beschriebenen Methode, mit der die Verteilung von TF-Sites innerhalb von annotierten Promotoren untersucht werden kann.

3.3 Software mit webbasierter Benutzeroberfläche

Um mit den von mir entwickelten Algorithmen auch in Zukunft noch weitere Analysen von Verteilungen und Distanzkorrelationen genomischer Elemente durchführen zu können, habe ich zwei Programme mit webbasierter Benutzeroberfläche entwickelt. Beide Programme wurden in Perl implemetiert, für die Erstellung der Benutzeroberfläche wurde die CGI-Technik verwendet.

3.3.1 Tool für Cluster-Analysen

Dieses Programm implementiert den in Abschnitt 3.1 beschriebenen, auf r-Scans basierenden, iterativen Algorithmus. Damit in einem Genom nach Clustern genomischer Elemente gesucht werden kann, werden alle Chromosomen des zu untersuchenden Genoms

konkateniert und als eine lange Sequenz betrachtet, auf der die Elemente verteilt liegen. Die Elemente, die der Benutzer analysieren möchte, werden in einer Datei[2] (BED-File) eingelesen. Anschließend werden die Positionen der Elemente auf der Sequenz bestimmt, die die aneinandergereihten Chromosomen darstellt, und die Elemente entsprechend sortiert. Dann wird die Clusteranalyse gestartet. Zusätzlich wird berechnet, welcher Anteil des gesamten Genoms von den Elementen abgedeckt wird.
Für die Clusteranalyse können folgende Paramter gesetzt werden:

- r (Größe des r-Fragments)
- Irrtumswahrscheinlichkeit
- Anzahl der Iterationen
- Elemente auf beiden Strängen, nur auf Vorwärts- oder nur auf Rückwärts-Strang
- Mindestanzahl von Elementen in einem Cluster (Default: r)

Das Programm hat eine Weboberfläche und wird über einen Internetbrowser aufgerufen. Bevor ein Benutzer eine Analyse starten kann, muss er aus einer Liste den Organismus auswählen, aus dessen Genom die zu untersuchenden genomischen Elemente stammen. Danach können auf der Hauptseite der Anwendung ein BED-File hochgeladen und die für die Clusteranalyse benötigten Parameter gesetzt werden. Nach Klicken auf den Start-Button werden dem Benutzer die Positionen der Bereiche auf dem Genom angezeigt, die als Cluster identifiziert wurden. Außerdem werden die Anzahl der für die Analyse verwendeten Elemente, der berechnete Schwellenwert[3] und der prozentuale Anteil auf dem Genom, der von Elementen abgedeckt wird, angezeigt. Wurde ein falscher Organismus ausgewählt[4] oder keine Datei hochgeladen, so wird eine Fehlermeldung angezeigt.

3.3.2 *GenomeInspector*

Zur Berechnung von Distanzkorrelationen zwischen beliebigen genomischen Elementen wurde der in Abschnitt 3.2.1 beschriebene Algorithmus von mir in Perl implementiert.

[2] Jede Zeile einer solchen Datei entspricht einem genomischen Element und hat vier durch Tabulator getrennte Einträge: Chromosom, Start- und Endposition (ausgehend vom Vorwärtsstrang) sowie optional Strang

[3] siehe Abschnitt 3.1

[4] Das Programm erkennt dies, wenn hochgeladene Elemente auf einem Chromosom liegen, das der Organismus gar nicht besitzt. Beispielsweise gibt es im menschlichen Genom kein Chromosom 25.

Die Software wurde aus historischen Gründen nach dem Programm *GenomeInspector* [23, 22] benannt, mit dem in kurzen Genomen (wenige Megabasenpaare) Distanzkorrelationen berechnet werden konnten.
Um es dem Benutzer zu ermöglichen selbst genomische Elemente einzulesen um sie miteinander zu korrelieren, habe ich eine Methode geschrieben, die BED-Files parst und die eingelesenen Elemente ordnet und gegebenenfalls sortiert. Ebenso besteht die Möglichkeit Elemente aus der Datenbank von *Genomatix* auszulesen[5] um sie miteiander zu korrelieren. Genauso können die Elemente des einen Typs aus einer Datei stammen und die des anderen Typs aus der Datenbank. Außerdem kann der Benutzer folgende Parameter setzen:

- Bezugspunkt[6] in Elementen von Typ 1
- maximal anzuzeigender Abstand zum Bezugspunkt
- Größe des Sliding-Windows
- Reduzieren von Elementen von Typ 2 auf einen Punkt oder Betrachtung in ihrer vollen Ausdehnung
- optional: Gehalt einzelner Nukleotide

Nach der Berechnung der Distanzkorrelationen, wird die Häufigkeitsverteilung der Distanzen graphisch dargestellt. Dazu habe ich auf das Modul GD zur Erstellung von Graphiken zurückgriffen.
Nun kann der Benutzer ein Intervall von Distanzen vorgeben und sich für einen der beiden Typen von Elementen entscheiden, um alle Elemente des ausgewählten Typs auszulesen, die zu mindestens einem Element des anderen Typs einen Abstand haben, der in diesem Intervall liegt. Dazu wird erneut eine Distanzanalyse gestartet. Allerdings werden jetzt nicht die Abstände gezählt, sondern lediglich die Elemente, die das Distanzkriterium erfüllen, gespeichert. Wenn beispielsweise im Graph ein deutlicher Peak zu sehen ist, kann so untersucht werden, welche Elemente diesen Peak verursacht haben. So lässt sich feststellen, ob es zwischen diesen Elementen Gemeinsamkeiten gibt, zum

[5] Von TF-Sites werden außerdem die absoluten Positionen berechnet und überlappende konservierte Regionen werden zusammengefasst
[6] 5'-Ende, 3'-Ende oder Mitte des jeweiligen genomischen Elements

Beispiel in funktioneller Hinsicht. Die extrahierten Elemente können auch für eine weitere Analyse verwendet werden.

Auch für die Berechnung von Distanzkorrelationen muss der Benutzer vor der Analyse

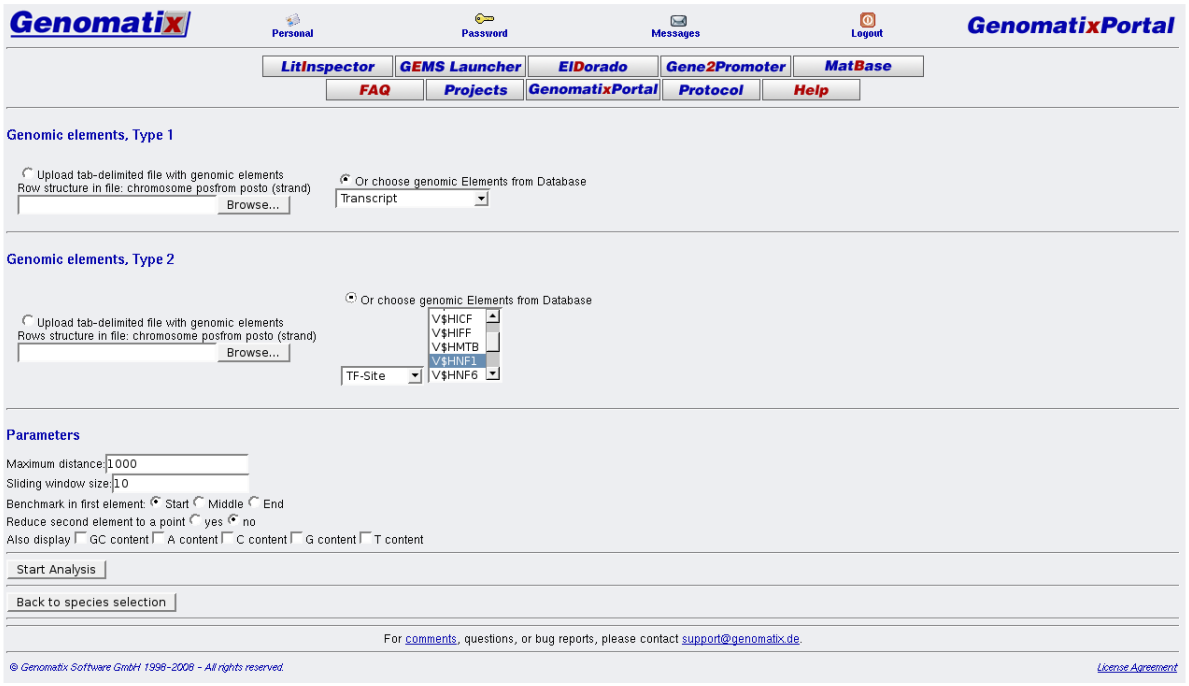

Abbildung 3.5: Screenshot der Hauptseite von *GenomeInspector*

aus einer Liste einen Organismus auswählen, aus dessen Genom die zu untersuchenden Elemente stammen. Danach wird im Webbrowser die Hauptseite der Anwendung (siehe Abbildung 3.5) angezeigt. Hier können Dateien hochgeladen, genomische Elemente aus der Datenbank ausgewählt und sämtliche Parameter gesetzt werden. Ferner wird hier durch Klicken auf den entsprechenden Button die Analyse gestartet. Die Distanz-Korrelationen werden berechnet und der Ergebnis-Graph wird dem Benutzer angezeigt. Konnten keine Distanzen berechnet werden, so wird anstatt des Graphen eine Fehlermeldung angezeigt. Dies ist der Fall, wenn

- ein falscher Organismus ausgewählt wurde[4]
- das Feld „Upload“ markiert, aber keine Datei hochgeladen wurde

- alle Elemente zum Bezugspunkt einen größeren Abstand als den maximal zulässigen haben

Nach erfolgreicher Berechnung der Distanzkorrelationen können nach Vorgabe eines Intervalls Elemente von Typ 1 oder 2 extrahiert werden. Die Positionen der extrahierten Elemente werden dann aufgelistet. Sie können auch als BED-File heruntergeladen werden.

