Modul 10B:
Übungen zur Bioinformatik

Teil 1 – DNA-Sequenzanalyse und Gensuche

31.08.-04.09.2015

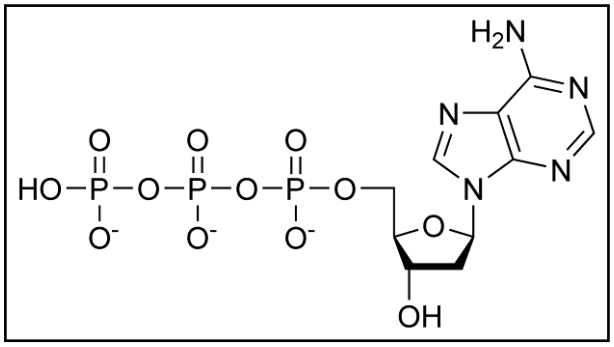
AG Hankeln
Institut für Molekulargenetik
J. J. Becherweg 30a
55128 Mainz

1. OLD SCHOOL: mit Fred Sanger zur ersten DNA-Sequenz

Die Sequenzierung von DNA ist die Grundlage der Genomforschung. Im ersten Kapitel beschäftigen wir uns daher mit der klassischen DNA-Sequenzierung nach Sanger. Die Sanger-Sequenzierung kennen Sie bereits aus der Vorlesung. Wir rufen uns zunächst einige Grundlagen in den eigenen „Arbeitsspeicher" zurück...

→ Warum wird diese Art der Sequenzierung auch „Kettenabbruchsequenzierung" genannt?

Betrachten Sie das unten abgebildete Nukleotid.

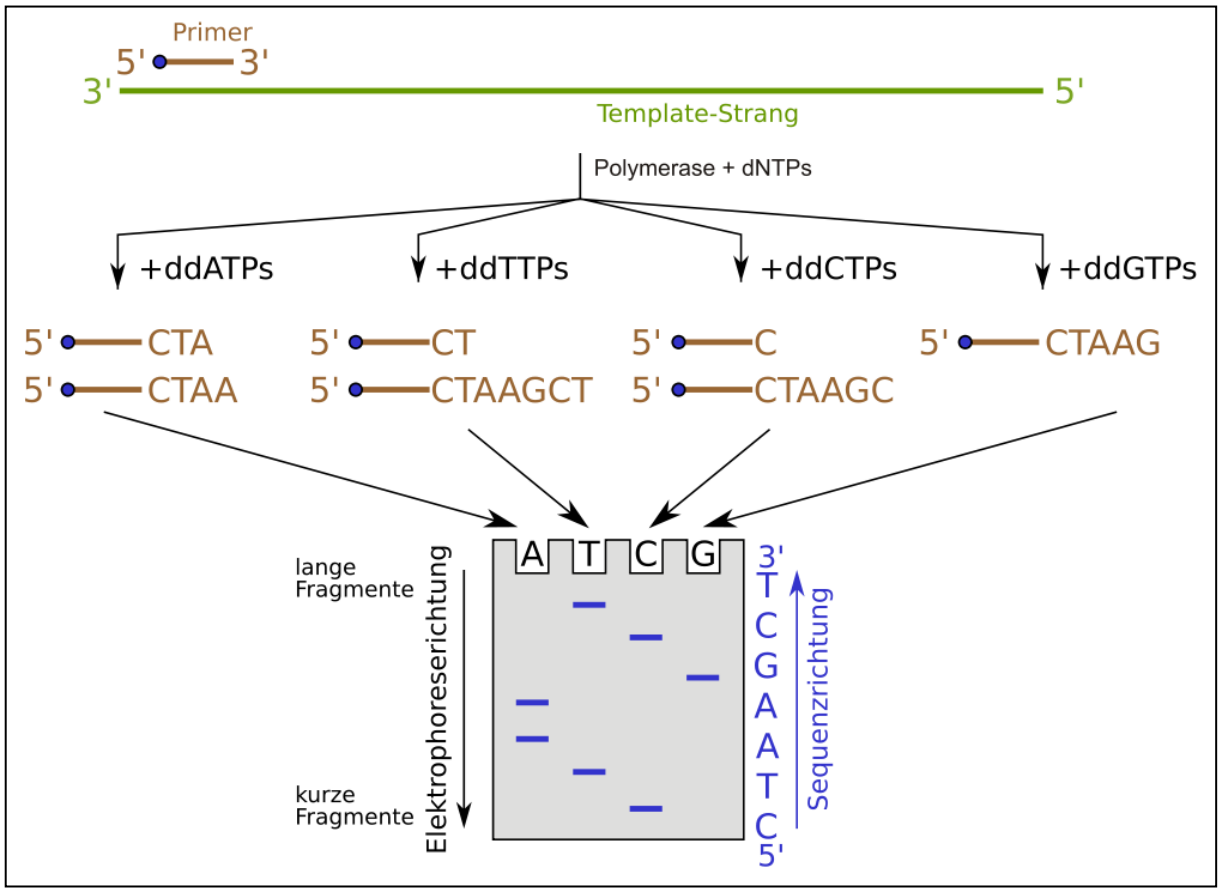


→ Um welches der vier DNA-Nukleotide handelt es sich hier?

