Daniel Hoffmann. Foto: Timo Bobert

*Die Evolutionstheorie ist die grundlegende Beschreibung der Entwicklung biologischer Vielfalt. Ihre mathematische Formulierung hat viele Wissenszweige befruchtet und wird immer breiter eingesetzt. Mathematische Modelle der Evolution helfen beispielsweise bei der Entwicklung antiretroviraler Therapien und beim Design von Biomolekülen.*

# Evolution: Von der Biologie zur Mathematik und wieder zurück

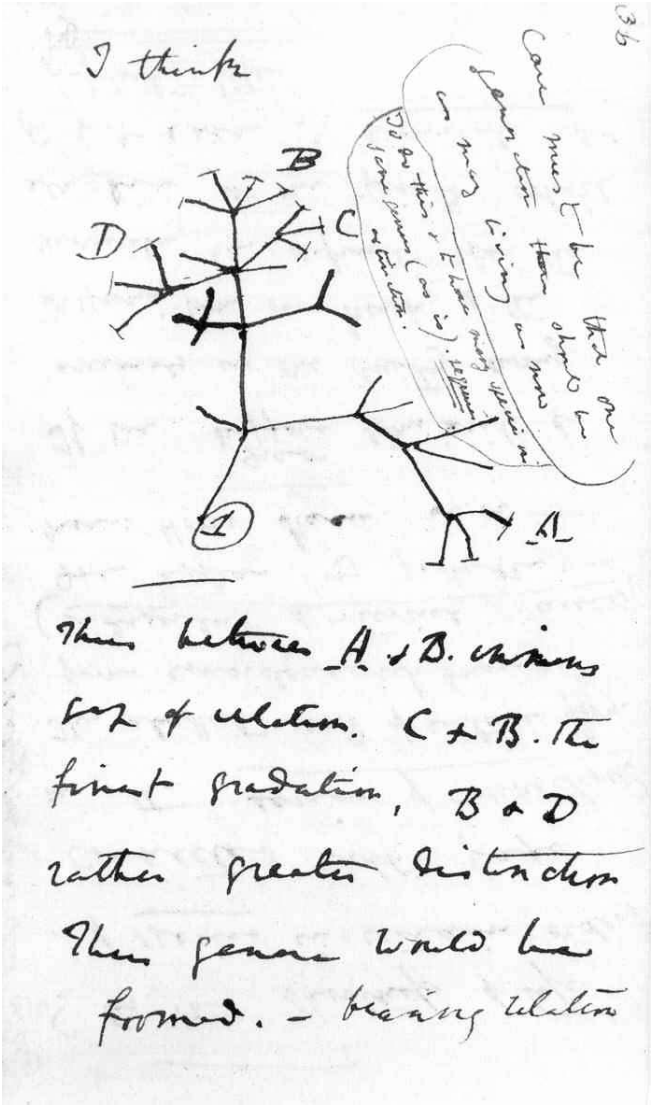
## Die Ergebnisse der Evolutionstheorie sind heute nicht nur Naturbeschreibung, sondern auch mathematisches Werkzeug

Von Daniel Hoffmann

**Die nichtmathematischen Anfänge**

In seinen Memoiren schreibt Charles Darwin:[1] "I attempted mathematics, and even went during the summer of 1828 with a private tutor to Barmouth, but I got on very slowly. The work was repugnant to me, chiefly from my not being able to see any meaning in the early steps of algebra. This impatience was very foolish, and after years I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics, for men thus endowed seem to have an extra sense." Tatsächlich hat Darwins fehlende mathematische Bildung ihn und seinen Zeitgenossen Alfred Russel Wallace nicht davon abgehalten, ein grundlegendes, seiner Natur nach mathematisches Muster in der Entwicklung des Lebens zu entdecken und dieses Muster in einer eleganten Theorie zu beschreiben, der Evolutionstheorie.[2] Man kann vermuten, dass ironischerweise gerade Darwins fehlende formale Mathematikkenntnisse dazu geführt haben, dass seine

