Bioinformatik

Seminar WS 2012/2013

Einführung in die Genomanalyse

Kim Weißer

inf9378

Inhaltsverzeichnis

Einleitung

In diesem Seminar geht es darum wie anhand der Struktur der DNA Gene analysiert werden können. Es wird zunächst der Aufbau der DNA erläutert und dann verschiedene Herangehensweisen zur Analyse erläutert.

Von Bakterien bis Mehrzeller die DNA ist überall enthalten. Sie enthält den individuellen Bauplan des jeweiligen Lebewesens.
Dass die DNA Träger der Erbinformation ist, weiß man jedoch erst seit 1944. Damals isolierten Oswald Avery und seine Mitarbeiter sowohl die DNA als auch Proteine von S-Pneumokokken und gaben sie zu je einer Kultur von R- Pneumokokken. Nur bei der Kultur in der die DNA gegeben wurde, ließen sich danach S-Pneumokokken nachweisen.
Dies schloss die Theorie, Proteine seien Träger der Erbinformation aus.
Seit dem ist die DNA immer genauer erforscht worden. [2]

Ein weiterer Begriff der eng der Erbinformation verbunden ist, ist der des Gens. Klassisch-biologisch ist das Gen als Teilstück eines Chromosoms definiert, dass eine einzelne sichtbare Eigenschaft des Phänotyps definiert. Heute weiß man das ein Gen aus mehreren DNA Stücken besteht.

Die zentrale Fragestellung des Seminars ist also, wie Gene in DNA-Sequenzen erkannt werden können.

Grundlagen

DNA-Aufbau

Wenn es darum geht die DNA zu analysieren, muss man zu nächst einmal wissen, was sie codiert. Dies können Polypeptide, Proteine, funktionelle RNA oder andere spezifische Genprodukte sein. Gene die z.B. funktionale RNA, also keine Proteine codieren, werden auch non-coding Gene (ncRNA) genannt. Etwa 5-10% des menschlichen Genoms hat solch eine Funktion.
Um die Sequenzen zu finden, die diese Polypeptide codieren (Gene). Muss zu nächst, die Struktur der DNA genauer erläutert werden. [1]

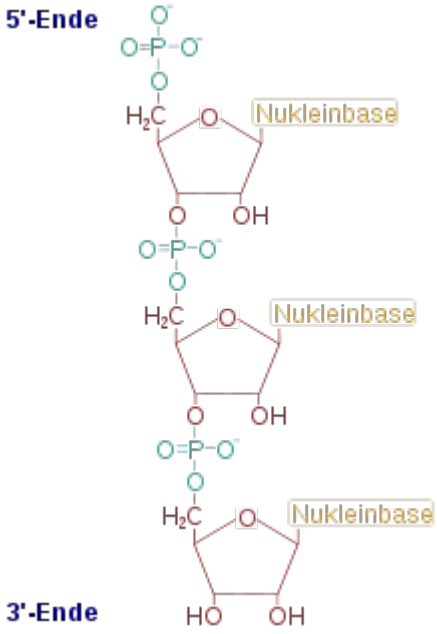


Abbildung 1: Quelle [11]

Die DNA ist ein fadenförmiges Makromolekül. Es besteht aus verketteten Nukleotiden. Ein Nukleotid setzt sich zusammen aus dem Zucker Desoxyribose, Phosphorsäure und aus einer Basen Adenin, Thymin, Cytosin oder Guanin.
In eukaryotischen Zellen sitzt die DNA im Zellkern.
In der DNA wechseln sich Zucker und Phosphatreste regelmäßig ab. Die Enden des fadenförmigen Moleküls, das sie bilden, bezeichnet man als 3' und 5'-Ende. Diese werden so bezeichnet, da die sich im Zucker befindenden Kohlenstoffatome nummeriert werden. Dabei wird am Sauerstoffatom gestartet.

Die DNA liegt in Form einer Doppelhelix vor. Es liegen sich zwei DNA Stränge gegenüber und bilden die Struktur einer Wendeltreppe. Sie sind verbunden über Wasserstoffbrücken an den Basen. Dabei paaren sich immer die Basen Adenin und Thymin, sowie Cytosin und Guanin. [2]