→ An welcher Stelle müssen Sie das Nukleotid modifizieren, um einen Kettenabbruch zu verursachen?

Im Folgenden machen wir einen kleinen Zeitsprung in die Anfangsphase der Sanger-technik, denn erst dann lernen Sie die modernen Methoden ein wenig schätzen :-))

Die folgende Abbildung zeigt das Prinzip der Sanger-Sequenzierung VOR der Erfindung fluoreszenzmarkierter Terminator-Nukleotide. In jeden der 4 Sequenzieransätze, aufgeteilt auf 4 Eppendorfgefäße, gab man alle dNTPs aber je nur 1 ddNTP hinzu. Die Markierung der entstehenden Sequenzierprodukte erfolgte durch radioaktive Nukleotide, der Nachweis durch Gelelektrophorese und Autoradiographie. Wegen des Nachweises durch Filmschwärzung brauchte man 4 Gelspuren, um die 4 basenspezifischen Sequenzierungsansätze auszulesen.



→ Wozu dient der sogenannte „Primer"?

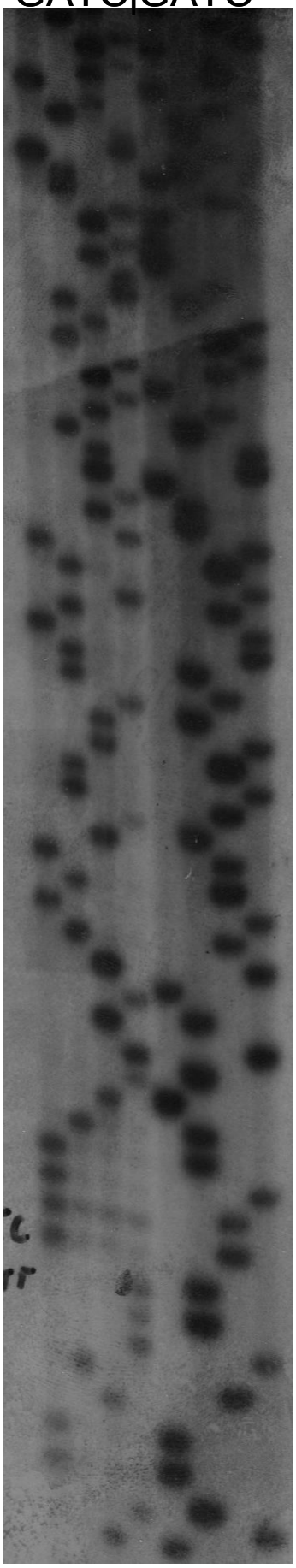
→ Wo befinden sich beim Gellauf Anode und Kathode? Machen Sie sich klar, wie die DNA geladen ist und warum!

→ Wie lautet die Sequenz des <u>Matrizen</u>-Stranges?

5' - _ _ _ _ _ _ _ - 3'

Doch jetzt eine echte OLD SCHOOL-Übung (heute kann das kaum jemand mehr :-)...

GATC|GATC



Das nebenstehende Bild eines Sequenziergels zeigt 8 Spuren (also 2 Sequenzierungen). Wie Sie sehen, sind Ergebnisse in echt selten so schön wie in gezeichneten Schemata. Versuchen Sie trotzdem einmal, das gezeigte Gelbild zu interpretieren!

→ Notieren Sie die beiden Sequenzen, wie sie aus dem abgebildeten Gel abgelesen werden können. Achten Sie auf die richtige Leserichtung und geben Sie die Orientierung der Sequenz an!

Wenn Sie später im Kurs noch Zeit haben, können Sie durch BLAST-Datenbanksuchen versuchen herauszufinden, was für DNA hier *anno* 1985 analysiert wurde...

Sequenz 1:

☐- _

_ _

_ -☐

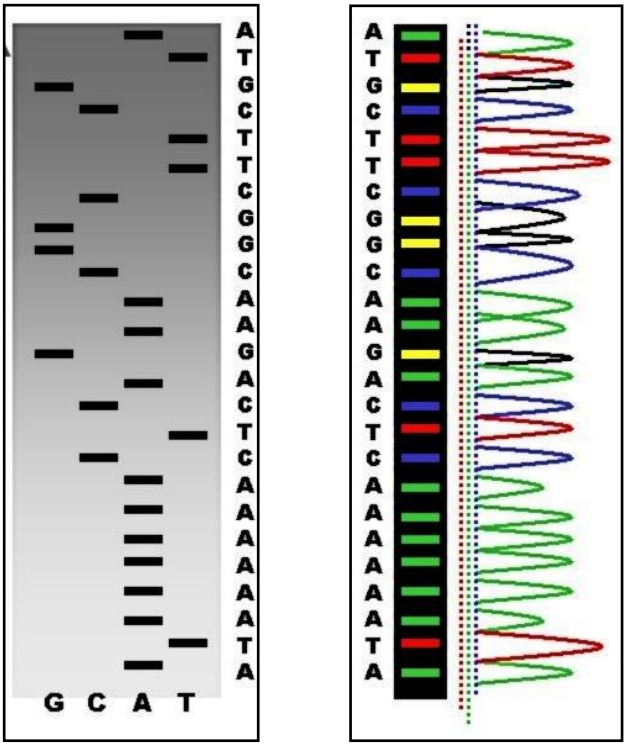
Sequenz 2:

☐- _

_ _

_ -☐

Wie Sie nun selbst gemerkt haben, war das Sequenzieren mit der Sanger-Methode aufwändig und mühsam (dabei waren Sie noch nicht einmal im Isotopen-Labor!). Heutzutage erfolgt die Markierung der Nukleotide für gewöhnlich nicht mehr radioaktiv, sondern mit Fluoreszenzfarbstoffen. Da jedes der vier Nukleotide einen anderen Fluoreszenzfarbstoff trägt, kann die Sequenzierungsreaktion mit allen vier ddNTPs in einem Gefäß erfolgen („one tube"). Die Gelelektrophorese wird in sehr feinen Kapillaren durchgeführt (auch hier nur noch 1 Gelspur pro Sequenz → „one lane") und die Auswertung der Gele, das sogenannte „base-calling" erledigt ausgeklügelte Software.

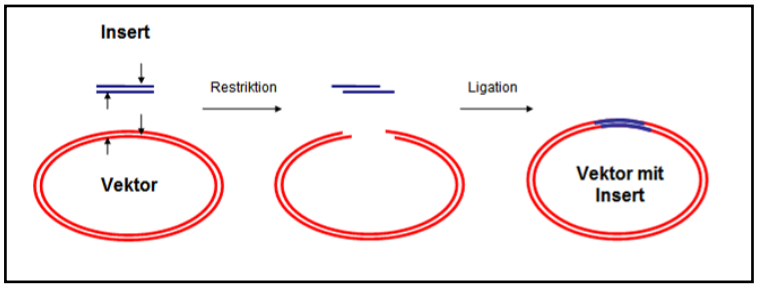


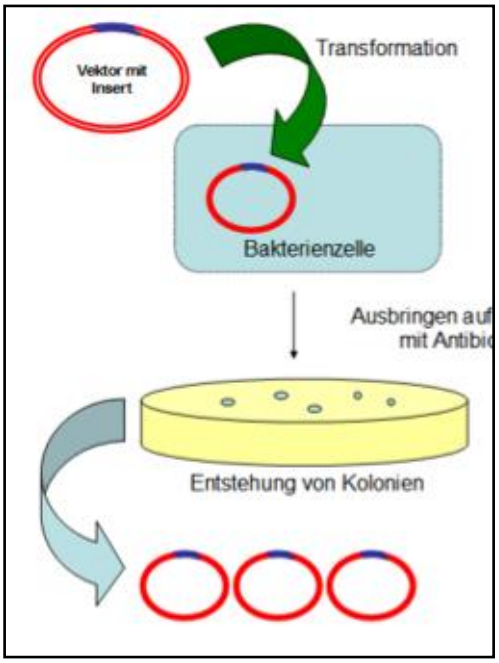
Sanger-Sequenzierung „damals und heute"
Links: 4 Reaktionen, in jeder Reaktion nur 1 bestimmtes ddNTP
Rechts: „one tube, one lane"-Prinzip, die ddNTPs sind mit verschiedenen Floureszenzfarbstoffen markiert. Die Gel-Auswertung erfolgt automatisch, Ausgabe sind sogenannte „Chromatogramme" (hierbei entspricht die peak-Höhe der Fluoreszenz-Instensität)

Die Auswertung von Sequenzierungen ist trotz dieser Erleichterungen noch immer höchst wichtig und keineswegs trivial! Im nächsten Abschnitt werden Sie daher moderne Chromatogramme bearbeiten und auswerten. Für das Verständnis ist es wichtig zu wissen, wie überhaupt die Ausgangs-DNA-Moleküle (Matrizen) für eine DNA-Sequenzierung gewonnen werden. Betrachten Sie also einmal den folgenden Exkurs...

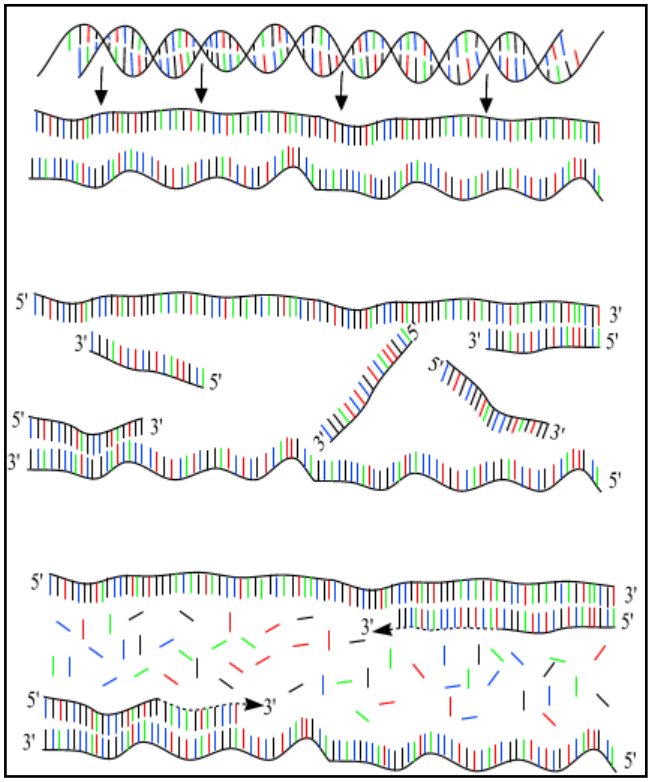
Exkurs:
Wie bekomme ich genügend Matrizen-DNA für eine Sequenzierungsreaktion?