Theorie so populär geworden ist wie kaum eine zweite, denn Darwin formulierte sie nicht mit Hilfe mathematischer Symbole, sondern in gewöhnlichem Englisch und mit einfachen Bildern (Abb. 1), sodass die Grundzüge der Theorie von vielen auf Anhieb verstanden wurden. Die Evolutionstheorie lässt sich folgendermaßen zusammenfassen (Abb. 2). Eine Population von Individuen verschiedener Arten hat Kinder (*Replikation*), die meist sehr ähnlich zu ihren Eltern sind. Einige der Kinder zeigen dagegen Änderungen (*Mutationen*) gegenüber ihren Eltern. Manche Eltern pflanzen sich stärker fort als andere (*Selektion*), weil sie besser angepasst sind an die jeweiligen Umweltbedingungen, oder wie es in der Evolutionstheorie genannt wird: weil sie eine höhere *Fitness* haben. Die Kinder einer Generation sind jeweils die Eltern der nächsten Generation, sodass der Prozess eine zyklische Aufeinanderfolge von Replikation, Mutation und Selektion ist. Die Population kann sich auf diese Weise dynamisch entwickeln, und zwar nicht nur ihrer Größe nach, sondern auch was die Verteilung über die Arten angeht, denn verschiedene neue Arten entstehen aus alten Arten, andere Arten sterben aus. Etwa in dieser Form beschrieb Darwin die Evolution, für deren Gültigkeit er zuvor zahlreiche Belege gesammelt hatte, nicht zuletzt auf seiner mehrjährigen Expedition auf dem Forschungsschiff Beagle. Manche Aspekte der Evolution blieben auch Darwin ein Rätsel, zum Beispiel warum in der Entwicklung einer Population sich durch Mischung nicht ein einziger Phänotyp herausbildet. Dass dies schon mit einfachen mathematischen Mitteln erklärbar ist, zeigten 1908 der Mathematiker G. H. Hardy[3] und der Arzt W. Weinberg,[4] aufbauend auf den Arbeiten von Mendel.

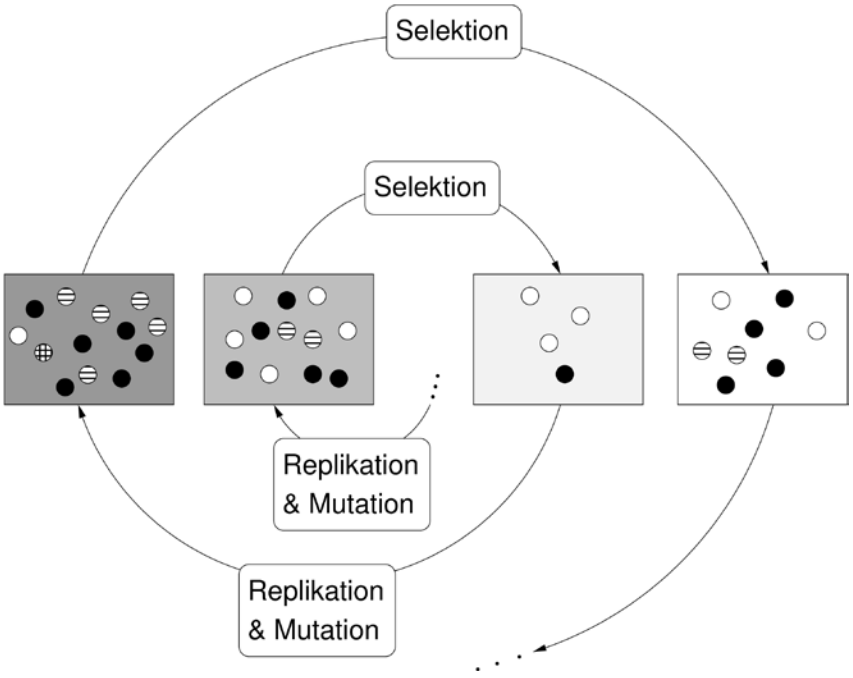



(1) Darwins erste dokumentierte Skizze eines evolutionären Baums (First Notebook on Transmutation of Species, 1837).
Quelle: http://en.wikipedia.org/wiki/Charles_Darwin

## Der zweite Durchbruch für die Evolutionstheorie

Trotz dieser Fortschritte war noch lange nach Darwin unverstanden, *wie* die Information über die Eigenschaften eines Individuums von den Eltern ererbt werden. Der Durchbruch dazu gelang in der Mitte des zwanzigsten Jahrhunderts. Zu dieser Zeit hatten sich die Hinweise verdichtet, dass die Information in einem langen Kettenmolekül gespeichert ist, der Deoxyribonukleinsäure (DNA), die aus vier Sorten von Bausteinen besteht, den Monophosphaten von Adenosin, Cytosin, Guanosin und Thymidin, häufig abgekürzt als A, C, G und T. Schließlich bündelten Watson und Crick die experimentellen Vorarbeiten vieler Wissenschaftler in einem geometrischen, also wiederum mathematischen Modell der Struktur der DNA, der berühmten Doppelhelix.[5] Damit wurden die molekularen Grundlagen der Informationsübertragung von Eltern auf Kinder verstehbar. Das Strukturmodell wurde so gedeutet, dass die Erbinformation in der Reihenfolge der vier verschiedenen Sorten von Bausteinen der molekularen Kette besteht, vergleichbar einer Reihe von Buchstaben oder Zeichen, die einen informationstragenden Text bilden. Der Text, das so genannte *Genom*,