3.4 Interpretation von Kurven

Für die Interpretation der Kurven, die Distanzkorrelationen darstellen, sollen der arithmetische Mittelwert und die Standardabweichung der Distanzhäufigkeiten einen Anhaltspunkt liefern, wann ein Peak signifikant ist. Zwar wird bei normalverteilten Daten gefordert, dass Werte mindestens zwei Standardabweichungen vom Mittelwert entfernt sein müssen, damit von Signifikanz gesprochen werden kann, jedoch ist bei den Datensätzen, die im Rahmen dieser Diplomarbeit für Analysen verwendet wurden, die Verteilung unbekannt.
Generell ist es bei biologischen Daten häufig so, dass ein geschultes Auge für die Interpretation benötigt wird. Jedoch kann es auch vorkommen, dass selbst erfahrene Biologen Daten unterschiedlich interpretieren. Deshalb sollen Mittelwert und Standardabweichung lediglich als Anhaltspunkte dienen und nicht als Richtwerte.
Für die im Rahmen dieser Arbeit als signifikant bezeichneten Distanzen wurde gefordert, dass sie mindestens häufiger als die Standardabweichung der Distanzhäufigkeiten auftreten. Außerdem muss aus der Kurvenform ersichtlich sein, dass es sich bei diesen Peaks nicht um zufälliges Rauschen handelt. Wenn auch die geglättete Kurve die Standardabweichung übersteigt und der Peak in dieser Kurve sich von denen in seiner Umgebung deutlich abhebt, desto sicherer ist, dass signifikante Distanzen gefunden wurden.

4 Ergebnisse

4.1 Anwendung von r-Scans

4.1.1 Cluster von Solexa-Elementen zum Auffinden unbekannter Transkripte

Der in dieser Diplomarbeit entwickelte Algorithmus zur Auffindung von Clustern genomischer Elemente mit Hilfe von r-Scans wurde an Datensätzen getestet, die mittels der Solexa-Sequenzier-Technik [3] aus mRNAs gewonnen worden waren (Im folgenden als Solexa-Elemente bezeichnet). In einem Experiment wurden die mehr oder weniger stark exprimierten mRNAs aus menschlichen B-Zellen und Nierenzellen gewonnen und sequenziert, wobei die mRNAs in Sequenzen von jeweils exakt 27 bp Länge fragmentiert wurden, die stark überlappen können. Nachdem diese Fragmente auf das Genom gemappt worden waren, wurde mit dem r-Scan-Algorithmus nach Häufungen dieser Elemente auf dem Genom gesucht. Dabei wurden die einzelnen Chromosomen separat analysiert. Der Wert für r war in dieser Analyse fest auf 10 eingestellt, die Anzahl der Iterationen betrug 30. Es wurden in jeder Iteration die 9 kleinsten r-Fragmente gesucht. Die Irrtumswahrscheinlichkeit betrug 1%.
Es wurde erwartet, dass Häufungen von Solexa-Elementen vor allem innerhalb von Regionen auf dem Genom zu finden sind, die bereits als Transkripte annotiert worden sind. Dies war auch der Fall. Allerdings wurden auch außerhalb von bekannten Transkripten signifikante Cluster von den Elementen gefunden, was darauf hindeutet, dass sich dort unbekannte, noch nicht annotierte Transkripte befinden könnten. Ob es sich bei diesen Regionen tatsächlich um noch unbekannte Transkripte handelt, müsste noch mit Hilfe von Programmen zur Auffindung von Genen überprüft werden, indem in diesen Regionen nach Open-Reading-Frames und Splice-Sites gesucht wird.
In Abbildung 4.1 ist ein intergenisches Cluster dargestellt. Dieses signifikante Cluster von Solexa-Elementen auf Chromosom 5 im Humangenom befindet sich mehrere Kilo-

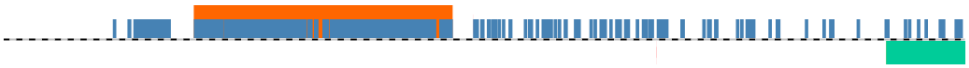

Abbildung 4.1: Cluster (orange) von Solexa-Tags (blau) aus der Niere in einem intergenischen Bereich von Chromosom 5 (bp 40861800-40863813) auf dem Vorwärtsstrang (oben). Downstream ist in türkis auf dem Gegenstrang im Abstand von 3374 bps eine annotierte UTR zu sehen.

basen von dem nächst näheren Gen entfernt. Die Tatsache, dass 323 *expressed sequence tags* [4] mit dem Cluster überlappen (nicht in der Abbildung dargestellt), ist ein sicherer Hinweis, dass das Cluster in einer codierenden Region auf dem Genom liegt, die bislang noch nicht als Transkript annotiert wurde. Bei dem Cluster selbst könnte es sich um ein Exon handeln.

4.1.2 Identifizieren von potentiellen Transkriptionstart-Regionen mit Hilfe der Cluster-Analyse

Transkriptionstart-Regionen werden bei *Genomatix* mit Hilfe von CAGE-Tags bestimmt. Befindet sich in einem 10-Basenpaar-Fenster auf der DNA eine ausreichende Anzahl von CAGE-Tags, so wird diese Region als mögliche Transkriptionstart-Region definiert. Anschließend wird versucht diese TSR zu erweitern, indem geprüft wird, ob sich in einem Abstand von 40 bp oder weniger weitere CAGE-Tags befinden.
Zum Vergleich sollte ermittelt werden, ob mit dem r-Scan-Algorithmus die selben Regionen auf dem Genom als TSRs identifiziert werden. Dazu wurde für jedes Chromosom des Humangenoms auf dem Vorwärtsstrang nach signifikanten Anhäufungen von CAGE-Tags gesucht. Für r wurden in dieser Analyse die Werte 3, 4 und 5 verwendet. Die Anzahl der Iterationen betrug 20, die Irrtumswahrscheinlichkeit 1%. In jeder Iteration wurden die 3 kleinsten r-Fragmente betrachtet. Die so gefundenen CAGE-Tag-Cluster wurden dann darauf überprüft, ob und mit wie vielen annotierten TSRs sie überlappen. Tabelle 4.1 zeigt die Anzahl der mit den einzelnen Scans gefundenen Cluster sowie den Anteil der Cluster, die mit genau einer, mehr als einer oder keiner TSR überlappen.

	Cluster	Davon überlappen mit 1 TSR	> 1 TSR	keiner TSR
3-Scan	895	889	2	4
4-Scan	888	859	27	2
5-Scan	882	699	181	2

Tabelle 4.1: Vergleich der mit 3-, 4- und 5-Scans gefundenen Cluster

Für jeden Scan wurden aus der Liste gefundener Cluster Stichproben genommen, um zu überprüfen für welchen Wert von r die höchste Übereinstimmung mit TSRs erzielt werden konnte. Es zeigte sich, dass die Übereinstimmung mit bekannten TSRs am größten ist, wenn $r = 4$ gewählt wurde. In diesem Fall entspricht nämlich der Schwellenwert für den Abstand, den ein CAGE-Tag zu einem Cluster haben darf um dazugenommen zu werden[1], am ehesten den oben erwähnten 40 bp. Deshalb ist die Mehrheit aus der Stichprobe der durch 4-Scans identifizierten Cluster deckungsgleich mit bereits annotierten TSRs (siehe zum Beispiel Abbildung 4.2), einige decken allerdings auch mehrere TSRs ab, besonders auf dem kürzesten menschlichen Chromosom, dem y-Chromosom. Dies liegt daran, dass der Betrag des Schwellenwerts von der Anzahl der Elemente auf dem betrachteten Chromosom abhängt und auf dem y-Chromosom nur wenige CAGE-Tags annotiert wurden. Die 3-Scans identifizierten fast ausnahmslos Cluster, die innerhalb einer TSR liegen (siehe Tabelle 4.1). Die Untersuchung der Cluster aus der Stichprobe zeigte allerdings, dass die mit 3-Scans gefundenen Cluster die TSRs meist nicht in ihrer vollen Länge abdecken, da mit den 3-Scans nur sehr dicht geclusterte Anhäufungen von CAGE-Tags detektiert wurden. Die mit den 5-Scans gefundenen Häufungen sind weniger dicht geclustert als in den beiden anderen Analysen, da die Größe des Schwellenwerts mit zunehmendem r ebenfalls zunimmt. Deshalb tritt bei den 5-Scans häufig der Fall auf, dass ein Cluster mehr als eine TSR abdeckt (siehe Tabelle 4.1).
Interessant ist, dass in allen drei Analysen auch einige wenige Cluster gefunden wurden, die nicht mit einer bekannten TSR überlappen (siehe Tabelle 4.1). Hier könnte es sich um noch nicht annotierte TSRs handeln.

[1] siehe Abschnitt 3.1

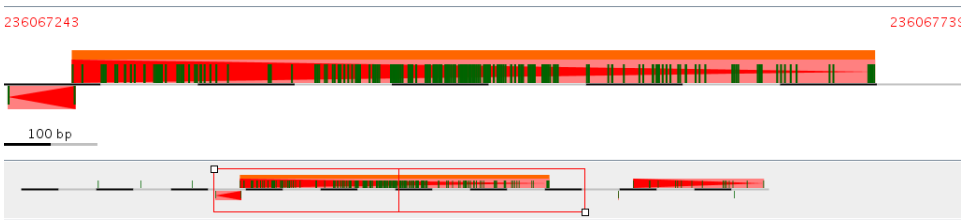

Abbildung 4.2: Mit 4-Scan identifiziertes Cluster (orange) von CAGE-Tags (dunkelgrün) auf dem Vorwärtsstrang von Chromosom 2. In rot ist eine TSR zu sehen, die deckungsgleich mit dem Cluster ist..