1. Möglichkeit: die klassische DNA-Klonierung (Gentechnik)





Bei einer Klonierung werden das zu sequenzierende DNA-Stück („Insert") und ein Plasmid („Vektor") mittels der DNA-Ligase verbunden.
Anschließend wird das entstandene Konstrukt in eine Bakterienzelle eingebracht. Das Bakterium vermehrt das Plasmid mitsamt dem Insert tausendfach, so dass man genug Material für eine Sequenzierung erhält. Da die Sequenz des Plasmids vollständig bekannt ist, können Primer für die Sequenzierung hergestellt werden, ohne dass man Teilsequenzen des Inserts kennen muss. Solche Sequenzierprimer können dann entweder links oder rechts des Integrats binden und von dort aus eine Sequenzierung initiieren.



2. Möglichkeit: die PCR (Polymerase-Kettenreaktion)

Bei einer PCR wird die DNA durch Hitze einzelsträngig gemacht, anschließend binden Primer flankierend zu der unbekannten Region (Sequenz-bereiche rechts und links der unbekannten Region müssen also VORHER bekannt sein!) und eine DNA-Polymerase verlängert dann die Primer komplementär zum template-Strang. Diese Abfolge an Reaktionen wird bis zu 40x wiederholt, die Amplifikation des gewünschten Abschnitts erfolgt exponentiell, da in jeder Runde die Anzahl der Zielmoleküle verdoppelt wird.
→ Können Sie die Abbildung der PCR ein wenig besser beschriften?

→ Machen Sie sich zudem die fundamentalen Unterschiede zwischen einer PCR und einer Sequenzierungsreaktion klar! Viele (insbesondere Studierende benachbarter Fachrichtungen der Lebenswissenschaften :-) verwechseln das gerne mal...

	PCR	Sequenzierung
Amplifikation		
Primer		
Nukleotide		

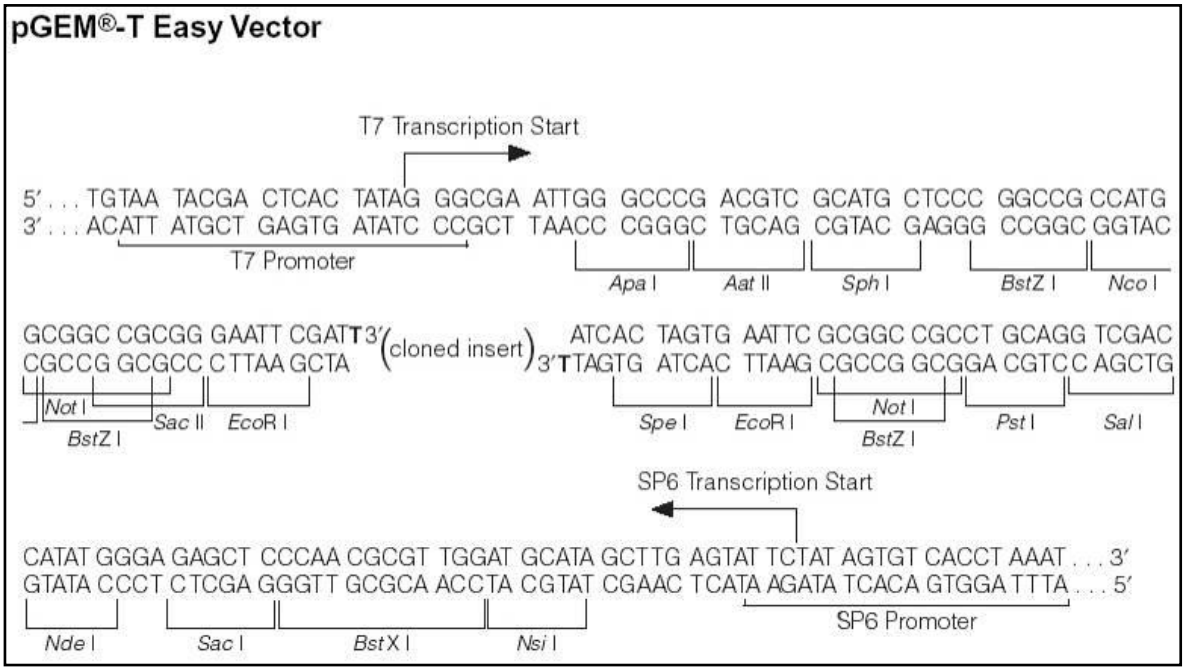
2. Was steckt im Klon? Auswertung moderner Sequenz-Chromatogramme