wird von der biologischen Maschinerie jeder Zelle abgelesen und wie ein Plan zum Aufbau und Management des Organismus verwendet; beispielsweise steckt der Bauplan jedes Proteinmoleküls verschlüsselt in diesem Text. Das Genom kann, wie ein gewöhnlicher Text auch, kopiert und an Nachkommen weitergegeben werden. Replikation ist also in diesem Bild die Herstellung von Kopien der Elternzeichenketten. Auch der Mutation kann ein molekularer Vorgang zugeordnet werden, denn im Allgemeinen kommt es bei der Fortpflanzung nicht zu einer genauen Kopie der Bausteine des Elterngenoms, sondern zu einer geänderten Version, weil Fehler im Kopierprozess auftreten, oder weil die Kindergenome Mischungen der Elterngenome sind. Wie bereits angedeutet, sind die eben geschilderten molekularen Vorgänge leicht übertragbar auf mathematische Operationen auf abstrakten Zeichenketten. In diesem Sinne war die Doppelhelix auch ein Durchbruch im Hinblick auf die mathematische Beschreibung der Evolution.




(2) Schema der Evolution. Replikation und Mutation: Eine Population von Individuen (Kreise) unterschiedlicher Spezies (gefüllt, leer, gemustert) repliziert, wobei auch Mutanten entstehen, also Individuen mit neuen Eigenschaften. Selektion: In Abhängigkeit von ihren Eigenschaften erreichen mehr oder weniger Individuen die nächste Replikation. Der Prozess aus Replikation, Mutation und Selektion wiederholt sich über viele Zyklen, wobei in jedem Zyklus eine neue Population entsteht, die neuartige Mutanten enthalten kann. Über die Zeit können sich die Randbedingungen der Population ändern (verschiedene Grautöne) wodurch sich die Fitness der Spezies ändert.

**Eine Theorie, die viele Fragen beantwortet**

Die Erkenntnis, dass sich die Erbinformation als Zeichenkette beschreiben lässt, hat die Evolutionsforschung beflügelt und ihr neue Anwendungsfelder erschlossen. Neben der klassischen Anwendung auf die Entstehung neuer biologischer Arten, gehören dazu das Verständnis wichtiger Aspekte von Krebserkrankungen und Infektionen, der Entstehung menschlicher Sprachen, sowie die Entwicklung einer großen Gruppe von Methoden zur mathematischen Optimierung. Einen Überblick geben zum Beispiel die Werke von Nowak[6] oder Deb.[7] Im Folgenden soll auf zwei Anwendungen näher eingegangen werden: Wirkstoffresistenz beim humanen Immundefizienzvirus (HIV) und Optimierung von Biomolekülen.

**Evolution von HIV im Laufe einer Infektion**

Im Körper von Infizierten tritt HIV in großer Zahl auf und in vielen, immer wieder neuen Mutanten. Das Virus befällt Zellen des Immunsystems, nutzt diese Zellen zur eigenen Replikation und tötet sie schließlich, was langfristig zum Immunversagen und zum Tod des Infizierten führt. Umgekehrt wird das Virus auch von Zellen des Immunsystems bekämpft, denn Moleküle auf der Oberfläche der Viruspartikel werden von Immunzellen erkannt, die dann Viruspartikel vernichten. Neben diesem Selbstschutz des Körpers, der auf Dauer leider nicht ausreicht, werden gegen das Virus auch Medikamente eingesetzt, die sich an die viralen Proteinmoleküle lagern und dadurch direkt oder indirekt deren Funktion hemmen und auf diese Weise auch die Virusreplikation vermindern. Aus der Vielfalt der viralen Mutanten werden solche bevorzugt replizieren, die vom Immunsystem nicht erkannt werden und die resistent sind gegen Medikamente. Allerdings können die viralen Proteine nicht beliebig mutieren, da sie für das Virus eine wohldefinierte biologische Funktion ausüben, die Voraussetzung für die Replikation ist, und die durch Mutationen leicht zerstört werden kann. Die Erhaltung der Replikationsfähigkeit liefert also einen weiteren, intrinsischen Selektionsdruck. Damit sind die wesentlichen Voraussetzungen für Evolution gegeben: die Population mutierender und sich replizierender Viren und der Selektionsdruck des Immunsystems und der Medikamente. Tatsächlich wird beobachtet, dass HIV schnell Resistenzen