4.2 Bindungspräferenzen von TF-Sites innerhalb von Promotoren

4.2.1 Analyse mit allen Promotoren

Für alle 153 Familien von menschlichen Transkriptionsfaktoren, die in der *Genomatix*-Datenbank abgelegt sind, habe ich die Distanzkorrelationen zu den 3'-Enden von Promotoren berechnet. Dabei fiel auf, dass es im Wesentlichen 5 verschiedene Arten von Kurven gibt, die in leichten Variationen immer wieder auftreten. Die einzelnen Typen sind im folgenden mit Beispielen aufgelistet und näher beschrieben.

Typ 1 Es kann keine Tendenz festgestellt werden, dass Distanzen in einem bestimmten Bereich verstärkt auftreten, die Kurve ist stark verrauscht und nähert sich einer Gleichverteilung an. Dies ist beispielsweise bei PAX4 der Fall.

Typ 2 Man sieht eine Überrepräsentation der Bindungswahrscheinlichkeit im proximalen Promotorbereich und eine Unterrepräsentation im Bereich der 5'-UTR. In Richtung des distalen Promotorbereichs nimmt die Bindungshäufigkeit ab. Bei Kurven von diesem Typ sind somit deutlich Bereiche im Promotor erkennbar, an die besonders oft gebunden wird. Ein Beispiel für eine solche Verteilung ist die des Transkriptionsfaktors SP1F.

Typ 3 Die Bindungswahrscheinlichkeit nimmt in Richtung des distalen Promotorbereichs zu. So konnte es etwa bei MYT1 beobachtet werden.

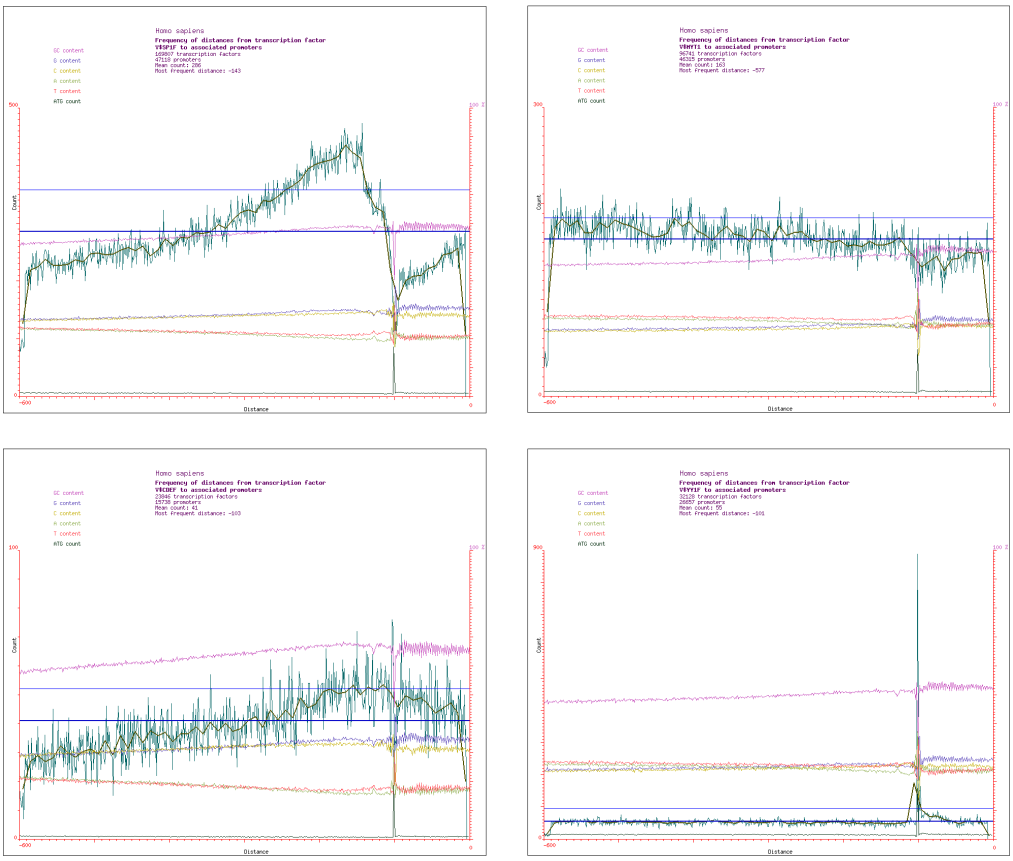

Abbildung 4.3: Übersicht über die verschiedenen Typen von Kurven bei Distanzkorrelationen zwischen TF-Sites und Promotoren
Links oben: SP1F (Typ 2), rechts oben: MYT1 (Typ 3), links unten: CDEF (Typ 4), rechts unten: YY1F (Typ 5)

Typ 4 Es lässt sich eine Überrepräsentation der Bindungshäufigkeiten im proximalen Promotorbereich und in der Nähe der UTR beobachten, während im distalen Bereich eine Unterrepräsentation zu sehen ist. Auch bei Kurven von Typ 4 sind also Bereiche im Promotor zu erkennen, an die häufig gebunden wird, jedoch nicht so deutlich wie bei Typ 2. CDEF ist ein Beispiel für diesen Fall.

Typ 5 Beim letzten Typ ist die Kurve an sich einem der oben beschriebenen Typen zuzuordnen, aber ein paar wenige Peaks in der Nähe der Transkriptionstart-Site überragen die Häufigkeiten der anderen Distanzen bei weitem. Dies ist beispielsweise bei YY1F der Fall. Der Graph beschreibt im wesentlichen eine von Rauschen überlagerte Gleichverteilung, der Abstand von 101 bp liefert aber einen überdeutlichen Peak, der die Häufigkeiten der anderen Distanzen um ein vielfaches überragt.

In Abbildung 4.3 sind für die Typen 2-5 exemplarisch die Kurven von den vier Transkriptionsfaktoren SP1F, MYT1, CDEF und YY1F dargestellt.

Tabelle 4.2 zeigt, wie oft die einzelnen Typen von Kurven beobachtet werden konnten. Obwohl es nicht immer eindeutig bestimmt werden konnte, da vereinzelt auch Mischformen auftraten, wurde jede Kurve genau einem Typ zugeordnet. Zusätzlich sind die relativen Häufigkeiten des Vorkommens der verschiedenen Kurventypen in Abbildung 4.4 in einem Balkendiagramm dargestellt.

Typ 1	43
Typ 2	17
Typ 3	24
Typ 4	14
Typ 5	55
Total	153

Tabelle 4.2: Häufigkeit des Vorkommens der einzelnen Typen von Kurven

Um sicherzustellen, dass die Bindungswahrscheinlichkeit der Transkriptionsfaktoren tatsächlich von bestimmten Abständen zur Transkriptionstart-Site abhängt, und nicht von anderen Einflüssen, habe ich zudem die Verteilung des GC-Gehalts in den Promotoren graphisch dargestellt. Es ist bekannt, dass der GC-Gehalt in regulatorischen Bereichen wie Promotoren höher ist als anderswo im Genom [14, 25]. Bei einem Zusammenhang zwischen Bindungswahrscheinlichkeit eines Transkriptionsfaktors und dem

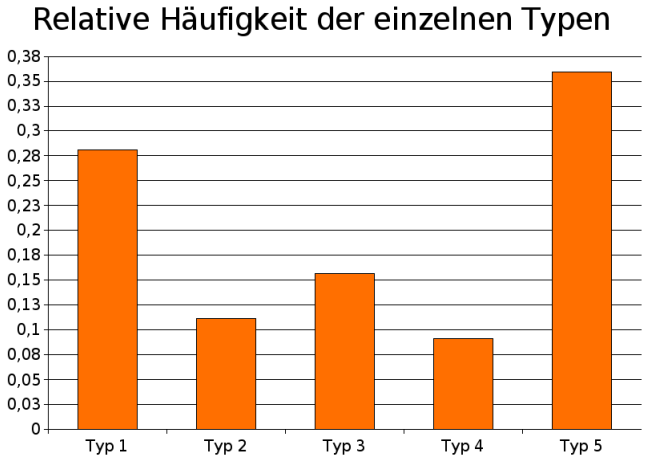

Abbildung 4.4: Relative Häufigkeit der einzelnen Kurventypen

GC-Gehalt in den Promotoren hätte man erwartet, dass die Verteilung der Distanzhäufigkeiten stark mit der Kurve des GC-Gehalts korreliert, was bedeuten würde, dass dort, wo der GC-Gehalt am größten ist, der Faktor auch am wahrscheinlichsten bindet. Besonders bei dem GC-Box-Faktor SP1F wäre ein starker Zusammenhang zwischen Bindungswahrscheinlichkeit und GC-Gehalt zu erwarten gewesen, da der am besten konservierte Bereich des Bindungsmotivs dieses Faktors nur Gs und Cs beinhaltet (siehe Abbildung 4.5). Allerdings konnte diese Vermutung nicht bestätigt werden. Zwar gibt es auch bei der Kurve des GC-Gehalts ein Minimum, das im Bereich der Transkriptionstart-Site liegt (siehe Abbildung 4.3), und die Kurve steigt danach wieder an bis ein Maximum erreicht wurde, woraufhin sie wieder abfällt, allerdings gibt es bezüglich der Steigung sehr große Unterschiede zwischen der Kurve des GC-Gehalts und der der Distanzhäufigkeiten. Interessant ist hingegen, dass bei allen Transkriptionsfaktoren die GC-Gehalt-Kurve der Promotoren einen sehr ähnlichen Verlauf hat. Im Bereich zwischen der Transkriptionstart-Site und dem Promotorende ist der GC-Gehalt großen Schwankungen ausgesetzt, weil hier bei manchen Promotoren schon der codierende Bereich beginnt und bestimmte Codons bevorzugt verwendet werden. Im Bereich der Transkriptionstart-Site gibt es einen starken und steilen Abfall der Kurve, die sofort danach wieder stark ansteigt. Kurz danach gibt es einen weiteren Einbruch in der Kurve, der allerdings bei weitem nicht so stark und auch viel weniger steil ist. Ab diesem Punkt nimmt der GC-Gehalt in Rich-