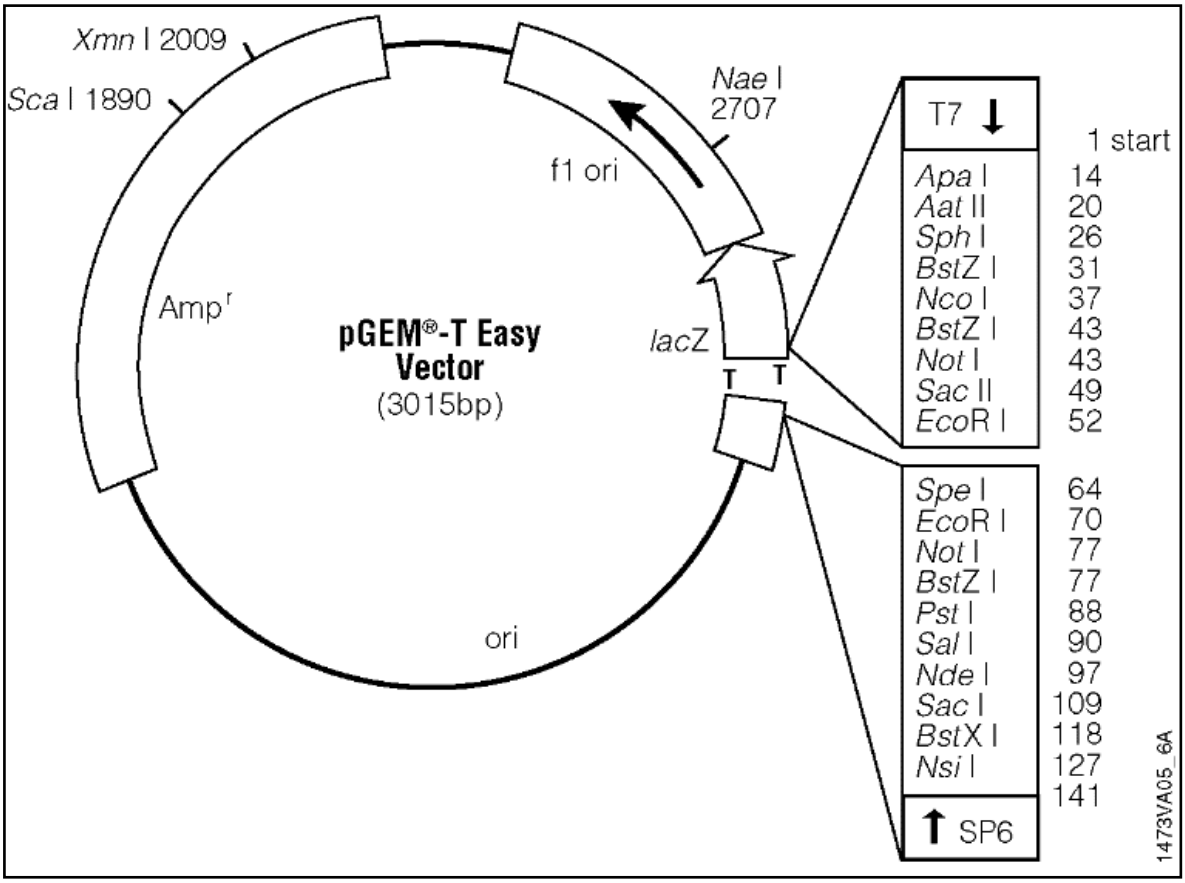
Sie werden zunächst das „Chromatogramm" einer modernen, fluoreszenzbasierten Sanger-Sequenzierung erhalten, das die Sequenz eines in einen Plasmidvektor klonierten Genabschnitts zeigt. Ihr Ziel ist es zunächst, per Auge „schlechte" Sequenzbereiche und Abschnitte der Vektor-DNA zu erkennen und aus der Sequenz zu entfernen. Mit der neu entschlüsselten Sequenz des Klon-Integrats werden Sie anschließend in Datenbanken nach ähnlichen Sequenzen suchen, um herauszufinden, um welches Gen es sich handelt.

Öffnen Sie die Datei „Klonierung1_T7.ab1" mit dem Shareware-Programm FinchTV[1]. Die Datei enthält die Sequenz eines Genabschnitts, der in den Plasmid-Vektor „pGEM T Easy" kloniert wurde. Wenn Sie sich die Vektor-Karte anschauen, wird klar, dass die Sequenz Vektoranteile enthalten wird (die natürlich entfernt werden müssen).