entwickelt. Typischerweise setzen sich unter Gabe eines bestimmten Medikaments HIV-Varianten durch, die Mutationen aufweisen, die für das jeweilige Medikament charakteristisch sind und Resistenz gegen dieses Medikament vermitteln. Dies passiert so verlässlich, dass die Liste dieser Resistenzmutationen im dominierenden Virusstamm eines Patienten als Basis für therapeutische Entscheidungen dient: je nachdem welche Mutationen vorliegen, werden bestimmte Medikamente nicht mehr gegeben, während andere noch wirksam sind.

**Wie funktioniert eine Resistenzmutation?**

Der oben eingeführte Begriff der Resistenzmutation soll noch etwas ausgeleuchtet werden, um dem Leser einen Eindruck von der Komplexität des Problemfeldes zu geben, auf dem wir uns bewegen. Unter den Medikamenten gegen HIV sind zum Beispiel Hemmstoffe gegen das virale Protein HIV-Protease (PR), so genannte Protease-Inhibitoren (PIs). PR ist ein Enzym, das in der Reifung des Virus eine entscheidende Rolle spielt, indem es aus einer Vorstufe des Virus, dem so genannten Polyprotein, virale Proteine herausschneidet und dadurch aktiviert. Das Polyprotein wird dazu in einen Kanal der PR (Abb. 3 oben) eingelegt und dann an speziellen Stellen geschnitten. Ist der Kanal durch einen Hemmstoff blockiert, so können die Schnitte nicht stattfinden (Abb. 3 Mitte). Dieser Mechanismus erklärt die Wirkung der PIs. Es ist klar, dass die beschriebene Hemmung durch PIs die Replikation des Virus stark behindert. Wenn man dem Patienten PIs verabreicht, werden daher vor allem solche Varianten des Virus sich vermehren können, bei denen der PI nicht fest genug im Kanal des Enzyms haftet, zum Beispiel in dem eine Mutation den Kanal verformt (Abb. 3 unten). Andererseits kann eine Verformung des Kanals dazu führen, dass das Polyprotein nicht mehr in den Kanal eingelegt und geschnitten werden kann, was auch wieder die Replikationsfähigkeit des Virus vermindern würde. Es werden also insgesamt solche PR-Mutanten selektiert, die Polyprotein schneiden, die aber nicht effektiv durch PIs gehemmt werden.

**Mutagenetische Bäume**

Bei der Evolution im Patienten löst das Virus ein Optimierungsproblem: Jene Mutanten setzen sich durch, die unter den gegebenen Bedingungen (Hemmstoffe, etc.) optimal replizieren. Man beobachtet, dass diese optimalen Virusmutanten nicht in einem Schritt entstehen, sondern dass sich über die Zeit eine Serie von Mutationen ergibt, man sieht also *Mutationsdynamik*. Das Verständnis der Mutationsdynamik kann dabei helfen, bessere Therapien zu entwickeln. Von einer langfristig wirksamen Therapie erwartet man beispielsweise, dass der Weg des Virus zum Optimum möglichst lang ist und über Zwischenzustände führt, die für das Virus ungünstig sind.