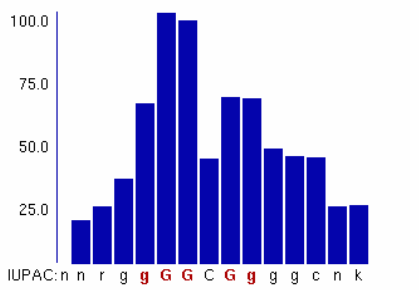

Abbildung 4.5: Graphische Darstellung des Profils der Positions-Gewichts-Matrix von SP1F

tung des distalen Promotorbereichs langsam aber stetig ab.
Da die Kurven des A-, G- und T-Gehalts an derselben Stelle einen deutlichen Peak zeigen, an der die GC-Gehalt- und die C-Gehalt-Kurve ihr jeweils größtes Minimum haben, wurde dieses Phänomen näher untersucht. Es stellte sich heraus, dass der Grund hierfür die Tatsache ist, dass das Startcodon ATG an dieser Stelle überdurchschnittlich oft vorkommt, wie in der entsprechenden Kurve in den Graphiken zu sehen ist. Der Grund für das häufige Auftreten des Startcodons an dieser Position ist die Tatsache, dass viele Transkripte, die noch nicht vollständig annotiert sind, keine 5'-UTR haben und direkt mit dem Startcodon beginnen. Insgesamt beginnen mehr als 15% der menschlichen Transkripte in den Datenbanken von *Genomatix* direkt mit dem Startcodon, wovon fast 80% Bronze-Transkripte sind. Dadurch konnte auch erklärt werden, wieso manche Transkriptionsfaktoren überdurchschnittlich hohe Peaks an dieser Stelle liefern. Bei diesen Transkriptionsfaktoren enthält nämlich das bestkonservierte Motiv die Sequenz ATG oder die zu dieser revers komplementäre Sequenz CAT. Diese Theorie konnte am Beispiel des YY1F bestätigt werden, dessen Kernmotiv die Sequenz CAT beinhaltet (siehe Abbildung 4.6). Wenn man hier nur die Distanzen zu den Promotoren darstellt, bei denen der Faktor an denselben Strang bindet, auf dem auch der Promotor liegt, so ist der Peak bei weitem nicht so hoch wie im gegenteiligen Fall, bei dem die Promotoren betrachtet werden, bei denen der Faktor nicht an den Strang bindet, auf dem der Promotor liegt, sondern an den Gegenstrang. Im letzteren Fall handelt es sich vermutlich um falsch positive Bindungsstellen, die mit dem Startcodon zusammenfallen, das tatsächlich an den berechneten Stellen liegt.

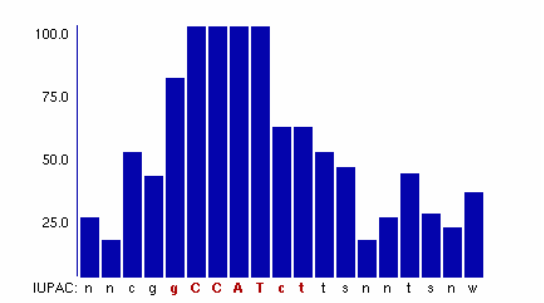

Abbildung 4.6: Graphische Darstellung des Profils der Positions-Gewichts-Matrix von YY1F

Zusammenfassend kann man sagen, dass nur bei den Kurven von Typ 2 oder 4 Bereiche im Promotor zu erkennen sind, in denen sich besonders häufig TF-Sites befinden, wobei diese Bereiche in den Kurven von Typ 2 wesentlich deutlicher zu sehen sind als in den Kurven von Typ 4. Dies legt nahe, dass in den Promotoren, in denen sich in diesen Intervallen TF-Sites befinden, der zugehörige Transkriptionsfaktor regulierende Wirkung hat. In den Kurven von Typ 2 sind diese Bereiche wahrscheinlich deshalb besonders deutlich zu erkennen, weil überproportional viele Transkripte von dem entsprechenden Transkriptionsfaktor reguliert werden.

4.2.2 Analyse mit Promotoren vollständig annotierter Transkripte

Auch die Analyse der Distanzkorrelationen von TF-Sites zu den 3'-Enden von hypothetischen Promotoren um CAGE-Tag-Cluster wurde für alle 153 Familien von in der Datenbank von *Genomatix* hinterlegten menschlichen Transkriptionsfaktoren durchgeführt. Bei der Interpretation der Graphen, die das Ergebnis der Analyse bilden, konnten sofort einige wesentliche Unterschiede zur vorherigen Analyse ausgemacht werden, in die noch alle Promotoren eingingen (siehe Abbildung 4.7).

Dass für diese Analyse fast ausnahmslos Promotoren betrachtet wurden, deren Transkripte vollständig annotiert sind, konnte durch die Graphen bestätigt werden. Da kein Peak in der Startcodon-Kurve beobachtet wird, darf davon ausgegangen werden, dass der Großteil der hinter den betrachteten Promotoren liegenden Transkripte eine UTR besitzt. Dies zeigt sich auch in der GC-Gehalt-Kurve, die nun im Bereich der

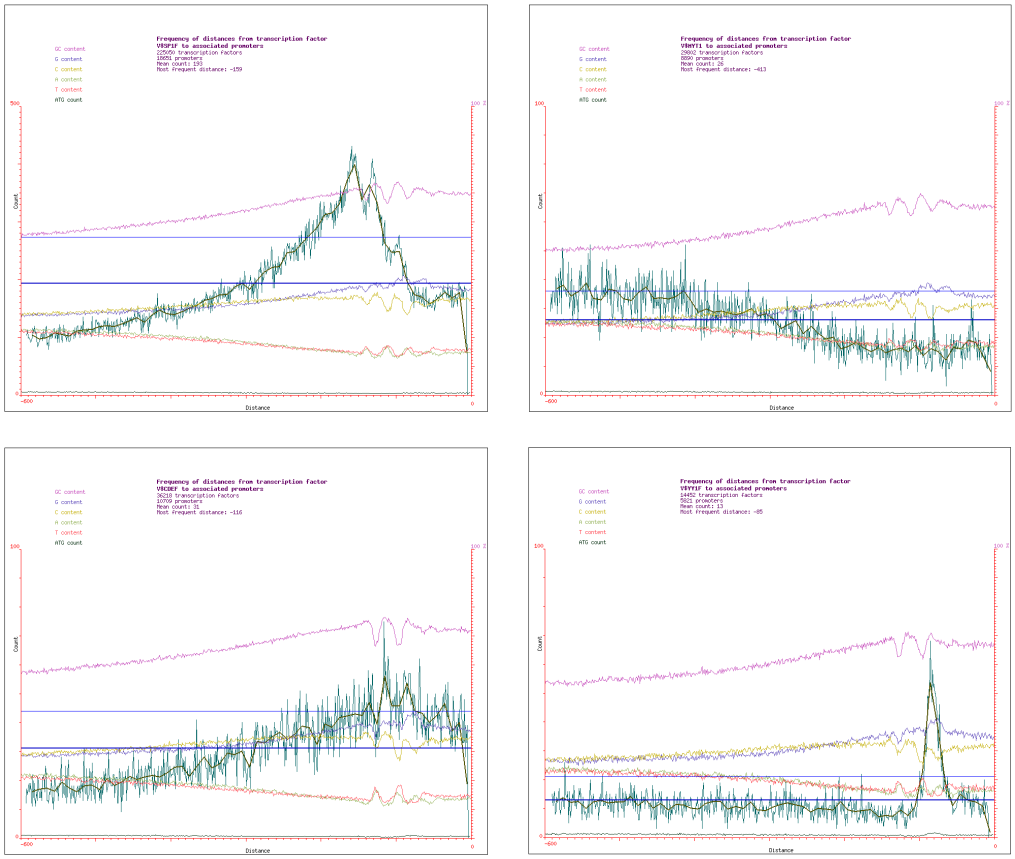

Abbildung 4.7: Distanzkorrelationen zwischen verschiedenen TF-Sites und Gold-Promotoren
Links oben: SP1F, rechts oben: MYT1, links unten: CDEF, rechts unten: YY1F

Transkriptionstart-Site keinen scharfen Abfall nach unten mehr aufweist. Zudem ist der durchschnittliche GC-Gehalt in den Promotoren höher als in der vorherigen Analyse. Auch die übermäßig hohen Peaks bei den Distanzhäufigkeiten sind nun nicht mehr zu sehen. Somit machen die Kurven nun biologisch gesehen insgesamt mehr Sinn und lassen sich auch besser interpretieren. In Abbildung 4.7 sind zum Vergleich nochmals die Kurven für die vier Transkriptionsfaktoren SP1F, MYT1, CDEF und YY1F zu sehen. Alle Kurven zeigen, verglichen mit den Kurven in Abbildung 4.3, ein deutlicheres Bild. Ausgehend vom distalen Promotorbereich ist ein Anstieg des GC-Gehalts in Richtung der Transkriptionstart-Site zu beobachten, während der GC-Gehalt im Transkript wieder abnimmt. Das Phänomen, dass der durchschnittliche GC-Gehalt ansteigt, bis ein Maximum erreicht ist, je mehr man sich der Transkriptionstart-Site nähert, wurde auch schon in [25] beschrieben. Darüber hinaus sind in unmittelbarer Nähe dieses Maximums auch deutlich zwei Bereiche zu erkennen, in denen der GC-Gehalt signifikant niedriger ist als in der Umgebung dieser Bereiche. Zumindest der erste Abfall des GC-Gehalts, der im Bereich von ca. 30 bp upstream vom Beginn des Transkripts zu finden ist, lässt sich biologisch interpretieren. Hier bindet bei sehr vielen Promotoren ein Transkriptionsfaktor, der unter dem Namen TATA-Box (TBPF) bekannt ist und der, wie der Name schon sagt, an ein AT-reiches Motiv im Promotor bindet. In Abbildung 4.8 ist der entsprechende Kurvenverlauf zu sehen.
Darüber hinaus treten nun bei den Transkriptionsfaktoren, die bestimmte Bereiche zur Bindung bevorzugen, diese Bereiche deutlicher hervor als in der ersten Analyse. Ein Beispiel hierfür ist der GC-reiche Faktor SP1F, der jetzt auch eine stärkere Korrelation mit der GC-Kurve zeigt (siehe Abbildung 4.7). Im Gegensatz sind nun aber auch weit mehr Kurven verrauscht und somit nicht eindeutig interpretierbar, da die falsch-positiven Peaks nicht mehr auftreten.

