Korrelationsanalyse, Molekularbiologie und Zufälligkeit

Johannes Nübler

Mein Ziel ist es, zu untersuchen, ob die Reihenfolge der in der DNA verwendeten Aminosäuren zufällig ist, oder irgendwelchen Auswahlregeln unterliegt.

Im ersten, biochemischen, Teil wird dabei die Bedeutung und der Aufbau der DNA erläutert. Die DNA ist Träger der Erbinformation. Es handelt sich um ein langkettiges Molekül aus 4 chemischen „Buchstaben“. Eine andere Klasse chemischer Makromoleküle, die Proteine, sind in jedem Organismus u.a. für Regelung und Katalysation fast aller chemischer Reaktionen verantwortlich. Diese sind Wörter aus einem anderen, 20-buchstabigen, chemischen Alphabet. Der Zusammenhang zwischen der DNA und den Proteinen besteht darin, dass die Reihenfolge der in Proteinen verbauten Buchstaben mittels eines Übersetzungscodes von der DNA abgeschrieben wird.

Im zweiten, mathematischen, Teil wird nun untersucht, ob in der DNA die Reihenfolge der verwendeten vier Buchstaben völlig zufällig ist. Sowohl durch einfaches Abzählen, als auch durch Erzeugung gewisser fraktaler Bilder sieht man, dass das Paar „cg“ deutlich seltener vorkommt, als es einer zufälligen Auswahl entspräche.

1 Teil I: Biochemie

1.1 Was ist DNA

Aufbau aus 4 Nukleinsäuren (a, c, g, t) Die DNA ist der Träger der Erbinformation. Sie ist ein (sehr!) langes Makromolekül, das in jeder Zelle eines Organismus im Zellkern verstaut ist. In allen Zellen ist, unabhängig von der Funktion der Zelle als Muskel-, Haut-, Gehirn- oder Ei- oder Spermazelle, exakt die gleiche DNA gespeichert.

Sie sieht aus, wie eine gewundene Strickleiter (die „Doppelhelix“). Halbe Sprossen der Strickleiter bestehen aus jeweils einer von 4 Nukleinsäuren: Adenin, Cytosin, Guanin und Thymin (abgekürzt als a, c, g und t). Das sind 4 Stoffe, deren chemische Zusammensetzung prinzipiell völlig irrelevant ist – wie hoffentlich klar werden wird.

Es genügt, die Hälften der Strickleitersprossen zu betrachten, da dann die andere Hälfte eindeutig festgelegt ist. Gegenüber von Adenin liegt immer Thymin und gegenüber von Cytosin immer Guanin (Merkregel: die eckigen und die runden Buchstaben bilden ein Paar).

Die DNA kann also betrachtet werden, als ein langes Wort aus vier Buchstaben. Diese lineare Folge ist die einzige Information, die in der DNA steckt. Im Vergleich zu den später betrachteten Proteinen ist beispielsweise die dreidimensionale Form, in die sich das Molekül faltet, völlig irrelevant.

Universalität Die DNA ist bei allen Lebewesen, von Menschen über Säugetiere und Insekten bis zu Bakterien, genau gleich aufgebaut. Es werden die selben vier Nukleinsäuren verwendet. Sogar Viren, bei denen man sich – da sie keinen eigenen Stoffwechsel haben, sondern den des Wirtes parasitieren – fragen kann, ob sie Lebewesen sind, haben prinzipiell die gleiche DNA.

1.2 Was sind Proteine

Wozu werden sie gebraucht Proteine sind ebenfalls langkettige Makromoleküle (aus anderen Bausteinen als die DNA). Sie werden verwendet als

- Baustoffe: Membranen, ...