Kann man für ein bestimmtes Medikament die Abfolge der Mutationen genau vorhersagen? Das ist leider nicht der Fall, zum Beispiel ist aus Beobachtungen bekannt, dass auf eine Mutation verschiedene, sich ausschließende Mutationen folgen können. Ein einfaches mathematisches Modell, das ein solches Verhalten wiedergeben kann, ist ein sogenannter Baum, der in diesem Fall *mutagenetischer Baum* genannt wird.[9] Bäume sind eine spezielle Klasse von *Graphen*, das heißt mathematische Verknüpfungsmuster, in denen Knoten über Kanten miteinander verbunden sind. Ein Baum ist ein gerichteter Graph, also ein Verknüpfungsmuster, das sich von einem Wurzel-Knoten über verschiedene Verzweigungen zu Blatt-Knoten erstreckt. Abbildung (4) zeigt einen mutagenetischen Baum, wie er sich aus Daten von Patienten ergab, die mit dem Medikament Zidovudine behandelt wurden. Der Baum beginnt beim Wildtyp des Virus als Wurzel-Knoten und setzt sich über mehrere Kanten bis zu den beiden Blättern *210W* und *67N* fort. Dieser Baum sagt also zwei verschiedene Abfolgen von Mutationen voraus: $WT \rightarrow 70R \rightarrow 215F,Y \rightarrow 41L \rightarrow 210W$ oder $WT \rightarrow 70R \rightarrow 219E,Q \rightarrow 67N$. Welches dieser beiden Muster in einem speziellen Fall auftritt, kann nicht mit Sicherheit gesagt werden, sondern es handelt sich um einen probabilistischen Graphen. Jede Mutation tritt mit einer bestimmten bedingten Wahrscheinlichkeit auf, die als Zahl neben der jeweiligen Kante steht. Also entwickelt sich unter der Bedingung, dass *70R* aufgetreten ist, mit einer Wahrscheinlichkeit von 0.46 die Mutation *215F,Y* und mit einer Wahrscheinlichkeit von 0.43 die Mutation *219E,Q*. Die bedingten Wahrscheinlichkeiten gewinnt man aus der Beobachtung von Mutationshäufigkeiten in Patienten.

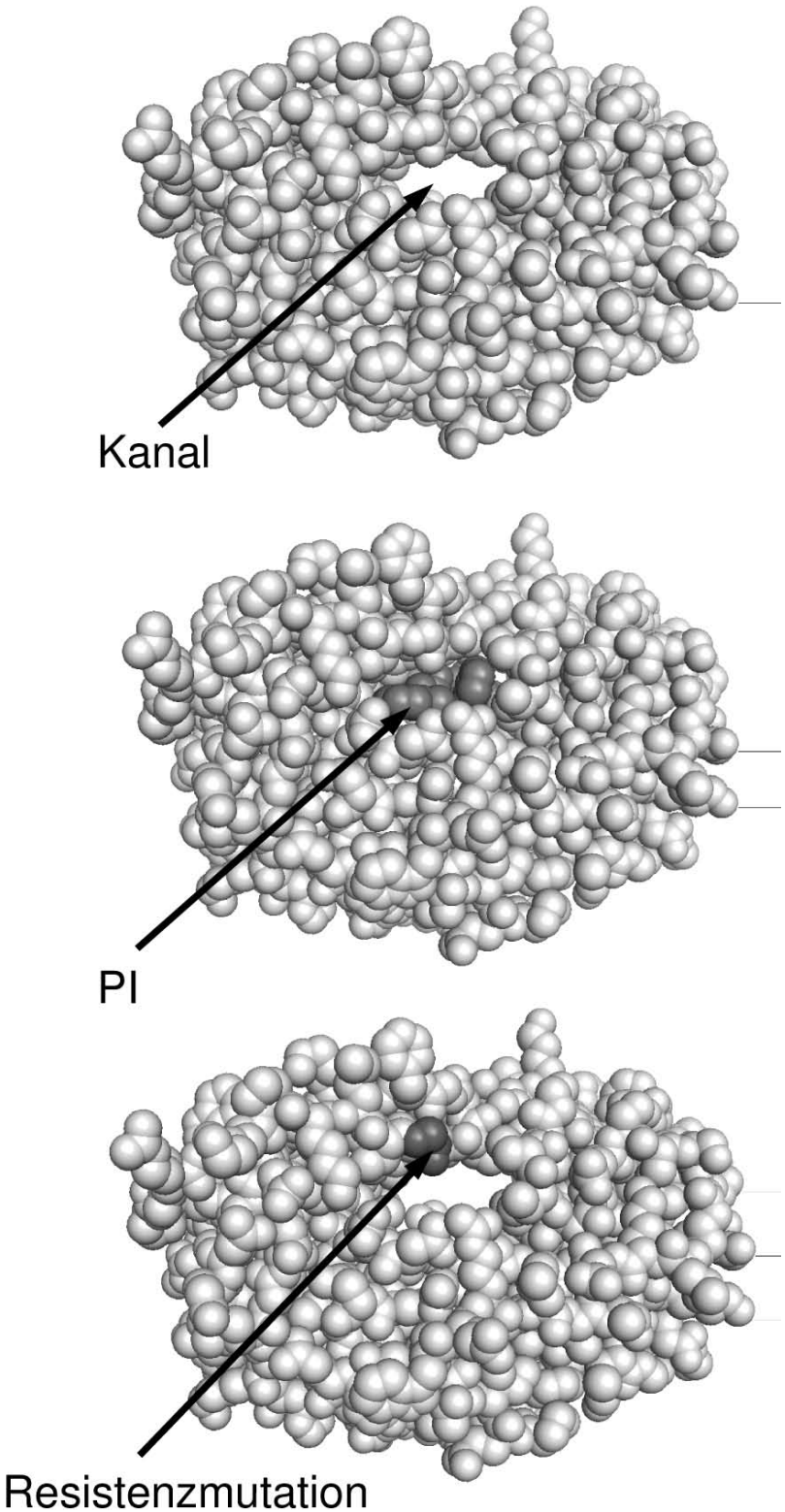
Ein mutagenetischer Baum wie in Abbildung (4) gezeigt ist ein probabilistisches Modell für die Evolution von Resistenzmutationen. Dieser Baum ist eine quantitative Variante des qualitativen Baum-