4.2.3 Gewebespezifische Analyse

Von einigen Transkriptionsfaktoren ist bekannt, dass sie bei der Bindung an den Promotor einen bevorzugten Abstand zur Transkriptionstart-Site zeigen, zum Beispiel HNF1, E2FF und CREB [24, 20, 13, 10]. Bei E2FF und CREB konnte dies in den beiden oben beschriebenen Analysen bestätigt werden[2], bei HNF1 jedoch nicht. In [24] wird ein An-

[2]Beide Transkriptionsfaktoren zeigen eine erhöhte Bindungspräferenz zwischen der TSS und der Position 100 bps upstream von der TSS

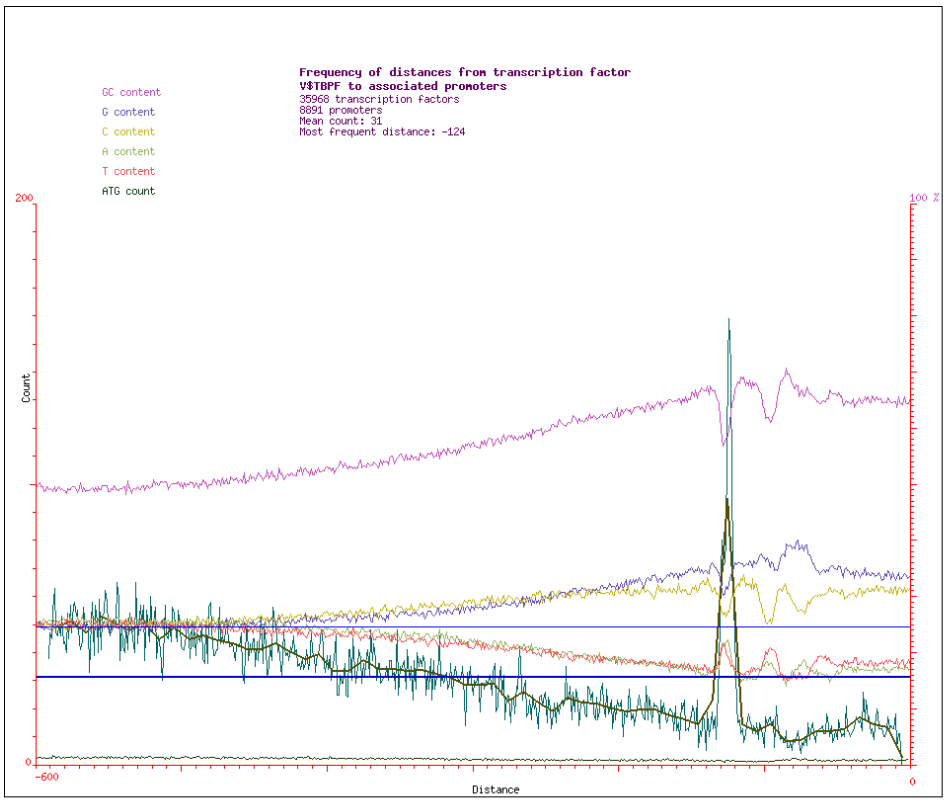

Abbildung 4.8: Distanzkorrelationen zwischen den Bindungsstellen der TATA-Box und CAGE-Tag-Clustern
Die Abnahme des GC-Gehalts in dem Bereich, in dem am häufigsten Bindungsstellen der TATA-Box vorhergesagt wurden, ist deutlich zu erkennen.

satz beschrieben, bei dem kleine Teilbereiche von Vertebraten-Genonem nach potentiellen Bindungsstellen von HNF1 durchsucht wurden. Dabei wurde festgestellt, dass HNF1-Sites besonders häufig in den Promotoren von Genen vorkommen, von denen bekannt ist, dass sie in der Leber exprimiert werden. Ferner wurde gezeigt, dass der Transkriptionsfaktor nur bei Leber-spezifischen Genen Bindungspräferenzen im Promotor zeigt. Deshalb wurde in dieser Diplomarbeit in einer Hochdurchsatzanalyse die Verteilung von HNF1-Sites in den Promotoren Leber-spezifischer Gene im Humangenom untersucht.
Dazu wurde eine Liste von Transkripten erstellt, von denen es Hinweise in Form von CAGE-Tags [12] gibt, dass sie in der Leber exprimiert werden. Aus dieser Liste wurden dann alle Transkripte entfernt, die auch in anderen Geweben exprimiert werden. Zum Vergleich wurden außerdem mehrere Listen derselben Größe mit zufällig aus dem Humangenom gewählten Transkripten gebildet. Mit *GenomeInspector* wurden alle Listen mit HNF1-Sites korreliert. Einerseits wurde die Site auf einen Punkt reduziert, andererseits wurde sie in ihrer vollen Ausdehung betrachtet. Der Bezugspunkt war in diesem Fall das 5'-Ende der Transkripte. Außerdem wurde die Größe des Sliding-Windows variiert.
Bei der Korrelationsanalyse der HNF1-Sites mit Transkripten zeigte sich, dass HNF1-Sites in Leber-assoziierten Genen an bestimmten Stellen im Promotor besonders häufig binden. Diese Tendenz ist nicht erkennbar, wenn man alle Transkripte betrachtet oder zufällig ausgewählte (siehe Abbildung 4.9). Es ist zu beachten, dass für die Analyse nur HNF1-Sites verwendet wurden, die innerhalb von Promotoren liegen. Da Promotoren sich in 3'-Richtung nur bis zu der Position 100 bps downstream vom 5'-Ende der Transkripte erstrecken, handelt es sich bei allen TF-Sites, die mehr als 100 bps downstream vom Anfang der Transkripte liegen, um statistisches Rauschen. Deshalb gingen in dieser Analyse alle Distanzen > 100 nicht in die Berechnung des Mittelwerts und der Standardabweichung der Distanzhäufigkeiten mit ein.
Im Fall der Leber-assoziierten Transkripte ist eine erhöhte Bindungshäufigkeit im proximalen Bereich der Position 70 bp upstream vom 5'-Ende der Transkripte auszumachen. Auch bei 400 bis 500 bps upstream vom 5'-Ende kann eine Überrepräsentation der Bindungswahrscheinlichkeit von HNF1 beobachtet werden, jedoch nicht so deutlich wie in der Nähe der Transkriptionstart-Site. Wenn zufällig ausgewählte oder alle Transkripte mit HNF1-Sites korreliert wurden, sind dagegen in den Kurven keine signifikanten Peaks zwischen 100 und 50 bps upstream vom 5'-Ende der Transkripte erkennbar. Diese Beobachtungen decken sich mit den Erwartungen [24].