RNA

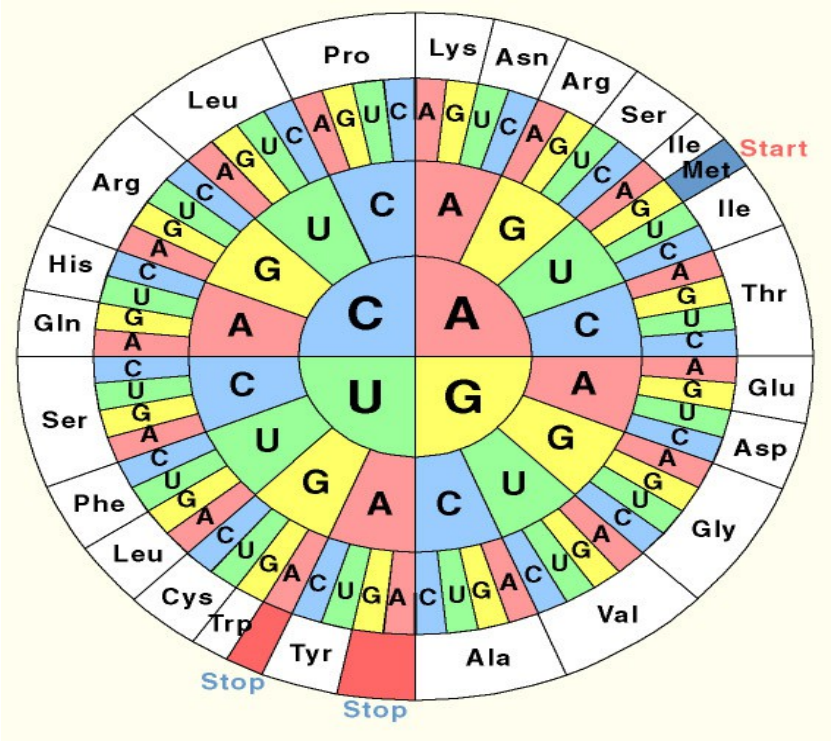
Neben der DNA existiert noch eine zweite Nukleinsäure in der Zelle. Die Ribonukleinsäure (RNA) besitzt statt Desoxyribose den Zucker Ribose und statt Thymin die Base Uracil.
Es gibt verschiedene Arten von RNA.
Die mRNA (messenger RNA) wird erzeugt in dem das Protein RNA-Polymerase ein proteincodierendes Stück der DNA vom 3' zum 5' Ende entlang fährt und dabei entsprechende RNA-Nukleotide zusammensetzt.
Die tRNA (transfer RNA) dient dazu anhand von Nukleotiden Proteine zusammenzusetzen. Ihre Länge liegt zwischen 73 und 95 Nukleotiden.
Neben diesen beiden gibt es noch viele weitere RNA-typen.

Kodierung von Aminosäuren

Es gibt 20 Aminosäuren. Sie sind die Bestandteile von Proteinen. Damit die DNA diese codieren kann, wird eine Folge von drei Nukleotiden benötigt. Dies ergibt 64 (4 hoch 3) Möglichkeiten. Da dies mehr als die eigentlichen 20 minimal benötigten Kombinationen sind, sind einige Codons mit der gleichen Aminosäure belegt.

Außerdem gibt es Start- und Stopcodons, welche den Beginn und das Ende eines Gens markieren.

Der genetische Code ist beinahe universell. Es gibt ein paar Organismen , bei denen die Aminosäure welche das Codon darstellt, abweicht.



Zeichnung 1: Quelle [13]

Ein Gen ein Polypeptid Hypothese

Beadle und Tantum schlossen aus einer Reihe von Experimenten, dass jedes Gen die Synthese eines bestimmten Enzyms codiert. Sie formulierten die Ein-Gen-Ein-Enzym-Hypothese. Diese wurde später zur Ein-Gen-Ein-Polypeptid-Hypothese verfeinert.

Die Eigenschaften des genetischen Codes:

- Er ist ein Triplett-Code.
- Er gilt für beinahe alle Lebewesen, ist also universell.
- Jedes Triplett codiert genau eine Aminosäure.
- Die Tripletts schließen lückenlos aneinander.
- Jede Base gehört nur zu einem Codon. Es gibt keine Überlappungen.
- Er wird in 5'->3' Richtung gelesen.

Vom Gen zum Polypeptid

In der Zelle wird aus der DNA ein Protein, in dem die DNA zunächst in mRNA übersetzt wird. Dieser Prozess heißt Transkription. Die Umwandlung von mRNA in ein Polypeptid wird Translation genannt.

Die Transkription startet am Promoter. Hier bindet sich das Enzym RNA-Polymerase an die DNA und spaltet sie blasenartig auf. Das Enzym fährt nun den DNA Strang entlang. Dabei kann es anhand der Basenpaarung die mRNA-Sequenz zusammenbauen.
Prokaryoten besitzen nur einen Polymerasetyp, Eukaryoten hingegen mehrere.

Die Translation geschieht an den Ribosomen. Sie kann nur mit Hilfe von tRNA stattfinden.

Introns

Seit 1977 ist bekannt, dass eukaryotische Gene nicht aus einer einzigen DNA Sequenz bestehen. Sie werden immer wieder durch nicht codierende Sequenzen unterbrochen. Diese Sequenzen werden Introns genannt, die codierenden hingegen Exons. Wie die meisten nicht codierenden DNA-Sequenzen, besitzen Introns kurze Abschnitte mit Aminosäuren und viele Stopcodons. Sie sind außerdem nicht an die Codons gebunden, können also inmitten eines Codons beginnen und enden.