→ Sie wissen, dass die Sequenzierung mit dem Primer T7 gemacht wurde. Die Primersequenz lautet 5' - TAATACGACTCACTATAGGG - 3'. Suchen Sie die Bindungsstelle des Primers auf dem Vektor (Abbildung nächste Seite). Welchen DNA-Strang werden Sie im Chromatogramm-file also vor sich sehen? Welcher Strang dient als Matrize? (oben = Watson, unten = Crick)

[1] http://www.geospiza.com/Products/finchtv.shtml


pGEM®-T Easy Vector
T7 Transcription Start
5′ . . . TGTAA TACGA CTCAC TATAG GGCGA ATTGG GCCCG ACGTC GCATG CTCCC GGCCG CCATG
3′ . . . ACATT ATGCT GAGTG ATATC CCGCT TAACC CGGGC TGCAG CGTAC GAGGG CCGGC GGTAC
T7 Promoter
Apa I
Aat II
Sph I
BstZ I
Nco I
GCGGC CGCGG GAATT CGATT 3′
CGCCG GCGCC CTTAA GCTA
(cloned insert)
ATCAC TAGTG AATTC GCGGC CGCCT GCAGG TCGAC
3′TTAGTG ATCAC TTAAG CGCCG GCGGA CGTCC AGCTG
Not I
Sac II
EcoR I
BstZ I
Spe I
EcoR I
Not I
BstZ I
Pst I
Sal I
SP6 Transcription Start
CATAT GGGA GAGCT CCCAA CGCGT TGGAT GCATA GCTTG AGTAT TCTAT AGTGT CACCT AAAT . . . 3′
GTATA CCCT CTCGA GGGTT GCGCA ACCTA CGTAT CGAAC TCATA AGATA TCACA GTGGA TTTA . . . 5′
SP6 Promoter
Nde I
Sac I
BstX I
Nsi I


Xmn I 2009
Sca I 1890
Nae I
2707
f1 ori
Amp^r
pGEM®-T Easy
Vector
(3015bp)
lacZ
T T
ori
T7
1 start
Apa I 14
Aat II 20
Sph I 26
BstZ I 31
Nco I 37
BstZ I 43
Not I 43
Sac II 49
EcoR I 52
Spe I 64
EcoR I 70
Not I 77
BstZ I 77
Pst I 88
Sal I 90
Nde I 97
Sac I 109
BstX I 118
Nsi I 127
141
SP6
1473VA05_6A

Suchen Sie nun im Chromatogramm-File nach dem Beginn des Inserts (syn. Integrats) und schneiden Sie alle davor (=5`) liegenden Bereiche (=Vektor) weg (markieren + `Entf`). Suchen Sie per Auge und obiger Vektorkarte auch auf der anderen Seite des Inserts nach Vektorsequenzen und entfernen Sie diese manuell!

→ Wie lang ist die verbleibende Sequenz, also das „reine" Insert?

Speichern Sie das editierte Chromatogramm („Klonierung1_edit_Gruppennummer.ab1"). **Exportieren Sie** die verbleibende Sequenz in dem universell von vielen Computertools lesbaren „Fasta"-Format (File → Export → DNA-Sequence: FASTA → „Klonierung1_edit_Gruppennummer.seq"). Dieser Fasta-File enthält nur noch die eigentliche DNA-Sequenz, nicht mehr jedoch die originalen Fluoreszenzsignale („traces").

Öffnen Sie jetzt die Datei „Klonierung2_sp6.ab1". Diese zeigt eine Sequenzierung des gleichen Plasmids, diesmal allerdings mit dem Sp6-Primer (5' - ATTTAGGTGACACTATAG - 3').

→ Welchen DNA-Strang werden Sie jetzt als Sequenz sehen?

Gehen Sie beim Editieren genauso vor wie bei der bereits bearbeiteten Sequenzierung, **speichern Sie das Ergebnis als *.ab1 und *.seq-file**.

Sie erinnern sich, die beiden Stränge der DNA, die sie jetzt einzeln sequenziert haben, sollten wegen der Komplementarität der Basen eigentlich perfekt zueinander passen. Als nächstes wollen wir daher überprüfen, ob die beiden Sequenzierungen tatsächlich das gleiche Ergebnis liefern (das ist nicht immer so, da die Sequenzierung manchmal falsche oder unklare Signale liefert). Vergleichen Sie die beiden entstandenen Sequenzen und versuchen Sie, durch „Untereinanderschreiben" ein doppelsträngiges Insert-DNA-Molekül zu generieren (d.h., Sie machen ein „Alignment"). Als Hilfestellung erhalten Sie Teile der Sequenzen in ausgedruckter Form. Beachten Sie unbedingt die 5'→3'-Orientierung der Stränge und die Komplementarität der Basen!

Bsp: 5'-GAGTC...insert...TTGCA-3'
3'-CTCAG...insert...AACGT-5'

Wie Sie merken ist eine solche manuelle Alignierung selbst von nahezu perfekt zueinander passenden und relativ kurzen „forward"- und „reverse"-Sequenzierungen recht zeitaufwändig. Es gibt natürlich Programme, die dies in wesentlich schnellerer Zeit erledigen können. Eines dieser Programme ist Blast2Seq (zu finden auf der Seite des NCBI[2]). Geben Sie ihre beiden editierten Sequenzen (*_edit_Gruppennummer.seq) bei Blast2Seq ein und betrachten Sie das Ergebnis.

→ Was fällt Ihnen im Vergleich zu ihren von Hand alignierten Sequenzen auf?

→ Gibt es Stellen, an denen sich die forward- und reverse-Sequenzierungen widersprechen? Falls ja, suchen Sie diese Stellen in den Chromatogrammen und versuchen Sie eine Erklärung zu finden!