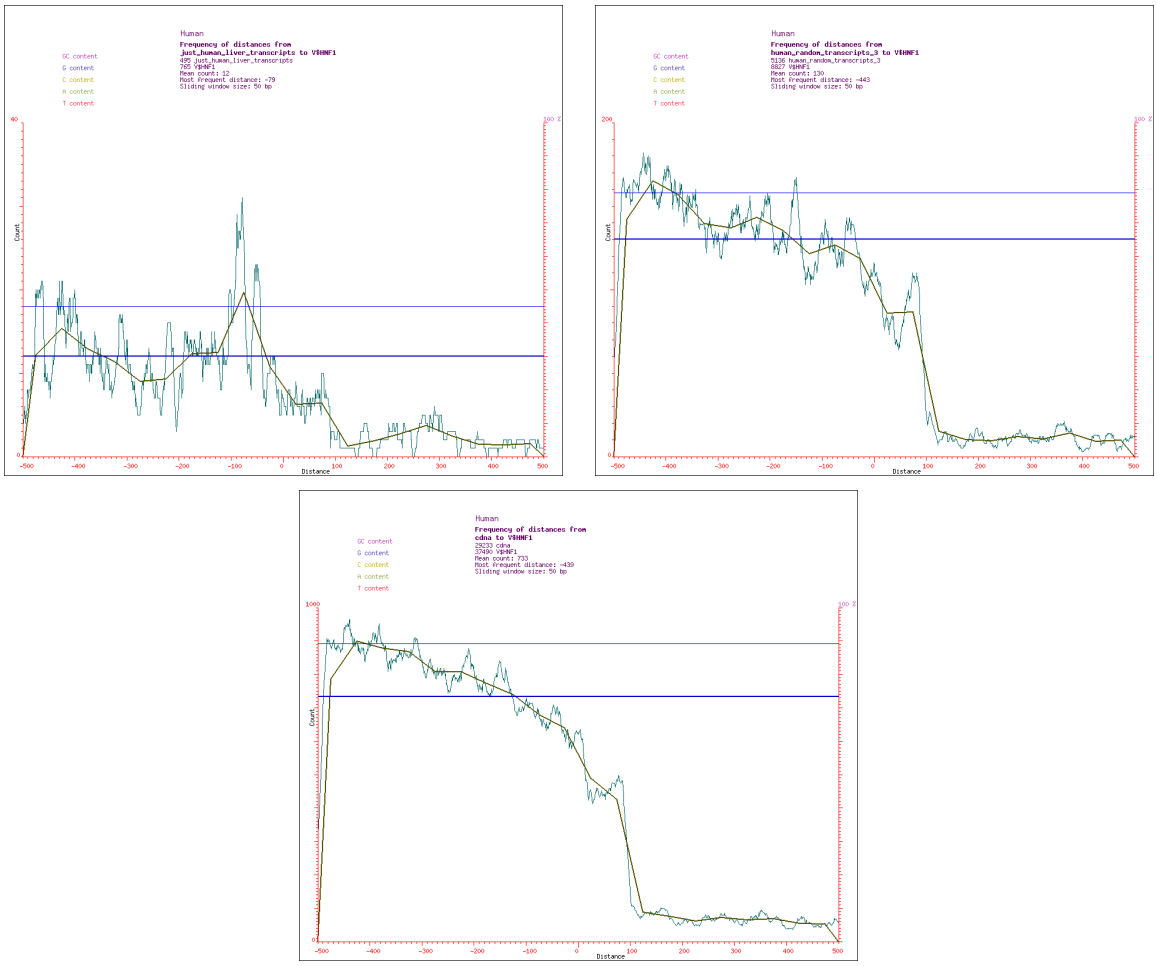

Abbildung 4.9: Distanzkorrelationen zwischen Transkripten und HNF1-Sites
Oben links: nur Transkripte, von denen es Hinweise in Form von CAGE-Tags gibt, dass sie ausschließlich in der Leber exprimiert werden. Oben rechts: zufällig ausgewählte Transkripte. Unten: alle Transkripte.
Das Sliding-Window hat bei allen Graphen eine Größe von 50 bp. Die HNF1-Sites wurden in ihrer vollen Ausdehnung betrachtet.

Jedoch ist bei der Auswertung der entsprechenden Graphen Vorsicht geboten. Zwar sind die Peaks in dem Graphen, für den nur Leber-assoziierte Gene betrachtet wurden, deutlich zu sehen, in dieser Deutlichkeit jedoch nur, wenn man die TF-Site in ihrer vollen Ausdehnung betrachtet. Dazu kommt, dass die häufigste Distanz weniger als 40 mal gezählt wurde. Andererseits ist derzeit noch von wenigen Transkripten bekannt, ob sie Gewebe-spezifisch exprimiert werden.

4.3 Distanzkorrelationen mit konservierten Regionen

Um festzustellen, welche Bereiche im Humangenom besonders stark Spezies-übergreifend konserviert sind, wurden die Distanzen von konservierten Regionen zu den Enden diverser genomischer Elemente bezüglich ihrer Häufigkeit analysiert. Dabei wurde von jedem Nukleotid der konservierten Region der Abstand zum jeweiligen Bezugspunkt gezählt.

4.3.1 Abstände zum 5'-Ende von Transkripten und microRNAs

Für diese Analyse wurden jeweils die 5'-Enden von Transkripten und microRNAs mit unterschiedlich stark konservierten Regionen korreliert. Der minimale Grad der Konservierung betrug dabei 80, 85, 90, 95 und 100%.
Bei den Distanzkorrelationen von konservierten Regionen zu Transkripten konnte festgestellt werden, dass sich die Abstände im proximalen Bereich des 5'-Endes des Transkriptes signifikant häufen. Je höher der Grad der Ähnlichkeit einer Sequenz zwischen zwei Spezies ist, desto näher befindet sich der Bereich höchster Übereinstimmung zwischen konservierter Region und Transkript an der TSS. Des weiteren zeigte sich, dass auch solche Bereiche des Genoms, die in microRNAs übersetzt werden, hochkonserviert sind. Die Distanzkorrelationen von Transkripten beziehungsweise microRNAs zu konservierten Regionen sind in Abbildung 4.10 zu sehen.

4.3.2 Abstände zum 5'-Ende von Exons

Da in Transkripten vor allem die Bereiche um das 5'-Ende als hochkonserviert identifiziert worden waren, habe ich auch die Distanzen der ersten Exons von Genen zu konservierten Regionen mit einem Konservierungsgrad von 100% korreliert. Da Single-Exon-Gene häufig für regulierende Elemente wie microRNAs codieren, die oft stark

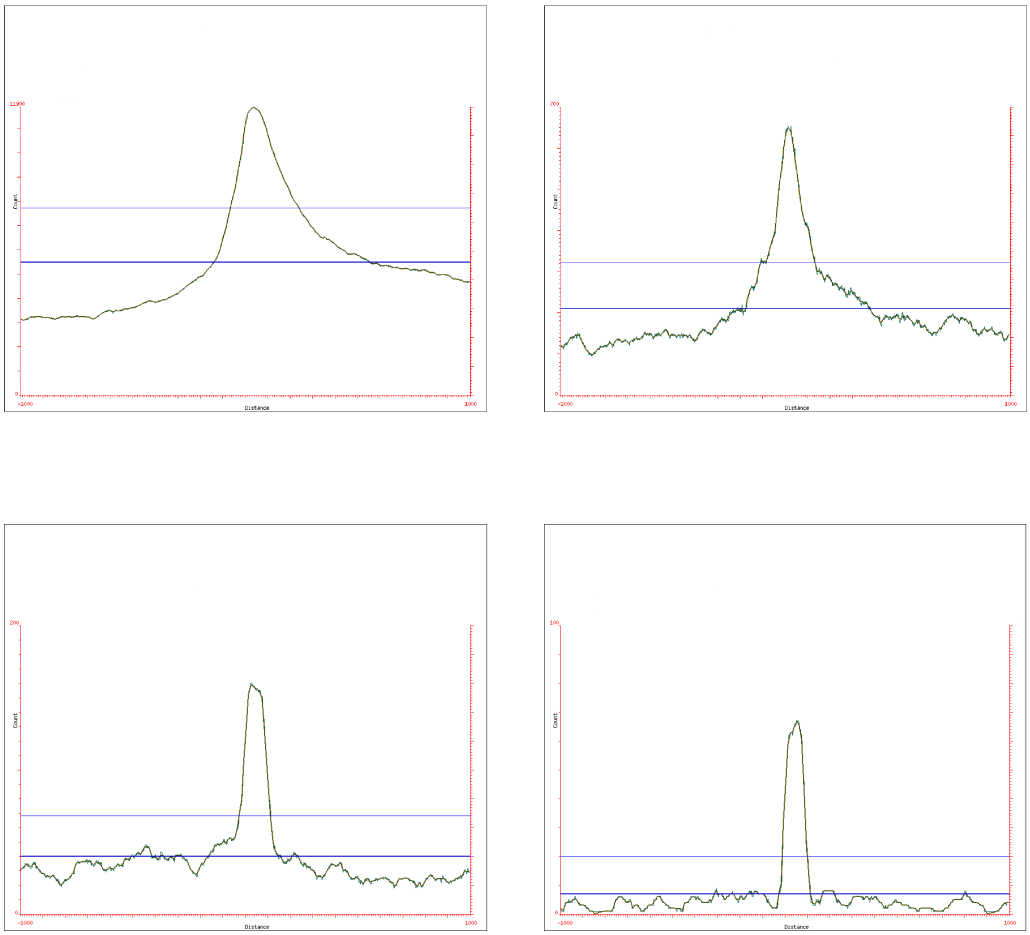

Abbildung 4.10: Distanzkorrelationen der 5'-Enden von Transkripten und microRNAs zu konservierten Regionen
Oben: Transkripte, unten: microRNAs, links: Konservierungsgrad von mindestens 80%, rechts: Konservervierungsgrad von 100%

konserviert sind, war davon auszugehen, dass auch kurze Transkripte ohne Introns generell einen hohen Konservierungsgrad aufweisen. Deshalb wurden die ersten Exons von Multi-Exon-Genen und das einzige Exon von Single-Exon-Genen jeweils separat betrachtet (siehe Abbildung 4.11). Zum Vergleich wurden auch die Abstände der inneren Exons von Multi-Exon-Genen zu konservierten Regionen berechnet.

Bei der Analyse, bei der Abstände vom 5'-Ende des ersten Exons eines Gens zu konser-

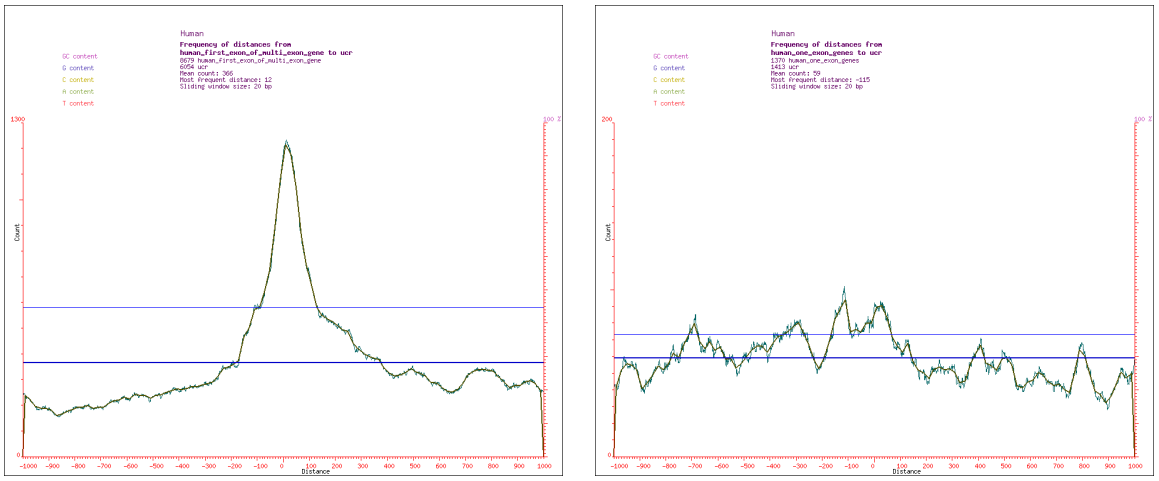

Abbildung 4.11: Distanzkorrelationen zwischen dem 5'-Ende des jeweils ersten Exon eines Gens und konservierten Regionen
Links: erste Exons aus Multi-Exon-Genen, rechts: Single-Exon-Gene.

vierten Regionen betrachtet wurden, zeigte sich, dass vor allem der Bereich um das erste Exon von Multi-Exon-Genen konserviert ist. Diese Tendenz ist bei Single-Exon-Genen bei weitem nicht so deutlich zu erkennen. Die entsprechenden Kurven sind in Abbildung 4.11 einander gegenüber gestellt.