Spleißen

In der Zelle werden die Introns aus der prä-mRNA durch das Spleißen herausgelöst. Es entsteht die fertige mRNA. Diesen Vorgang nennt man mRNA-Reifung.

Die Herausforderung für die DNA-Analyse ist dabei also, die richtigen Schnittstellen (Splice Sites) zu finden.

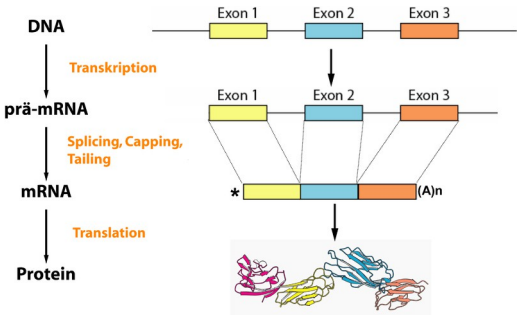


Abbildung 2: Quelle [12]

Eigenschaften von Pro- und Eukaryoten für die Genanalyse

Um den Aufbau und die Funktion von Genen heraus zu finden, gibt es verschiedene Genanalyseprogramme. Einige davon sind online zugreifbar.
Für die Analyse der DNA ist es weiterhin wichtig zwischen von Pro- und Eukaryoten zu unterscheiden. Für sie gibt es verschiedene Programme. [1]

Prokaryoten

Das prokaryotische Genom enthält in der Regel zwischen 1 und 6 Millionen Basenpaare und ist meist kleiner als Eukaryotisches, daher kann es in einem Stück verarbeitet werden. Das Genom von Eukaryoten muss hingegen erst geteilt werden.
Prokaryotische DNA besitzt fast keine Introns, daher sind Gene hier leichter zu identifizieren. Außerdem ist die Promotorregion bei Prokaryoten normalerweise klar definiert und liegt meist näher am eigentlichen Gen.
Für Prokaryoten gibt es daher eine Reihe von Analyseprogrammen, z.B. ORPHEUS und GLIMMER.
Zum Teil sind diese auf verschiedene Arten spezialisiert. Es gibt unter den Arten z.B. Vorlieben für bestimmte Codons. Daher ist es sinnvoll, Analyseprogramme zu verwenden, die bereits auf die entsprechende Art trainiert sind.

Eukaryoten

Bei Eukaryoten gibt es einige Faktoren, welche die Analyse erschweren. Zum einem ist die Gendichte geringer. Außerdem müssen Exons und die genauen Gebiete, an denen das Splicen statt findet, lokalisiert werden. Dann wird aus den Exons ein Protein synthetisiert und gesehen, ob dieses vergleichbar mit bereits bekannten Proteinen ist.
In seltenen Fällen, in den bei Eukaryoten ebenfalls wenige Introns vorkommen, sind Analyseprogramme für Prokaryoten anwendbar.

Programme, die sowohl für Eukaryoten, als auch für Prokaryoten geeignet sind, gibt es ebenfalls. Sie werden aber für Prokaryoten weniger genutzt. Grail (Gene Relationships Across Implicated Loci) ist so eines, es basiert auf dem Vergleich von Genen.

Ein typischer Ablauf bei der Analyse von Eukaryotischen Genen ist, die Sequenz zunächst einem Exonerkennungsprogramm zu geben. Für alle Exons wird dabei das wahrscheinliches Exon in eine Proteinsequenz übersetzt. Dies wird für alle Leserahmen gemacht. Die Sequenz mit den wenigsten Stopcodons wird mit in Datenbanken erfassten Sequenzen verglichen. [1]

Offener Leserahmen

Als Offenen Leserahmen bezeichnet man die DNA Sequenz, die das eigentliche Polypeptid codiert. Die Region liegt also zwischen Start- und Stopcodon.

Die Sequenz 'ATCTGUGCTTAT', kann auf drei verschiedene Weisen in Tripletts unterteilt werden:
ATC TGA CCT TAT
TCT GAC CTT
CTG ACC TTA

Dies sind die 3 der möglichen 6 Leserahmen. Die anderen 3 befinden sich auf dem komplementären DNA-Strang.

Bei der Genanalyse ist es nun wichtig, den richtigen Leserahmen zu finden. Eine sehr einfache Herangehensweise wäre dabei lange Sequenzen zu suchen, die nicht von Stopcodons unterbrochen werden. Das Problem dabei ist, dass kurze codierende Sequenzen nicht entdeckt werden.
Auch Introns erschweren den richtigen Leserahmen zu finden.