Wahrscheinlichkeit eines Mutationsmusters

Die Wahrscheinlichkeit (likelihood L) eines Mutationsmusters x unter der Bedingung eines mutagenetischen Baumes T ist gegeben als das in der Formel angegebene Produkt zweier Produkte:

$$L(x \mid T) = \prod_{e \in x} p(e) \prod_{e \notin x} (1-p(e))$$

Dabei läuft der Index e über alle Kanten des Mutationsbaums – jedes e entspricht einem Übergang zwischen zwei Mutanten. Das erste Produkt geht über alle Wahrscheinlichkeiten p(e) von Mutationen die zum Muster x führen, das zweite Produkt steuert einen Faktor von 1 – p(e) bei für jede Mutation, die nicht zu x führt.

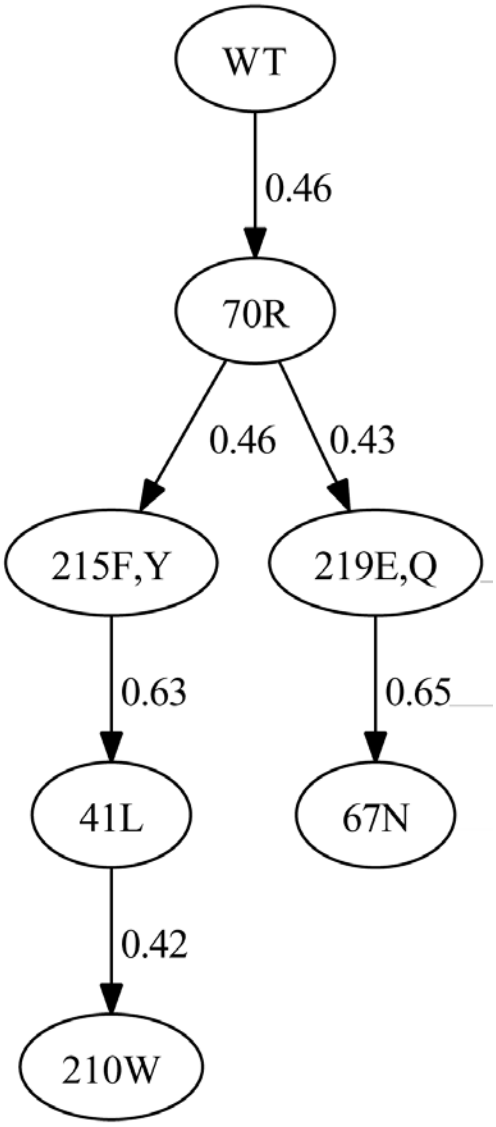



(3) Resistenzmutation in HIV-Protease. Oben: HIV-Protease (PR) mit dem Kanal, in dem das virale Polyprotein geschnitten wird – ein notwendiger Schritt der Virusreplikation. Jede der kleinen Kugeln stellt ein Atom dar. Die Lage der Atome wurde experimentell bestimmt[8]. Mitte: Der Kanal wird durch einen Hemmstoff blockiert (dunkelgrau). Unten: Eine Resistenzmutation (dunkelgrau) verformt den Kanal so, dass der Hemmstoff sich nicht mehr fest einlagern kann.