- Biokatalysatoren: Alle chemischen Reaktionen im Körper, von der Zersetzung der aufgenommenen Nahrung über Anlagerung des Sauerstoffes im Blut bis zu Muskelkontraktion, müssen auf der einen Seite sicher exotherm sein – schließlich soll Arbeit verrichtet werden (Muskel), bzw. die freiwerdende Energie in weiterverwertbarer Form gespeichert werden (Verdauung). Andererseits dürfen diese Reaktionen aber nicht spontan ablaufen, sobald die Reagenzien vorhanden sind – Muskeln sollen schließlich nicht immer DNA kontrahieren, wenn das Blut gerade Sauerstoff anschwemmt. Die chemische Lösung für dieses Problem sind Reaktionen mit einer gewissen Aktivierungsenergie, die von Katalysatoren bereitgestellt werden muss. Diese Katalysatoren sind durchweg Proteine.
- Hormone: Im Körper müssen an einer Stelle gemessene Informationen, z.B. Konzentration des Zuckers im Blut, der Blutdruck, Salzkonzentration, ... an anderer Stelle verarbeitet werden, z.B. in den Nieren, der Hypophyse,.... Falls das Signal zwar nicht schnell transportiert werden muss, aber an vielen Stellen im Körper benötigt wird, wird ein charakteristischer Stoff in den Flüssigkeitskreislauf des Körpers ausgeschüttet: Botenstoffe oder auch Hormone.

Aufbau aus 20 Aminosäuren Proteine sind – wie die DNA – Makromoleküle: Lange Ketten aus diesmal 20 verschiedenen chemischen Stoffen, den Aminosäuren. Im menschlichen Körper kommt eine Unzahl verschiedener Proteine zum Einsatz. Alle diese sind aber Wörter aus einem Alphabet mit 20 Buchstaben.

Bei Proteinen ist, im Gegensatz zur DNA, die sterische (räumliche) Struktur das (einzig) Wichtige.

Die Ketten aus Aminosäuren sind in einer ganz bestimmten Art räumlich zusammengefaltet. Der Austausch einer einzigen Aminosäure gegen eine andere, also die Veränderung eines einzigen Buchstabens im Wort, kann die „Bedeutung“ des Wortes in der Hinsicht völlig zerstören, dass das Protein sich räumlich anders faltet und so die Reaktion nicht mehr katalysieren kann, für die es bestimmt war.

1.3 (Protein-)Synthese

Was hat DNA mit Proteinen zu tun Die große Frage ist nun folgende: Angeblich ist ja die DNA die einzige Information, die bei Zeugung eines Organismus weitergegeben wird. Ein Individuum entwickelt sich aus einer einzigen Zelle: einer befruchteten Eizelle. In dieser sind zwar schon Proteine vorhanden, aber sicher wenige. Wo kommen die Proteine des erwachsenen Organismus her?
Die Antwort auf diese Frage ist:
Die meisten Proteine werden mittels eines komplizierten chemischen Apparates vom Körper selber gebaut. Es wird eine Aminosäure an die andere gehängt, und zwar *wird die Reihenfolge, in der das geschieht, von der DNA abgelesen*.

Nun hat aber das Alphabet der DNA-Wörter vier Buchstaben, das der Aminosäuren 20.

Deswegen werden jeweils Gruppen von drei Nukleinsäuren nach einem Code (das ist der vielzitierte „genetische Code“) in eine Aminosäure übersetzt. Diese werden – entsprechend der Reihenfolge der DNA – linear hintereinandergehängt. Das ist das Protein.

Es gibt $4^3 = 64$ verschiede Dreiergruppen aus vier Buchstaben, aber nur 20 verschiedene Aminosäuren. Deswegen ist der genetische Code degeneriert: Eine Aminosäure

kann durch mehr als ein Tripplet von DNA-Buchstaben „aufgeschrieben“ sein, und es werden auch alle 64 Tripplets verwendet.

Da wie oben gesagt, bereits die Änderung einer einzigen Aminosäure zur vollständigen Nichtfunktion eines Proteins führen kann, darf natürlich bei der Ablesung nicht geschlampt werden. Der chemische Apparat, der diese Proteinbiosynthese bewerkstelligt, ist ziemlich kompliziert, aber hier nicht weiter wichtig. Er funktioniert, und ist bis ins Detail bekann.