Die folgende Tabelle zeigt mögliche Konsequenzen für das konstruierte Protein bei der falschen Vorhersage von Exons

Start des Exons	Länge des Exons	Effekt auf die Übersetzung des Exons	Effekt auf die Übersetzung des nächsten korrekt startenden Exons
Korrekt	Korrekt	Korrekt	Korrekt
	Falsch, richtiger Rahmen	Korrekt, aber zusätzliche oder fehlende Teile	Korrekt, außer vielleicht dem ersten Stück
	Falsch, falscher Rahmen	Korrekt, aber zusätzliche oder fehlende Teile	Falsch
Falsch, richtiger Rahmen	Korrekt	Korrekt, aber zusätzliche oder fehlende Teile	Korrekt, außer vielleicht dem ersten Stück
	Falsch, richtiger Rahmen	Korrekt, aber zusätzliche oder fehlende Teile	Korrekt, außer vielleicht dem ersten Stück
	Falsch, falscher Rahmen	Korrekt, aber zusätzliche oder fehlende Teile	Falsch
Falsch, falscher Rahmen	Korrekt	Falsch	Falsch
	Falsch, richtiger Rahmen	Falsch	Falsch
	Falsch, falscher Rahmen	Falsch	Möglicherweise korrekt

Quelle: [1]

Vergleiche von Analyseprogrammen

Um die Genauigkeit von Genanalyseprogrammen zu bestimmen, wird häufig ein numerischer Wert ermittelt.

Die Genauigkeit von Analyseprogrammen lässt sich in drei Level beurteilen: Base-Level, wo die einzelnen Nukleotide betrachtet werden, durch die Exon-Struktur (Exonlevel) und die Korrektheit des Proteinprodukts (Protein-Level).

Auf dem Nukleotid-Level wird zwischen true-positive (TP), false-positive (FP), true-negative (TN) und false-negative (FN) Nukleotiden unterscheiden. TP sind Nukleotide, die zu einem Gen gehören und als solche erkannt werden. FP sind Nukleotide, die als Genbestandteile erkannt werden, jedoch keine sind. Entsprechend sind TN und FN nicht identifizierte Nukleotide, die richtig bzw. falsch als solche erkannt wurden.

Angegeben wird die Genauigkeit dann in Sensitivity (Sn) und Specificity (Sp):

$$Sn = \frac{TP}{TP + FN} \qquad Sp = \frac{TP}{TP + FP}$$

Die Sensitivity sagt also aus, wie viele der Nukleotide eines Gens als solche erkannt werden. Die Specificity hingegen sagt aus, wie genau das Programm arbeitet. So hat z.B. ein Programm, welches die komplette ihm gegebene Sequenz als Gen erkennt, eine Sensitivity von 1 aber eine sehr niedrige Specificity.

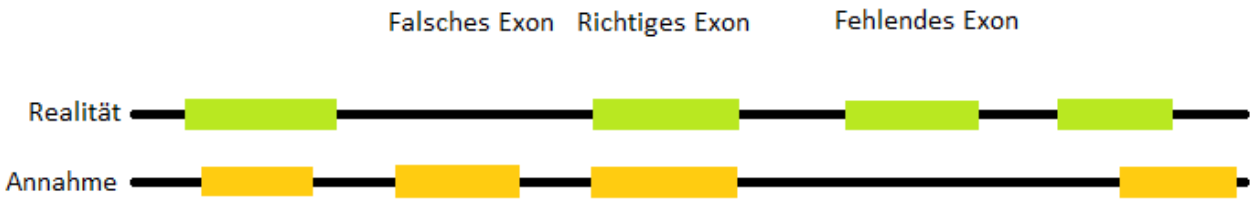
Die Genauigkeit kann auf diese Weise nur mit beiden Werten bestimmt werden. Umgangen wird dieses Problem mit dem approximate correlation coefficient (AC). Dieser wird mit Hilfe der average conditional probability(ACP) berechnet.

$$ACP = \frac{1}{4}\left[\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right]$$

$$AC = 2 * ACP - 1$$

Bestimmen der Genauigkeit von Programmen
Um die Genauigkeit von Exon Vorhersagen zu prüfen, werden die Exons in Proteinsequenzen übersetzt und diese werden mit bereits analysierten Sequenzen verglichen. Danach wird die wahrscheinlichste Variante gewählt. Sind die Exons entschlüsselt, ist immer noch nicht bekannt, welche Funktion die resultierenden Proteine haben.

Auf dem Exon Level unterscheidet man zwischen richtig vorhergesagten Exons, falsch vorhergesagten Exons, und fehlenden Exons.