Wurden die 5'-Enden von Exons, die im Inneren eines Gens liegen, mit konservierten Regionen korreliert, so zeigte sich, dass auch diese Exons mit konservierten Regionen korrelieren. Allerdings sind hier die Grenzen zwischen Exon und Umgebung noch viel schärfer, was die Konservierung betrifft, als bei den ersten Exons.

Die Beobachtung, dass die ersten Exons von Multi-Exon-Genen stärker konserviert sind als Single-Exon-Gene, stimmte nicht mit der Erwartung überein, dass vor allem Single-Exon-Gene stark konserviert sind. Dies konnte allerdings durch den entsprechenden Graphen weder bestätigt noch widerlegt werden, da hier lediglich nach konservierten Regionen gesucht, die in einem bestimmten Abstand zum 5'-Ende der Transkripte lie-

gen. Obwohl eine Abstandsspezifität zum 5'-Ende nicht bestätigt werden konnte, ist es durchaus möglich, dass Single-Exon-Gene häufig stark konserviert sind, da sowohl die Exons selbst als auch die konservierten Regionen unterschiedliche Ausdehnungen haben. Dass Exons im allgemeinen hoch konserviert sind, ist nicht überraschend, da es sich hier um codierende Bereiche im Genom handelt. Entsprechend ist auch nicht verwunderlich, dass die Peaks von inneren Exons sehr steil sind, da Introns stärker Mutationen ausgesetzt sind.
Dass bei den ersten Exons der Peak viel breiter ist, ist wahrscheinlich darauf zurückzuführen, dass hier zusätzlich bestimmte Motive mit regulierender Funktion wie die TATA-Box oder auch das Startcodon stark konserviert sind. Wie in Abschnitt 4.2.2 beschrieben hat auch die TATA-Box eine hohe Abstandsspezifität.

4.3.3 Abstände zum 3'-Ende von Transkripten und Exons

Wenn das 3'-Ende von Transkripten oder Exons mit zu 100% konservierten Regionen korreliert wurde, konnte in allen Fällen ein hoher Grad der Konservierung in diesem Bereich festgestellt werden. Abbildung 4.12 zeigt die Distanzkorrelationen zwischen den ersten Exons von Multi-Exon-Genen und Single-Exon-Genen zu konservierten Regionen. Die Beobachtung, dass die 3'-UTR von Transkripten stark konserviert ist, deckt sich mit

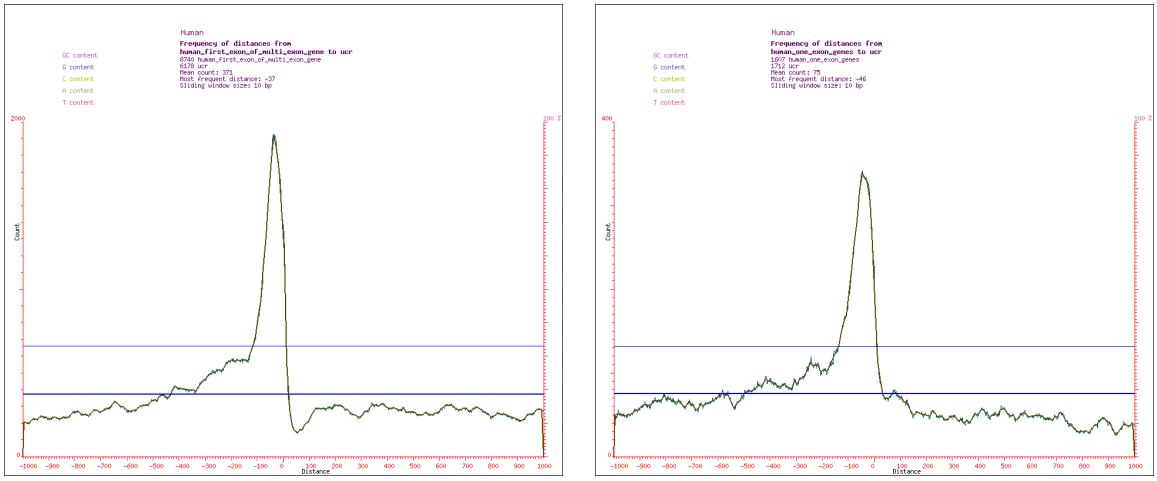

Abbildung 4.12: Distanzkorrelationen zwischen dem 3'-Ende des jeweils ersten Exon eines Gens und konservierten Regionen
Links: erste Exons aus Multi-Exon-Genen, rechts: Single-Exon-Gene

dem, was in der Literatur [11] beschrieben wird. Die 3'-Enden von Exons sind vermutlich deshalb Spezies-übergreifend konserviert, weil Exons einerseits kodierende Funktion haben und andererseits, damit die Information, wo das Exon endet und das Intron beginnt, erhalten bleibt. Letzteres könnte der Grund dafür sein, dass die Peaks in den Graphen in Richtung Intron sehr steil abfallen.
Beispielhaft soll dies am 3. Exon-Intron-Übergang innerhalb eines Transkripts veranschaulicht werden. In Abbildung 4.13 sind die Distanzkorrelationen der 3'-Enden der in Transkripten an dritter Stelle auftretenden Exons zu konservierten Regionen dargestellt. Um einen Anhaltspunkt zu haben, wo sich die Exon-Intron-Grenze befindet, wurde für diese Analyse auch der Gehalt an einzelnen Nukleotiden in den betrachteten Sequenzen berechnet und dargestellt. Ganz in der Nähe des 3'-Endes sind sehr scharfe signifikante Peaks in den G- und T-Gehalt-Kurven zu erkennen. Dies weist auf den Exon-Intron-Übergang hin, da Introns immer mit dem Motiv 'GT' beginnen. Downstream von diesen Peaks fällt auch die Kurve der Distanzhäufigkeiten ab, was darauf schließen lässt, dass Introns nur sehr schwach konserviert sind.

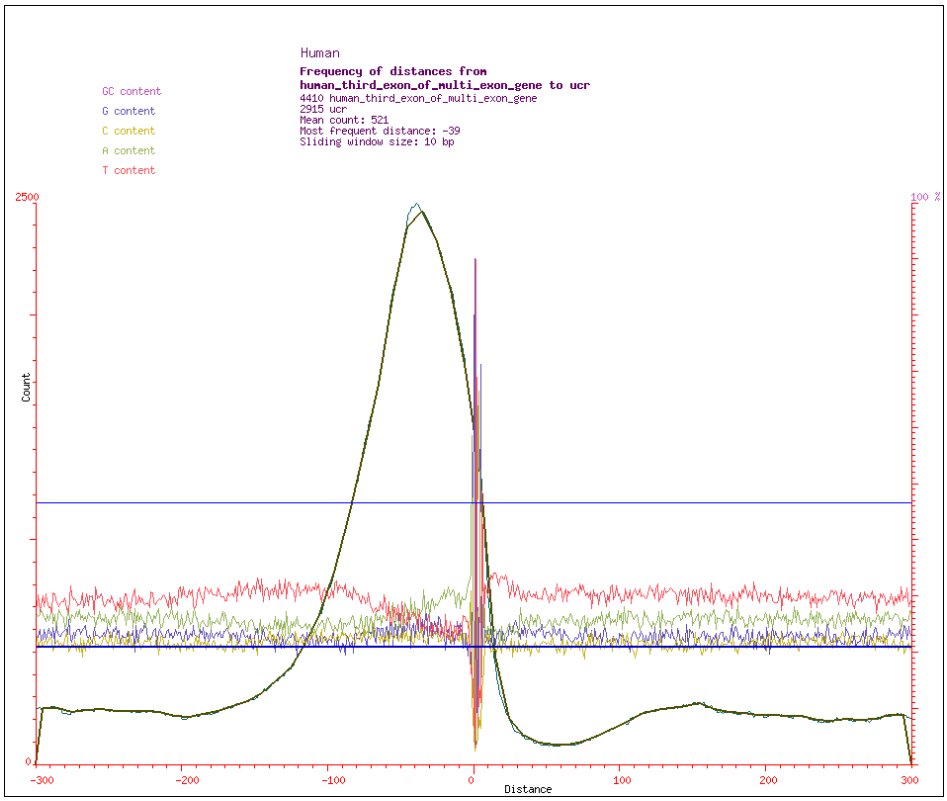

Abbildung 4.13: Distanzkorrelationen zwischen dem 3'-Ende des jeweils dritten Exons eines Gens und konservierten Regionen