Hingegen ist nur in Anfangszügen bekannt, wie aus der Information für Proteine ein Organismus entsteht.

Universalität des genetischen Codes Wie die Struktur der DNA, so ist auch der zur Übersetzung verwendete Code bei allen Lebewesen der gleiche! Das heißt natürlich nicht, dass alle Lebewesen die gleichen Proteine benutzen, aber doch, dass bei allen Lebewesen die Proteine aus Aminosäuren aufgebaut sind.

2 Teil II: Mathematik

Ist die DNA eine Zufallsfolge?

Die Folge der Nukleinsäuren a, c, g, t in der DNA hat in dem Sinne eine „Bedeutung“, dass sie die Reihenfolge der Aminosäuren in Proteinen codieren, und diese Reihenfolge legt die dreidimensionale Struktur fest, in der sich ein Protein faltet.

Man könnte nun auf die Idee kommen, die Reihenfolge der Nukleinsäuren in der DNA sei sicher nicht einfach eine Zufallsfolge, da sie sehr viel Information trägt. Letzteres ist sicher richtig, aber wenn man es genau bedenkt, ist es nur eine These, zu behaupten, deswegen könne die Folge keine Zufallsfolge sein. Und wenn man etwas nicht weiß, schaut man nach...

2.1 Nukleinsäuren abzählen

Theorie: Die Mathematik sagt: Wenn zwei zufällige Ereignisse unabhängig voneinander stattfinden, so ist die Wahrscheinlichkeit für eine bestimmte Kombination $P(x, y)$ gleich dem Produkt der Wahrscheinlichkeiten der Einzelereignisse $P(x)$ und $P(y)$.
Test auf Unabhängigkeit: Ist $P(x, y) = P(x)P(y)$ erfüllt?
Falls die DNA eine Zufallsfolge aus vier Buchstaben ist, so sollten die „Wahlen“ zweier aufeinanderfolgender Buchstaben unabhängige Ereignisse sein. Beispielsweise sollte die Wahrscheinlichkeit, dass nach einem c ein g folgt, genauso groß sein wie das Produkt der Wahrscheinlichkeiten dafür, dass der erste Buchstabe ein c, der zweite ein g ist.

1. Wir zählen, wie häufig in einem Stück DNA die verschiedenen Nukleinsäuren vorkommen, und erhalten die Wahrscheinlichkeit dafür, dass ein beliebig herausgegriffener Buschstabe beispielsweise ein „c“ ist:
   $P(c)$ = Anzahl der gefunden c / Gesamtlänge des untersuchten DNA-Stückes
2. DNA zählen wir, wie häufig Paare vorkommen, und erhalten die Wahrscheinlichkeit dafür, dass ein beliebig herausgegriffenes Paar beispielsweise das Paar cg ist:
   $P(cg)$ = Anzahl der gefunden Paare cg / (Gesamtlänge des untersuchten DNA-Stückes –1)
3. Nun rechnen wir die Differenz aus zwischen der im Falle einer Zufallsfolge vorhergesagten Wahrscheinlichkeit für ein Paar und der tatsächlich vorgefundenen Wahrscheinlichkeiten für dieses Paar.

Durchführung: Diese Schritte müssen natürlich mit relativ großen Stichproben durchgeführt werden. Deswegen bietet sich der Computer an. Ich habe zunächst aus dem Internet einfach DNA-Sequenzen

heruntergeladen. Dafür gibt es frei verfügbare Datenbanken, wo man entweder den Namen eines Proteins wie z.B. Insulin oder Hämoglobin eingibt und die dieses Protein codierende DNA-Sequenz bekommt, oder ganze Chromosomen nimmt.
Die DNA-Sequenz für Hämoglobin-Alpha ist 454 Buchstaben lang, die für Insulin liefert ein 1882 Zeichen langes file. Diese ASCII-files habe ich nun an ein Java-Programm verfüttert, das obige drei Schritte ausführt. Die Ausgabe des Programms sieht so aus:

Gesamtanzahl: 1882 - INSULIN

a: 364	P(a): 19.3		
c: 595	P(c): 31.6		
g: 583	P(g): 30.9		
t: 340	P(t): 18.0		
aa: 52	P(aa): 2.7	P(a)*P(a): 3.7	daa: 0.9
ac: 113	P(ac): 6.0	P(a)*P(c): 6.1	dac: 0.1
ag: 148	P(ag): 7.8	P(a)*P(g): 5.9	dag: -1.8
at: 51	P(at): 2.7	P(a)*P(t): 3.4	dat: 0.7
ca: 145	P(ca): 7.7	P(c)*P(a): 6.1	dca: -1.5
cc: 166	P(cc): 8.8	P(c)*P(c): 9.9	dcc: 1.1
cg: 126	P(cg): 6.6	P(c)*P(g): 9.7	dcg: 3.0
ct: 157	P(ct): 8.3	P(c)*P(t): 5.7	dct: -2.6
ga: 124	P(ga): 6.5	P(g)*P(a): 5.9	dga: -0.5
gc: 212	P(gc): 11.2	P(g)*P(c): 9.7	dgc: -1.4
gg: 174	P(gg): 9.2	P(g)*P(g): 9.5	dgg: 0.3
gt: 73	P(gt): 3.8	P(g)*P(t): 5.5	dgt: 1.7
ta: 43	P(ta): 2.2	P(t)*P(a): 3.4	dta: 1.2
tc: 104	P(tc): 5.5	P(t)*P(c): 5.7	dtc: 0.1
tg: 134	P(tg): 7.1	P(t)*P(g): 5.5	dtg: -1.5
tt: 59	P(tt): 3.1	P(t)*P(t): 3.2	dtt: 0.1

In der ersten Spalte stehen die gezählten Häufigkeiten für einzelne Buchstaben bzw. für Paare. In der zweiten steht die gezählte Wahrscheinlichkeit für Paare, in der dritten Spalte die im Falle der Zufälligkeit erwartete Wahrscheinlichkeit. Die letzte Spalte ist die interessante: Dort steht die unter Punkt drei definierte Differenz. Die Angaben der letzten drei Spalten sind in Prozent.

Zwei Abweichungen stechen besonders hervor: Das Paar cg kommt statt in den vorhergesagten 9,7% aller Fälle nur in 6,6% vor, kommt also seltener als erwartet vor. Das Paar ct tritt in der Sequenz häufiger auf als es sollte, wenn die Sequenz völlig zufällig wäre, und zwar sind 8,3% aller Paare das Paar ct, während es nur 5,7% sein sollten.

Schauen wir uns das ganze noch mal für eine etwas größere Stichprobe an. Unter „Chromosom 21“ habe ich eine Aminosäuresequenz mit 281387 Buchstaben bekommen:

Gesamtanzahl: 281378 - CHROMOSOM 21

a: 88881	P(a): 31.5		
c: 56520	P(c): 20.0		
g: 56122	P(g): 19.9		
t: 79855	P(t): 28.3		
aa: 30629	P(aa): 10.8	P(a)*P(a): 9.9	daa: -0.9
ac: 15220	P(ac): 5.4	P(a)*P(c): 6.3	dac: 0.9
ag: 19449	P(ag): 6.9	P(a)*P(g): 6.3	dag: -0.6
at: 22314	P(at): 7.9	P(a)*P(t): 8.9	dat: 1.0
ca: 20948	P(ca): 7.4	P(c)*P(a): 6.3	dca: -1.0
cc: 13567	P(cc): 4.8	P(c)*P(c): 4.0	dcc: -0.7
cg: 2763	P(cg): 0.9	P(c)*P(g): 4.0	dcg: 3.0
ct: 18418	P(ct): 6.5	P(c)*P(t): 5.7	dct: -0.8
ga: 16776	P(ga): 5.9	P(g)*P(a): 6.3	dga: 0.3
gc: 11229	P(gc): 3.9	P(g)*P(c): 4.0	dgc: 0
gg: 14221	P(gg): 5.0	P(g)*P(g): 3.9	dgg: -1.0
gt: 13150	P(gt): 4.6	P(g)*P(t): 5.6	dgt: 0.9
ta: 19280	P(ta): 6.8	P(t)*P(a): 8.9	dta: 2.1
tc: 15707	P(tc): 5.5	P(t)*P(c): 5.7	dtc: 0.1
tg: 18914	P(tg): 6.7	P(t)*P(g): 5.6	dtg: -1.0
tt: 24773	P(tt): 8.8	P(t)*P(t): 8.0	dtt: -0.7

Wieder kommt das Paar cg seltener vor als es sollte, und zwar diesmal sogar mit einer extrem großen relativen Abweichung vom im Falle der Zufälligkeit vorhergesagten Wert: In nur 0,9 statt 4 Prozent aller Fälle.

Die einzige andere Abweichung die über einem Prozent liegt, ist bei dem Paar ta. Hier allerdings ist die relative Abweichung nicht so stark: 6,8 statt 8,9 Prozent.

Ich habe das Programm mit noch drei anderen Proteinen und einem weiteren Chromosom gefüttert, und die Ergebnisse lassen sich wie folgt zusammenfassen:

Die einzigen Abweichungen vom unter Annahme der Zufälligkeit vorhergesagten Wert, die jedes Mal größer als ein Prozent waren, fanden sich bei den Paaren „ta“ und „cg“. Beide kommen seltener vor. Die Abweichung bei cg erscheint in sofern sehr viel signifikanter, als die relative Abweichung deutlich größer ist. Das Paar kommt über drei mal seltener vor, als es sollte.

Fazit: Betrachtet man nur Paare, so erscheint die Folge der Aminsäuren in der DNA weitgehend zufällig, mit der Ausnahme, dass auf ein c sehr selten ein g folgt.

2.2 Zufällige Iterierte Funktionensysteme

Es soll eine Methode gezeigt werden, mit der man sehr schön graphisch sieht, dass auf ein „c“ sehr selten ein „g“ folgt.

Selbstähnliche Bilder, Erzeugung und Beispiele Selbstähnliche Bilder lassen sich durch einfache Iterationsvorschriften erzeugen. Wohl jeder hat das Sierpinski-Dreieck schon einmal gesehen. An diesem Beispiel soll eine Methode vorgestellt werden, solche Bilder zu erzeugen.

Wir malen uns ein leeres Dreieck und benennen die drei Ecken irgendwie, z.B. 1,2,3. Wir starten mir irgendeinem Punk (in der Tat ist es egal, ob er innerhalb oder ausserhalb des Dreiecks liegt) und würfeln mit gleichen Wahrscheinlichkeiten eine 1,2 oder 3. Zu jeder gewürfelten Zahl befolgen wir folgende Regel:

„Halbiere die Stecke zwischen dem Punkt und der Ecke des Dreiecks, die die erwürfelte Nummer trägt. Dort wird der neue Punkt gemalt.“

Hier ist diese Methode vier Schritte lang vorgeführt. Man kann sich recht einfach überlegen (und sieht es diesen vier Bildern mit etwas gutem Willen auch schon an), dass man das Sierpinskidreieck erhält: Liegt der Punkt einmal auf dem Dreieck, so tun dies alle Folgenden, und startet man mit einem Punkt, der nicht darauf liegt, so konvergiert doch die Folge auf das Dreieck. Das Sierpinskidreieck ist ein Attraktor dieses iterierten Systems.