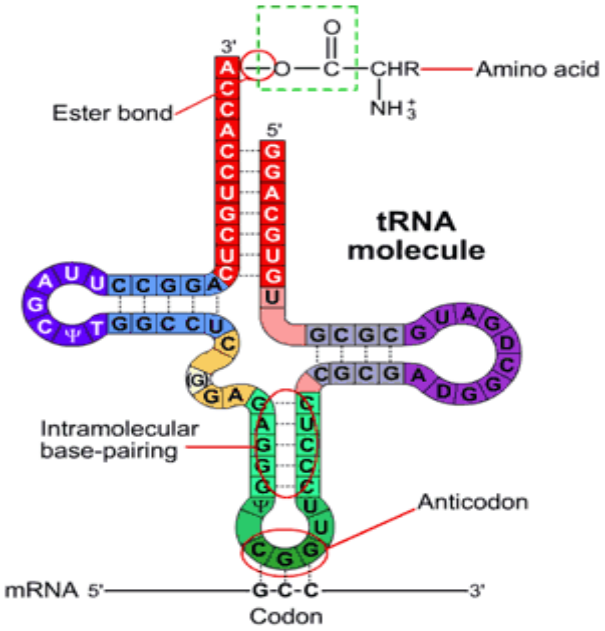
Sensitivity und Specificity berechnen sich folgendermaßen:

$$Sensitivity = \frac{Anzahl\ aller\ richtig\ vorhergesagten\ Exons}{Anzahl\ aller\ wirklichen\ Exons}$$

$$Specificity = \frac{Anzahl\ aller\ richtig\ vorhergesagten\ Exons}{Anzahl\ vorhergesagter\ Exons}$$

T-RNA

Bei der Suche nach Genen ist es leichter, zunächst nach funktionaler RNA und sich wiederholender Sequenzen zu suchen.
Ein Beispiel dafür ist die tRNA. Diese hat eine signifikante Form. Damit sich diese Struktur bilden kann, spielen Basenpaare eine signifikante Rolle. Zusammengefügt bilden sie typischerweise eine Kleeblattform. Daran können sich spezielle Analyseprogramme wie zB. tRNAscan bei der Suche orientieren. Standardprogramme hingegen überlesen diese Sequenzen leicht, da sie nicht die typische Codonstruktur besitzen.
Die t-RNA besitzt eine Anhaftstelle für eine Aminosäure sowie ein Anticodon, welches das Gegenstück zum Codon auf der mRNA bildet.
Bei der Translation fährt nun das Ribosom am mRNA-Strang entlang und setzt nach einander die Aminosäuren von der tRNA entsprechend dem mRNA-Strang zusammen.



Es ist sinnvoll zunächst tRNA Gene zu identifizieren, um herauszufinden welche Codons der Zelle zur Verfügung stehen. Dabei kann sich herausstellen, dass manche Codons nicht genutzt werden. Dies ist wichtig für die spätere Analyse.

Die Methode von tRNAscan beruht auf einem Entscheidungsbaum. Jeder einzelne Schritt wird überprüft. Dabei werden häufig unabhängig vom Ausgang des jetzigen Schritts weitere Schritte durchgeführt. Dabei wird gezählt, wie häufig die Tests bestanden werden. Diese ist am Ende ein Indikator für ein Gen.
Bei dieser Methode liegt die Identifizierungsrate bei etwa 97,5%. Dabei wird nur ein Gen auf 3 Millionen Basen falsch als Gen identifiziert. Insgesamt also ein sehr genaues Programm.
Für Eukaryoten hat tRNAscan jedoch eine zu hohe Rate von false-positive Identifizierungen. Todd Lowe und Sean Eddy entwickelten 1997 tRNAscan. Dies führt selber keine Genanalyse durch sondern gibt die Sequencen an Unterprogramme weiter. Zu diesen gehören tRNAscan, der Pavesialgorithmus, welcher nach Sequenzen sucht, die die Polymerase III kontrollieren und ein Algorithmus von Sean Eddy und Richard Durbin. Am Ende verbindet es diese zu einer genauen Vorhersage.

Homologie

Vergleiche mit bereits identifizierten Genen sind sowohl für Eukaryoten als auch Prokaryoten einsetzbar.
Vergleiche mit ähnlichen bereits identifizierten Genen können bei der Analyse von Genen helfen. Dabei lassen sich auch Rückschlüsse auf die jeweilige Funktion des Gens ziehen.

DNA-Sequenzen, die sich über mehrere Arten hinweg erhalten haben, sind mit großer Wahrscheinlichkeit wichtig für die Organismen.
Gene, die einzigartig für eine Art sind, können auf diese Weise nicht gefunden werden.
Die DNA-Sequenzen, die auf diese Weise gefunden werden, müssen nicht unbedingt ein Protein kodieren, sie können auch eine funktionale Bedeutung haben. Bei Eukaryoten ist die Suche nach Gleichheit schwieriger, da sich das gesuchte gen über mehrere Exons verteilen kann.
Ein Beispiel für immer wieder auftauchende DNA Sequenzen sind Transposons,.