Modells, das Darwin im vorletzten Jahrhundert für die Entwicklung der biologischen Arten skizziert hatte (Abb. 1). Wie kann man dieses quantitative mathematische Modell anwenden, um Wahrscheinlichkeiten von Mutationspfaden vorherzusagen? Die Wahrscheinlichkeit $L$ (vom englischen Begriff *likelihood*) für ein Mutationsmuster $x$ ergibt sich für einen gegebenen Baum $T$ aus dem Produkt der Wahrscheinlichkeiten $p(e)$ der Kanten $e$, die auf diesem Pfad beschritten werden, sowie für jede nicht beschrittenene Kante einem Faktor von $1 - p(e)$ (siehe Kasten *Wahrscheinlichkeit eines Mutationsmusters*). Je größer also die Wahrscheinlichkeiten der Kanten sind, die für ein Mutationsmuster $x$ beschritten werden, und je kleiner die Wahrscheinlichkeiten der Kanten, die für $x$ *nicht* beschritten werden, umso wahrscheinlicher ist das Mutationsmuster $x$.

**Von Bäumen zu Wäldern**

Wie sich bei weiterer Analyse von Patientendaten herausstellte, sind mutagenetische Bäume noch nicht ausreichend um die komplexe Mutationsdynamik genau zu beschreiben. Beispielsweise erscheint manchmal die Mutation 215*F,Y*, ohne dass vorher die Mutation 70*R* aufgetreten ist, während der Baum in Abbildung (4) nur den Weg über 70*R* erlaubt. Eine natürliche Verallgemeinerung eines einzelnen Baumes ist eine Mischung mehrerer verschiedener Bäume, wie zum Beispiel in Abbildung (5) gezeigt. Mit einer solchen Mischung $M$ aus mehreren Bäumen $T_k$ ergibt sich die Wahrscheinlichkeit eines Mutationsmusters $x$ als Summe über die Wahrscheinlichkeit des Musters für jeden der Bäume $T_k$, gewichtet mit einem Faktor $\alpha_k$ (siehe Kasten *Wahrscheinlichkeit eines Mutationsmusters im Mischmodell*). Die Parameter $\alpha_k$ werden ebenso aus den in der Realität beobachteten Mutationsdaten geschätzt wie die Verknüpfungsmuster und die Kantenwahrschein-




(4) Mutagenetischer Baum. Dieser Baum zeigt wie sich aus dem viralen Wildtyp (WT) bei Gabe des Medikaments Zidovudine Resistenzmutationen (Knoten) im viralen Protein Reverse Transkriptase bilden. Die Zahlen an den Kanten sind bedingte Mutationswahrscheinlichkeiten, zum Beispiel entwickelt sich aus dem Wildtyp mit einerWahrscheinlichkeit von 0.46 die Mutation 70R, wobei „70R" bedeutet, dass der Aminosäurebaustein an der Stelle 70 der Proteinkette der Reversen Transkriptase ausgetauscht wird durch die Aminosäure Arginin (abgekürzt „R").

lichkeiten *p(e)* der einzelnen Bäume. Offen bleibt dann nur noch, wie viele Bäume in die Mischung aufzunehmen sind, also der Parameter *K*. Es erscheint sinnvoll, nicht mehr Bäume aufzunehmen, als für eine gute Vorhersage notwendig sind. Im Fall von Resistenzmutationen gegen das Medikament Zidovudine stellte sich heraus, dass drei Bäume dafür genügen. Bei anderen Medikamenten oder Kombinationen von Medikamenten können mehr Bäume notwendig sein. Wie wird ein Mischmodell angewendet? Das Modell erlaubt die Berechnung von Wahrscheinlichkeiten von Mutationsmustern oder die Erzeugung repräsentativer Ensembles von Mutationsmustern. Beispielsweise besagt das Modell in Abbildung (5), dass 19 Prozent der beobachteten Daten durch den Baum auf der linken Seite erklärt werden, der ein gewisses Mutations-Grundrauschen modelliert, in dem jede Mutation allein auftreten kann. 47 Prozent der Daten werden durch eine lineare Sequenz von Mutationen vom Wildtyp zur Mutation 41*L* modelliert. Die verbleibenden 34 Prozent erklärt das verzweigte Modell auf der rechten Seite. Nutzt man das Mischmodell zur Erzeugung eines repräsentativen Ensembles von Mutationsmustern, so produziert es Muster, die sehr gut übereinstimmen mit den Mutationsmustern in Patientendaten, auch von solchen Patientendaten, die nicht in die Parametrisierung des Modells eingeflossen sind.