Um zu verstehen, wie man diese Methode der Erzeugung selbstähnlicher Bilder benutzen kann, um zu testen ob die Folge der Nukleinsäuren in der DNA eine Zufallsfolge ist oder nicht, wenden wir die gleiche Methode auf ein Quadrat an. Wir benennen alle vier Ecken des Quadrates irgendwie, z.B. mit den Buchstaben a, c, g, t – den Abkürzungen für die Nucleinsäuren. Wir starten mit irgendeinem Punkt, und verfahren

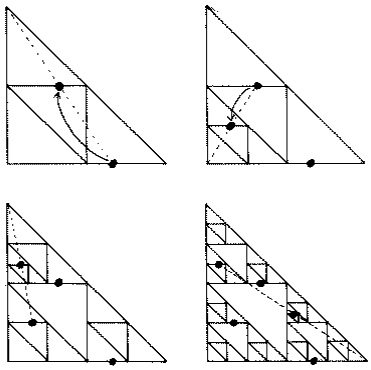

Abbildung 1.

genau wie bei dem Sierpinskidreieck: Falls die nächste Nucleinsäure Cytosin (Symbol c) ist, halbieren wir die Strecke zwischen dem aktuellen Punkt und derjenigen Ecke des Quadrates, die mit c bezeichnet ist und malen an diese Stelle einen Punkt. Dann sehen wir uns die nächste Nukleinsäure an und machen dasselbe mit dem eben gemalten Punkt, ...

Was erhält man?

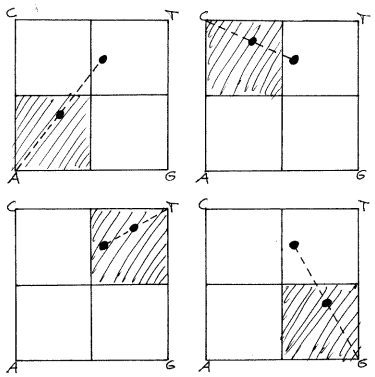

Abbildung 2.

Wird auf einen Punkt, der irgendwo im Quadrat liegt, Regel „a“ angewendet, so landet der nächste Punkt irgendwo im linken unteren Viertel. War die Nukleinsäure hingegen Cytosin (Symbol c) , so landet de Punkt irgendwo im linken oberen Viertel, usw., siehe Bild. Sind alle Nukleinsäuren

gleich häufig, so erhält man ein gleichförmig statistisch graues Quadrat.
Was passiert, falls die Folge keine Zufallsfolge ist?
Zunächst überlegen wir uns, was passiert, falls zwar keine Einschränkungen vorliegen, wie „nach einem c kommt kein g“, aber die Häufigkeiten für die vier Buchstaben nicht jeweils $1/4$ ist. Falls beispielsweise das „a“ deutlich seltener vorkommt als in $1/4$ aller Fälle, so wird klarerweise die linke untere Ecke weniger geschwärzt. Da aber nun der Punkt seltener in der linken unteren Ecke ist, wird auch die linke untere Ecke jedes der verbleibenden drei Teilquadrate weniger geschwärzt, denn um dorthin zu kommen, muss der Punkt ja vorher in der linken unteren Ecke gewesen sein, was ja eben nur selten der Fall ist.
Hier ist als Beispiel ein Bild erzeugt von einer Folge von vier Buchstaben mit folgenden Eigenschaften:

- Die vier Buchstaben kommen so oft vor, wie es den Häufigkeiten der vier Nukleinsäuren auf Chomosom 21 entspricht:
  1. A: 31 %
  2. C: 20 %
  3. G: 20 %
  4. T: 29 %
- Davon abgesehen ist die Folge eine Zufallsfolge (oder eine Quasi-Zufallsfolge, vom Computer generiert)

Die Ecken wurden benannt wie oben, also links unten a, links oben c, ... Man sieht deutlich, dass die a- und t- Ecken eines jeden Teilquadrates mehr geschwärzt werden, als die c- und g-Ecken.