Transposons auch als "springende Gene" bekannt, wurden 1948 von Barbara McClintock entdeckt. Sie können sich selbst vermehren und in andere DNA-Stücke springen. Dies kann dann gefährlich werden, wenn es in andere Gene springt und diese blockiert. Mindestens 45 Prozent unseres Genoms gehen auf ehemals springende, sowie teilweise noch springende Gene zurück.
Alu-Elemente sind Relikte solcher springenden Gene. Sie sind die sich am meisten wiederholenden Sequenzen in der menschlichen DNA und haben die Fähigkeit, sich zu reproduzieren und neu in die DNA einzufügen, verloren.
Sie werden auch als selfish DNA bezeichnet, weil sie kein Protein codieren und nur um ihrer selbst Willen zu existieren scheinen.
Alu-Elements sind in Introns weit verarbeitet. Manche werden beim Spleißen als Exon erkannt und verändern so das jeweilige Gen. [1,5,6]

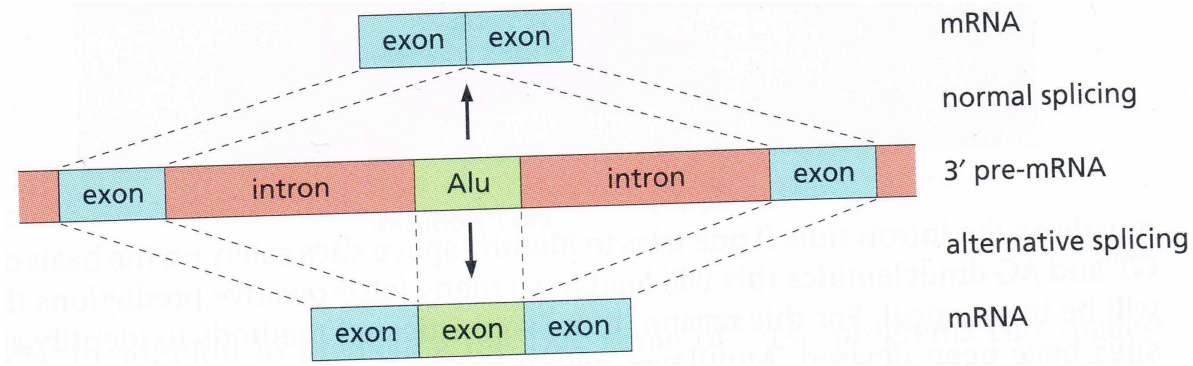


Abbildung 3: Quelle [1]

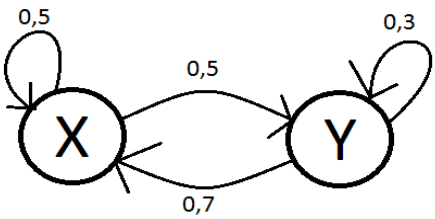
Lernende Programme

Eine andere Möglichkeit die DNA zu analysieren, ist durch so genannte „lernende" Programme.
Als Grundlage erhalten diese Programme zwei Mengen mit codierenden und nicht codierenden DNA Sequenzen.
Anhand dieser Sequenzen kann das Programm entscheiden, ob eine unbekannte Sequenz der einen oder andern Menge ähnlicher ist.

Markov Model

Ein Weg, um lernende Programme zu realisieren, sind Markov-Models. Sie werden in verschiedenen Abwandlungen verwendet, um Gene zu identifizieren.
Das einfachste Markovmodel ist ein Markov chain model. Dieses ist ein System von Zuständen mit einer gegebenen Wahrscheinlichkeit für die Übergänge zwischen diesen.

Hierzu ein Beispiel:
Ein System mit zwei Zuständen, welches eine Folge von A und B analysiert, wechselt beim Auftreten eines A in Zustand X, bei einem B in Y. Vom Zustand X in Y wechselt es mit einer Wahrscheinlichkeit 0,5. Die Wahrscheinlichkeit, in X zu bleiben, ist entsprechend genauso hoch. Umgekehrt von Y nach X wechselt man hier mit einer Wahrscheinlichkeit von 0,7.



Mit der Annahme, dass der Ausgangszustand X ist, würde die Folge ABA mit der Wahrscheinlichkeit 0,5*0,3*0,7 berechnet werden. [7]

Eine Erweiterung von Markov chain Models sind hidden Markov Models (HMM).