Nun wollen wir feststellen, um welchen Genabschnitt es sich bei den Sequenzierungen eigentlich handelt. Dazu verwenden wir den **BLAST-Algorithmus**, welcher eine Suchsequenz gegen eine Sequenz-Datenbank abgleicht und Alignments als Ergebnis ausgibt.

Gehen Sie also auf die Homepage des NCBI und folgen Sie dem Link zur Blast-Seite (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome).

→ Sie haben die Auswahl zwischen mehreren Algorithmen: welchen erachten Sie in unserem Fall als sinnvoll?

Folgen Sie dem Link zum ausgewählten Algorithmus und fügen Sie eine der beiden von Ihnen editierten DNA-Sequenzen ein.

→ Welche Datenbank wählen Sie und warum?

[2] http://www.ncbi.nlm.nih.gov/

Starten Sie nun die Blast-Suche. Das Ergebnis erscheint in wenigen Sekunden auf Ihrem Bildschirm:

- eine graphische Zusammenfassung der Ergebnisse gibt einen ersten Überblick über die Güte, Anzahl und Länge der Treffer.
- weiter unten finden sich die Beschreibungen der Datenbank-Treffer. Über die sogenannte „Accession number" ist jeder Datenbankeintrag direkt von der Startseite des NCBI aus zu erreichen, sie ist einzigartig für jede Sequenz. Die „Query coverage" gibt an, wie viel % der Gesamtlänge Ihrer Suchsequenz durch den Treffer abgedeckt sind. Der E-Value gibt an, mit welcher Wahrscheinlichkeit ein Treffer dieser Länge in dieser Datenbank durch Zufall zu erwarten wäre.
- Noch weiter unten findet man schließlich die Alignments mit allen notwendigen Informationen und Links.

→ Welches Ergebnis erhalten Sie? Erstaunt es Sie irgendwie, wenn Sie sich mal an die Zoologievorlesungen erinnern?

Wir wollen uns die **Genstruktur** dieses Gens genauer anschauen. Laden Sie hierzu bitte über die entsprechenden Links sowohl die genomische Sequenz als auch die mRNA-Sequenz des Gens aus der NCBI-Datenbank im FASTA-Format herunter.

→ Was erwarten Sie beim Vergleich dieser beiden Sequenzen? Wird ein Alignment einfach sein?

Zum Herunterladen der Sequenzen folgen Sie dem Link des ersten Blast-Treffers zur Sequenz des Chromosoms und suchen Sie dann über `Strg+F` nach dem Gennamen. Suchen Sie sich eine Transkript-Variante aus und klicken Sie auf die Links „gene" bzw. „mRNA". Auf den erscheinenden Seiten kann man sich über den Link „Display settings → FASTA" die Sequenz im gewünschten Format anzeigen lassen. Kopieren Sie die Sequenzen und speichern Sie diese unter einem sinnvollen Namen ab.

Um die Genstruktur zu visualisieren wenden wir das Tool „Spidey"[3] an. Dieser Algorithmus ermöglicht die Alignierung von Genom-DNA-Sequenz und mRNA-Sequenz. Wie Sie sehen können, gibt es hier neben der direkten Sequenzeingabe auch die Möglichkeit, die schon erwähnte „Accession number" zu verwenden. Kopieren Sie die genomische und die mRNA-Sequenz in die entsprechenden Eingabefelder, wählen Sie den entsprechenden Organismus und starten Sie das Alignment.

→ Was können Sie über die Genstruktur sagen? (Skizze)

3. Gendiagnostik durch Sequenzierung eines PCR-Produkts

Die PCR erlaubt es, innerhalb von etwa 2 Stunden jeden einmal bekannten Genomabschnitt aus einem anderen Individuum zu isolieren. Typischerweise wird in der Gendiagnostik also aus Blutzellen eines Probanden Genom-DNA isoliert und das interessierende GenX per PCR millionenfach amplifiziert. Die PCR-Amplifikate können dann direkt sequenziert werden. Im Folgenden sollen zwei derartige Sequenzierungen ausgewertet werden. Da bei diesem Versuch ja nicht kloniert wurde, können zur Sequenzierung auch keine Primer verwendet werden, die auf dem Vektor liegen. Daher zum Verständnis noch einmal die Frage…

→ Welche Primer wurden zum Sequenzieren benutzt? (Schauen Sie sich die Abb. der PCR noch einmal an)

Öffnen Sie die Dateien „PCR_for.ab1" und „PCR_rev.ab1" mit FinchTV. Editieren Sie die Dateien, indem Sie qualitativ minderwertige Bereiche entfernen und in unsicheren Bereichen „von Hand" nacheditieren. Speichern Sie Ihre Ergebnisse als „PCR_for_edit_Gruppennummer.ab1" und „PCR_rev_edit_Gruppennummer.ab1".

→ Fallen Ihnen Stellen in der Sequenz auf, an denen man nicht entscheiden kann, welches Nukleotid das „richtige" ist, weil es zwei gleich hohe Peaks an dieser Stelle gibt? Wie können Sie solche Stellen interpretieren?