Nun überlegen wir uns, was passiert falls die Einschränkung „nach einem c kommt nie ein g“ vorliegt: Bei Anwendung der Regel „c“ kommt der Punkt in der linken oberen Ecke zu liegen. Käme danach ein „g“, so würde der Punkt in die linke obere Ecke des rechten unteren Teilquadrates springen.

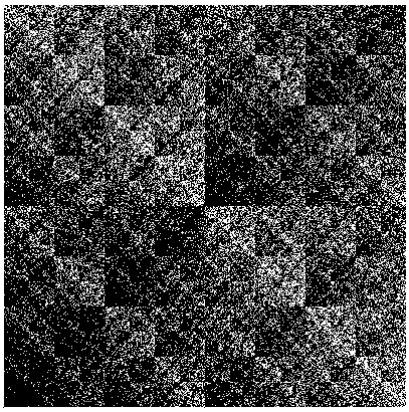

Abbildung 3.

Da dies („g nach c“) die einzige Möglichkeit ist, wie der Punkt in dieses Teilquadrat zu liegen kommen kann, bleibt es also leer. (linkes Bild)

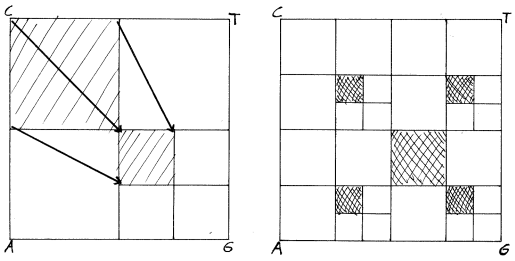

Abbildung 4.

Da es aber leer bleibt, bleiben auch die weiteren vier kleinen im rechten Bild eingezeichneten Teilquadrate leer, denn um in diese zu gelangen müsste der Punkt vorher im größeren Teilquadrat gelegen haben – was er ja eben nie tut. Das ergibt eine ganze weitere Hierarchie von immer kleineren Teilquadraten, die im Falle der Einschränkung alle leer bleiben.

Anwendung auf die DNA Nun haben wir alles zusammen, um eben diese Methode auf eine echte Nucleinsäurefolge anzuwenden. Das Programm liest aus einem Textfile (einer Nukleinsäurekette) einen Buchstaben nach dem anderen ein, befolgt die zu diesem Buchstaben gehörige Regel und malt

einen Punkt, wo es die entsprechende Regel fordert. Das Ergebnis ist folgendes:

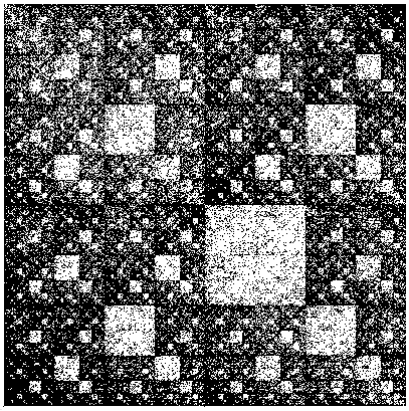

Abbildung 5.

Das Bild besteht aus den 281378 Punkten der Nukleinsäuresequenz von Chromosom 21. Man sieht sehr schön und auf einen Blick, dass die Teilquadrate, um die es oben ging, ziemlich leer bleiben. Mit anderen Worten, man sieht dem so erzeugten Bild sofort an, dass im betrachteten DNA-Strang nach der Nukleinsäure Cytosin sehr selten Guanyn folgt. Die relativ wenigen Punkte in den eigentlich „verbotenen“ Bereichen sind die 0,9 % aller Punkte, die doch „cg“ sind. Dass der Hintergrund nicht gleichmässig grau ist, liegt daran, dass nicht alle vier Nukleinsäuren gleich häufig, sondern mit der oben genannten Wahrscheinlichkeit vorkommen. Der Hintergrund ist genau der des obigen Bildes, wo extra diese Häufigkeiten respektiert wurden, mit einer ansonsten zufälligen Wahl.