GeneMark, arbeitet mit modifizierten HMM, mit inhomogenen Markov chain models. Es wurde 1993 am Georgia Institute of Technology in Atlanta, Georgia, USA entwickelt.
Es ist eines der wenigen Programme, die sowohl für Pro- als auch Eukaryoten anwendbar ist. [1,3]

Genutzt werden hierbei Markov chain model der 5ten Ordnung. Es wird also die Wahrscheinlichkeit für das Auftreten der Base a unter der Bedingung berechnet, das vorher die Sequenz $x_1x_2x_3x_4x_5$ aufgetreten ist. ($P(a|x_1x_2x_3x_4x_5)$)
Damit das Programm arbeiten, kann muss zunächst die Wahrscheinlichkeit $P(a|x_1x_2x_3x_4x_5)$ für alle möglichen Pentamere berechnet werden. [1]
Dies kann mit Hilfe einer Trainingsmenge geschehen. Die Wahrscheinlichkeit berechnet sich dabei folgendermaßen:

$$P(a \mid x_1 x_2 x_3 x_4 x_5) = \frac{n_{x_1 x_2 x_3 x_4 x_5 a}}{\sum_{z=A,C,G,T} n_{x_1 x_2 x_3 x_4 x_5 z}}$$

Dabei steht $n_{x_1 x_2 x_3 x_4 x_5 a}$ für die Anzahl mit welcher die Folge $x_1x_2x_3x_4x_5$ im Trainingsmenge vorkommt.

Jeder Leserahmen hat seine eigenen Wahrscheinlichkeiten. Diese werden hier durch einen Index gekennzeichnet: $P_1(a|x_1x_2x_3x_4x_5)$ für den 1.Leserahmen, $P_2(a|x_1x_2x_3x_4x_5)$ für den 2.Leserahmen, etc. Der Leserahmen wird dabei anhand der 5ten Base festgelegt.

Wenn nun also die Wahrscheinlichkeit berechnet werden soll, dass die Sequenz $x_1x_2x_3x_4x_5x_6x_7x_8x_9$ eine codierende Sequenz im 2. Leserahmen ist ($P(x|2)$), ergibt sich folgende Formel:

$$P(x \mid 2) = P_1(x_1 x_2 x_3 x_4 x_5) * P_1(x_6 \mid x_1 x_2 x_3 x_4 x_5)$$
$$* P_2(x_7 \mid x_2 x_3 x_4 x_5 x_6) * P_3(x_8 \mid x_3 x_4 x_5 x_6 x_7)$$
$$* P_1(x_9 \mid x_4 x_5 x_6 x_7 x_8)$$

Splice Site Erkennung

Um Eukaryotische Gene vollständig zu analysieren, erfordert das eine Erkennung ihrer Spleißingstruktur, also die Art, in der die RNA geteilt wird und Introns entfernt werden. Manche Programme integrieren Informationen über Splice Sites um Exons zu erkennen. Es gibt aber auch Programme, die speziell darauf ausgerichtet sind Splice Sites zu erkennen.
Es gibt Exon-Intron Splice Sites, sowie welche, wo das Intron ins Exon übergeht. Ein DNA-Strang wird immer vom 5' zum 3' Ende gelesen. Die Splice Site, bei der das 3'-Ende des Exons an das 5'-Ende des Introns grenzt, nennt man Donor. Die andere Akzeptor.

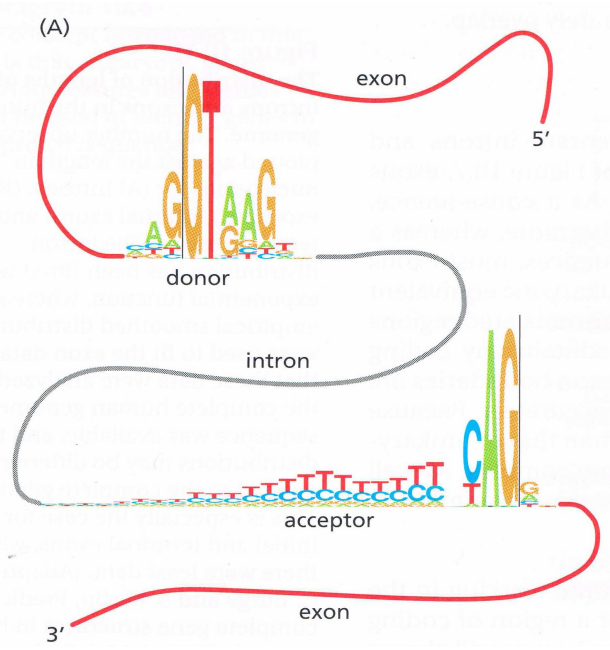


Abbildung 4: Quelle [1]

Viele Introns beginnen mit dem Dinukleotid GT und enden mit AG. Die U12-type Introns besitzen am 5'-Ende AT und am 3'-Ende AC.
Wenn also alle Positionen von GT und AG lokalisiert werden, erkennt man alle Splice Sites, auf die dieses Schema zutrifft, aber man hat auf etwa eine richtige Splice Site 30 – 100 falsche. Um dem entgegen zu wirken, muss man die Umgebung der Splice Site mit einbeziehen. Informationen, um eine Schnittstelle zu erkennen, befinden sich hauptsächlich auf der Intron Seite der Splice Site. [1]