[3] http://www.ncbi.nlm.nih.gov/spidey/

Tipps:

- die Sequenz stammt aus Homo sapiens
- verwenden Sie Blast zur Identifizierung des Gens X!
- für ganz schnelle: betrachten Sie in den Ergebnissen der Blast-Suche die Annotationen der gefundenen Genomsequenz, zu der Ihr GenX passt. Gibt es da einen auffälligen genetischen Unterschied („SNP" genannt)? Suchen Sie einmal in der Literaturdatenbank PubMed[4] mit dem Namen des GensX und dem Begriff „SNP": weshalb ist die sequenzierte Region für die Humangendiagnostik interessant?

4. Wie komme ich an eine bereits bekannte Sequenz? (Beispiel SARS Virus)

Suchen Sie auf der Ihnen nun schon bekannten NCBI-Seite nach der Genomsequenz des SARS Corona Virus. Stellen Sie hierzu die Suchauswahl von „All Databases" auf „Genome". Folgen Sie dem Treffer und schauen Sie sich die Annotation des Genoms an.

→ Was können Sie über das SARS-Genom sagen? (Größe, Anzahl der Gene etc.)

Wir wollen nun alle potentiellen **proteinkodierenden Gene** dieses Genoms finden. Dazu benutzen wir das Tool „**ORF Finder**"[5].

→ Was ist ein ORF (open reading frame)?

Wir wollen bei der Suche diesmal über die „Accession number" des Sequenzeintrags anstatt über die Sequenz selbst gehen. Kopieren Sie hierzu die GenBank Accession number und geben Sie sie in das entsprechende Feld auf der ORF-Finder-Seite ein. Starten Sie die Analyse.

[4] http://www.ncbi.nlm.nih.gov/pubmed
[5] http://www.ncbi.nlm.nih.gov/projects/gorf/

→ Warum gibt es 6 verschiedene Leserahmen?

→ Wie viele „open reading frames" findet das Vorhersageprogramm? Wie viele proteinkodierende Gene sind hingegen annotiert?

→ Finden Sie auf der ORF-Finder-Seite eine Möglichkeit, die Anzahl potentieller ORFs zu verringern? Was macht man?

Klicken Sie den längsten ORF an. Es erscheint nun die Nukleotid- mit zugehöriger Aminosäuresequenz. Im oberen Bereich der Seite erscheint zusätzlich die Möglichkeit, die Relevanz der ausgewählten ORF-Sequenz direkt per Blast-Suche zu überprüfen. Dabei wird ermittelt, ob es zu der potenziellen Proteinsequenz des ORFs ein bereits bekanntes, ähnliches Protein in den Sequenzdatenbanken gibt. Nutzen Sie diese Möglichkeit!

→ Um welches Protein handelt es sich?

5. „Wer schreibt, der bleibt": Suche nach biowissenschaftlicher Literatur

Die **„PubMed"-Literaturdatenbank**[6] bei NCBI ist eine der wichtigsten Internetadressen, wenn Sie recherchieren müssen, beispielsweise um Referenzen für eine wissenschaftliche Arbeit zu suchen oder sich über den aktuellen Stand der Forschung in einem bestimmtem Gebiet zu informieren. Bei PubMed findet man die Titel aller relevanten wissenschaftlichen Publikationen (meist mit Abstracts) sowie Links zu den Seiten der entsprechenden Zeitschriften. Durch ein ausgezeichnetes Suchsystem und viele nützliche Querverweise ist es häufig ein leichtes, schnell an die gewünschten Informationen zu einem Thema zu gelangen.

Tipps:

- durch setzen einer „wildcard" (*) können Sie mehrere Begriffe auf einmal suchen. Geben Sie zum Beispiel „phylogen*" als Suchbegriff ein, wird PubMed nach den Begriffen „phylogeny, phylogenomic, phylogenomics, phylogenetic, …" suchen.
- durch die Operatoren „AND, OR, NOT" können verschiedene Begriffe logisch miteinander verknüpft werden.
- Durch Klammersetzung können gezielt bestimmte Suchfelder wie Autor, Titel, Publikationsdatum, Journal etc. durchsucht werden. Beispiele:
  1. Hankeln[Author] AND globin[title]
     Alle Paper von Hankeln, die "globin" im Titel stehen haben
  2. Neuroglobin[title] AND "2004"[Publication Date] : „2010"[Publication Date]
     Alle Paper mit dem Wort "Neuroglobin" im Titel, die zwischen 2004 und 2010 veröffentlicht wurden.

Hier sind einige Übungen zur Literatursuche:

1. Suchen Sie in der PubMed-Datenbank nach der Publikation der SARS-Gesamtsequenz. Sie werden mehrere Publikationen finden. Warum gibt es verschiedene Publikationen zur Gesamtgenomsequenz?

2. Formulieren Sie eine Suchstrategie für: „allergies to eggs or peanuts". Wie lautet Ihre Such-Strategie? Wie viele Treffer gibt es?

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

3. Wann wurde das klassische Paper von Watson und Crick zur DNA-Struktur zuletzt noch einmal abgedruckt?

4. Wie viele Publikationen finden Sie, bei denen Thomas Hankeln Erstautor ist und in denen es um Globine geht?

5. Wie viele hingegen gibt es zu dem Thema, bei denen Thomas Hankeln unter den Autoren ist?