5 Diskussion

5.1 Was in dieser Arbeit erreicht wurde

In dieser Arbeit wurde eine Software namens *GenomeInspector* zur Hochdurchsatzanalyse von Distanzkorrelationen genomischer Elemente entwickelt. Benutzer können über eine Weboberfläche das Programm parametrieren und Distanzkorrelationen zwischen beliebigen Datensätzen berechnen lassen, die graphisch dargestellt werden. Aus dem Ergebnisgraphen können Elemente von Interesse extrahiert werden. *GenomeInspector* konnte von mir selbst an diversen Beispielen getestet werden und befindet sich derzeit bei der Firma *Genomatix* in der Testphase.
Des weiteren wurde ein Algorithmus entwickelt, mit dem die Verteilungen genomischer Elemente untersucht werden können. Dieser Algorithmus konnte erfogreich an mehreren Datensätzen getestet werden. Schwerpunktmäßig wurde in Verteilungen von CAGE-Tags sowie Solexa-Tags nach Clustern gesucht. Auch für diesen Algorithmus wurde ein Programm mit einer webbasierten Benutzeroberfläche entwickelt.
Distanzkorrelationen wurden vor allem anhand von zwei großen Beispielen untersucht: die Verteilung von Transkriptionsfaktorbindungsstellen in Promotoren und die Distanzkorrelationen zwischen konservierten Regionen und diversen anderen genomischen Elementen.
Dabei wurde festgestellt, dass es nur für wenige Transkriptionsfaktorbindungsstellen Regionen im Promotor gibt, an denen sie deutlich gehäuft auftreten, wenn in die Analyse alle Promotoren eingingen. Die Vermutung, dass es für TF-Sites gewebespezifisch Orte im Promotor gibt, an denen sie besonders häufig auftreten, konnte am Beispiel des Leber-assoziierten Faktors HNF1 bestätigt werden.
Als untersucht wurde, wie konservierte Regionen über das Genom verteilt sind, zeigte sich, dass vor allem die Regionen in der Nähe des 5'- und 3'-Endes von Transkripten stark konserviert sind. Von Exons konnte gezeigt werden, dass sie generell stark konser-

viert sind, während Single-Exon-Gene nur im Bereich ihrer 3'-UTR eine hohen Grad der Konservierung aufweisen. Auch die meisten microRNAs sind stark konserviert.

5.2 Verbesserungsmöglichkeiten

Verbesserungsmöglichkeiten gibt es vor allem in technischer Hinsicht. So wäre es unter Umständen sinnvoll, die Software anzupassen um Distanzkorrelationen von einem festen Element zu diversen anderen in einem einzigen Graphen darzustellen. Außerdem könnte es für den Benutzer hilfreich sein, wenn in 5'- und 3'-Richtung zwei unterschiedliche maximal zulässige Distanzen eingegeben werden können.
Es ist auch denkbar, die Graphen dynamischer zu gestalten. So ist es derzeit beispielsweise nicht möglich, einen bereits gezeichneten Graphen anders zu skalieren. Um einen Graphen anders zu skalieren, muss die Distanzanalyse erneut gestartet werden. Allerdings müssten, um die Graphen dynamischer zu gestalten, große Teile des Programms neu geschrieben werden, da mit dem GD-Modul für Perl nur statische Bilder dargestellt werden können.
Zudem könnte es sinnvoll sein, bei der Korrelationsanalyse die Stranginformation genomischer Elemente zu berücksichtigen. Dann wäre es beispielsweise möglich alle Elemente von Typ 1, die auf dem Vorwärtsstrang liegen, mit allen Typ-2-Elementen zu korrelieren, die auf dem Gegenstrang liegen. Denkbar wäre auch eine Auswahlmöglichkeit für Kombinationen, bei denen nur Typ-2-Elemente betrachtet werden, die auf dem selben Strang wie das Typ-1-Element liegen, oder nur Typ-2-Elemente, die auf dem Gegenstrang liegen.

5.3 Ausblick

Den in dieser Arbeit durchgeführten Analysen waren Grenzen gesetzt. Zum Zeitpunkt des Verfassens dieser Arbeit waren einige wichtige Daten nur in geringer Anzahl vorhanden. Die Aussagekraft einer Hochdurchsatzanalyse ist aber um so besser, je mehr Daten in die Analyse eingehen.
Bislang wurden noch sehr wenige microRNAs auf dem Humangenom annotiert. Wenn zu einem späteren Zeitpunkt mehr Daten vorliegen, können durch eine erneute Analyse der Distanzkorrelationen zwischen microRNAs und konservierten Regionen deutlichere Aussagen getroffen werden.

Auch ist derzeit nur von wenigen Transkripten bekannt, ob sie Gewebe-spezifisch exprimiert werden. Sobald von Transkripten mit mehr Bestimmtheit eine Zugehörigkeit zu einem bestimmten Gewebe bekannt ist, sollte die Analyse wiederholt werden, bei der Leber-assoziierte Transkripte mit HNF1-Sites korreliert wurden. Zudem wäre es sinnvoll, diese Analyse auch noch mit anderen Genen und Transkriptionsfaktoren durchzuführen, von denen ebenfalls die Zugehörigkeit zu einem bestimmten Gewebe bekannt ist. Ebenso sollte in spezifischeren Analysen die Verteilung von Transkriptionsfaktorbindungsstellen nur in solchen Promotoren untersucht werden, von denen bekannt ist, dass die Transkriptionsfaktoren dort die Transkription regulieren. Allerdings konnte bisher von keinem Transkriptionsfaktor die vollständige Anzahl der Gene bestimmt werden, deren Transkription von dem Faktor reguliert wird.
Darüber hinaus kann *GenomeInspector* selbstverständlich auch verwendet werden um Distanzkorrelationen zwischen genomischen Elementen zu berechnen, die im Rahmen dieser Diplomarbeit noch nicht analysiert wurden, zum Beispiel zwischen verschiedenen Transkriptionsfaktorbindungsstellen oder zwischen Enhancern und Promotoren. Auch lassen sich schnell allgemeine Aussagen zu neu entdeckten oder vorhergesagten Elementen im genomischen Maßstab treffen. So ist zu hoffen, dass im Feld der Genomik noch viele neue Erkenntnisse mit Hilfe dieser Software gewonnen werden können.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] http://genomatix.de/online_help/help_eldorado/eldorado_elements.

[2] http://microrna.sanger.ac.uk/.

[3] http://www.illumina.com/pages.ilmn?ID=203.

[4] http://www.ncbi.nlm.nih.gov/About/primer/est.html.

[5] http://www.ncbi.nlm.nih.gov/Genbank/.

[6] http://www.ncbi.nlm.nih.gov/RefSeq/.

[7] T. Blank. Diplomarbeit: In silico Vorhersage von MicroRNAs in konservierten Regionen. 2007.

[8] K. Cartharius. MatInspector: Analysing Promoters for Transcription Factor Binding Sites. *Analytical tools for DNA, genes and genomes: nuts & bolts by Arseni Markoff ed., The nuts & bolts series, DNA Press*, 2005.

[9] K. Cartharius, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, and T. Werner. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21:2933–2942, 2005.

[10] M. D. Conkright, E. Guzman, L. Flechner, A. I. Su, J. B. Hogenesch, and M. Montminy. Genome-Wide Analysis of CREB Target Genes Reveals A Core Promoter Requirement for cAMP Responsiveness. *Molecular Cell*, 11:1101–1108, 2003.

[11] A. Siepel et al. Evolutionary conserved elements in vertebrate, insect, worm and yeast genomes. *Cold Spring Harbor Laboratory Press*, 15:1034–1050, 2005.

[12] P. Carninci et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics Advanced Online Publication*, 2006.

[13] X. Zhang et al. Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. *PNAS*, 102:4459–4464, 2005.

[14] M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196:261–282, 1987.

[15] S. Karlin. Statistical signals in bioinformatics. *PNAS*, 102:13355–13362, 2005.

[16] S. Karlin, B. E. Blaisdell, L. Cardon R. J. Sapolsky, and C. Burge. Assessments of DNA inhomogeneities in yeast chromosome III. *Nucleic Acids Research*, 21:703–711, 1993.

[17] S. Karlin and V. Brendel. Chance and Statistical Significance in Protein and DNA Sequence Analysis. *Science*, 257:39–49, 1992.

[18] S. Karlin and C. Macken. Assessment of inhomogeneities in an E. Coli physical map. *Nucleic Acids Research*, 19:4241–4246, 1991.

[19] S. Karlin and C. Macken. Chance and Statistical Significance in Protein and DNA Sequence Analysis. *Journal of the American Statistical Association*, 86:27–35, 1991.

[20] A. E. Kel, O. V. Kel-Margoulis, P. J. Farnham, S. M. Bartley, E. Wingender, and M. Q. Zhang. Computer-assisted Identification of Cell Cycle-related Genes: New Targets for E2F Transcription Factors. *JMB*, 309:99–120, 2001.

[21] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, 23:4878–4884, 1995.

[22] K. Quandt, K. Grote, and T. Werner. GenomeInspector: a new approach to detect correlation patterns of elements on genomic sequences. *CABIOS*, 9:405–413, 1996.

[23] K. Quandt, K. Grote, and T. Werner. GenomeInspector: Basic Software Tools for Analysis of Spatial Correlations between Genomic Structures within Megabase Sequences. *Genomics*, 33:301–304, 1996.

[24] F. Tronche, F. Ringeisen, M. Blumfeld, M. Yaniv, and M. Pontoglio. Analysis of the Distribution of Binding Sites for a Tissue-specific Transcription Factor in the Vertebrate Genome. *JMB*, 266:231–245, 1997.

[25] L. Zhang, S. Kasif, C. R. Cantor, and N. E. Broude. GC/AT-content spikes as genomic punctuation marks. *PNAS*, 101:16855–16860, 2004.