Die SplicePredictor-Methode:

Schnittstellen, können mit Hilfe von Mustern und Basenstatistiken erkannt werden.
Beide Seiten der Schnittstelle werden betrachtet. Es gibt bestimmte Faktoren, die ein Spleißen beeinflussen. Spleißen geschieht also mit einer gewissen Wahrscheinlichkeit.
Ob es sich um eine Schnittstelle handelt, kann mit einem Grenzwert ermittelt werden.
Dabei ist es wichtig, die ungefähre Länge eines Introns zu kennen, da auch andere nahe Werte eine hohe Wahrscheinlichkeit erreichen können.[1]

Promotor

Die Promotorregion ist für die Analyse wichtig, da von hier die Transkription startet. Die Polymerase bindet sich an diese. Es gibt bestimmte Faktoren, die dabei fördernd oder hemmend auf die Transkription einwirken. [2]

Die Identifikation der Promotorregion ist jedoch schwierig, da es viele verschiedene Faktoren gibt. Diese treten nur sehr unterschiedlich auf und haben häufig unterschiedliche Vorlieben für bestimmte Sequenzen. Daher gibt es unterschiedlich Methoden, um sie zu lokalisieren.

Der δ-Faktor, ein Teil der bakteriellen Polymerase, bindet sich an die Pribnow-Box, ein Teil des Promotors, und erhöht an der Stelle entscheidend die Bindungswahrscheinlichkeit. Die Pribnow-Box besitzt die Konsensussequenz 5'-TATAAT-3'. Diese ist eine DNA-Sequenz die bei Vergleichen zwischen verschiedenen Promotorregionen immer wieder vorkommt. Sie liegt an Positon -10 von der Transkription Start Side aus.

Ein sich ebenfalls wiederholendes Element ist die Minus-35-Box mit der Konsensussequenz 5'-TTGACA-3'.
Ein weiteres Element, kann eine AT-reiche Region vor der Minus-35-Box sein.[1,8,9]

Eukaryoten hingegen besitzen häufig eine TATA-Box mit der Konsensussequenz 5'-TATAAT-3. Diese liegt in der Regel 25 Basenpaare stromaufwärts vom Startcodon.[2]

Ein weiterer Indikator für die Promotorregion sind CpG-Inseln. Diese sind eine Häufung von CG-Dinukleotiden. Das p in der Mitte rührt von dem Phosphatrest her, der sich zwischen den beiden Nukleotiden befindet. Es gibt sie hauptsächlich bei Eukaryoten. [10]

Viele Programme beziehen neben der Kernregion des Promoters auch den Transkriptionsstartpunkt mit ein (also der Region wo das eigentliche Gen beginnt).

Zusammenfassung

Die Region, welche ein Gen codiert, ist der Offene Leserahmen. Diesen müssen Analyseprogramme erkennen.
Es gibt verschiedene Programme für Pro- und Eukaryoten. Sie bedienen sich häufig einer Kombination verschiedener Methoden, wie der Suche nach Homologien und Mustern.
Um Programme vergleichen zu können, misst man deren Genauigkeit in Sensivity und Specificity.

Manche Programme arbeiten mit Methoden, die es ihnen ermöglichen „zu lernen“. Sie können anhand von Trainingsmengen bestimmen, ob eine DNA-Sequenz ein Gen ist oder nicht.
Des weiteren begibt man sich auf die Suche nach wichtigen und signifikanten Regionen, wie z.B. Promotorregion oder Splice Sites.

Da die Eukaryotische DNA von Introns unterbrochen wird, ist Exonerkennung besonders wichtig. Dieses kann durch das Erkennen der Splice Sites geschehen.

Quellen

[1] Understanding Bioinformatics, Zvelebil/Baum, Garland Science, 2008

[2] Grüne Reihe Materialien SII, Genetik, Schroedel, 2007

[3] www.exon.gatech.edu, GeneMark: Background information, Abrufdatum: 12.11.2012

[5] www.geneticorigins.org, Aluframeset, Abrufdatum 21.11.2012

[6] www.focus.de, DNA-Elemente: Das große Springen, Abrufdatum: 24.11.2012

[7] www.algorithm.cs.sunysb.edu, Abrufdatum: 22.11.2012

[8] www.wissenschaft-online.de, Pribnow-Box, Abrufdatum: 17.12.2012

[9] www.wissenschaft-online.de, TATA-Box, Abrufdatum: 17.12.2012

[10] Genetik, 4 .Auflage, Jochen Graw, Springer-Verlag, 2006

[11] www.wikipedia.org, Nukleinsäure, Abrufdatum: 9.10.2012

[12] www.wikipedia.org, Spleißen (Biologie), Abrufdatum: 17.12.2012

[13] www.chemgapedia.de, RNA- und DNA-Aufreinigung, Abrufdatum: 17.12.2012