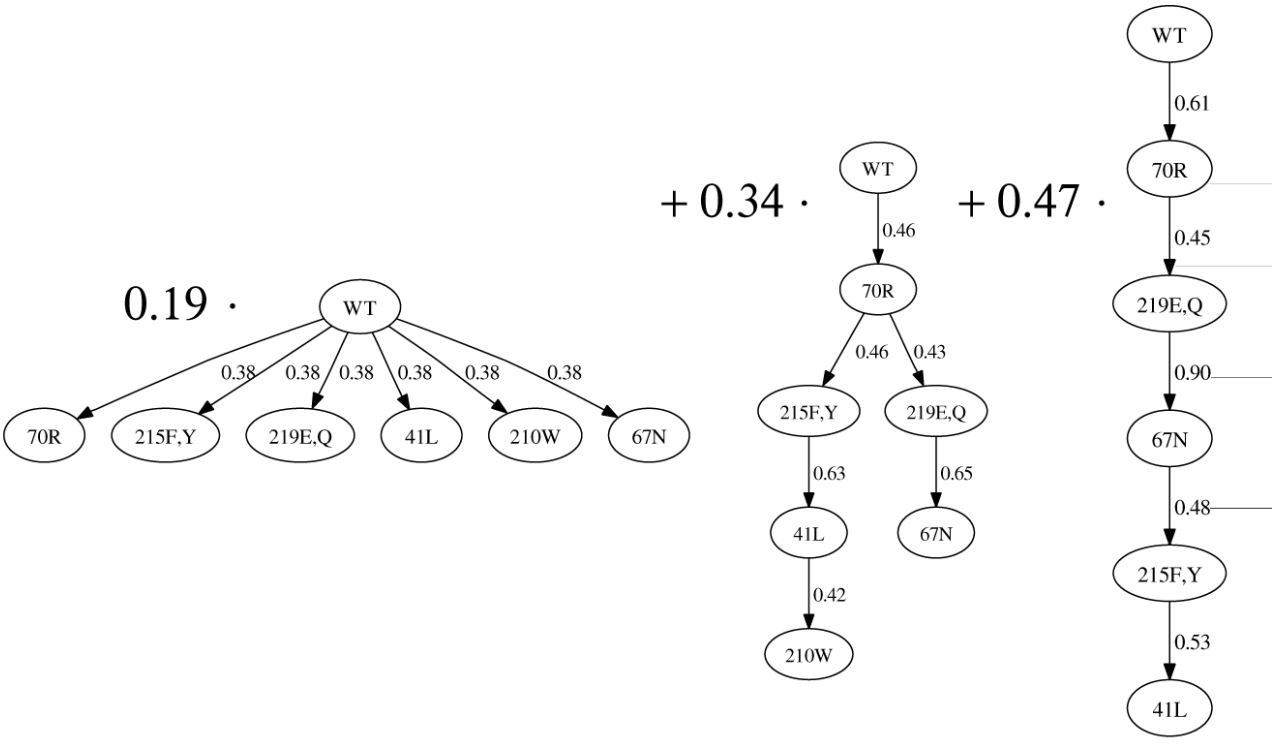
**Die genetische Barriere gegen virale Medikamentenresistenz**

Eine weitere Anwendung der Mischmodelle ist die Bewertung von Therapien und die Auswahl langfristig wirksamer Therapien. Ein geeignetes Maß für Wirksamkeit eines Medikaments gegen eine bestimmte Virusvariante ist der so genannte Resistenzfaktor, den man in Zellkulturexperimenten bestimmen kann. Wie der Name nahelegt, gilt: Je höher der Resistenzfaktor, umso höher die Resistenz der Virusvariante gegen das Medikament, im Verhältnis zur Resistenz des Wildtyps gegen das gleiche Medikament. Eine langfristig wirksame Therapie sollte also für eine lange Zeit die Resistenzfaktoren der evolvierten Mutanten unterhalb einer nach medizinischen Gesichtspunkten zu wählenden Schwelle halten. Man spricht in diesem Falle davon,

Wahrscheinlichkeit eines Mutationsmusters im Mischmodell

Das Mischmodell M ist eine lineare Überlagerung aus mehreren mutagenetischen Bäumen $T_k$, jeweils gewichtet mit Faktoren $a_k$, die aus klinischen Daten geschätzt werden. Damit ergibt sich die Wahrscheinlichkeit eines Mutationsmusters x zu: $L(x \mid M) = \sum_{k=1}^{K} a_k L(x \mid T_k)$

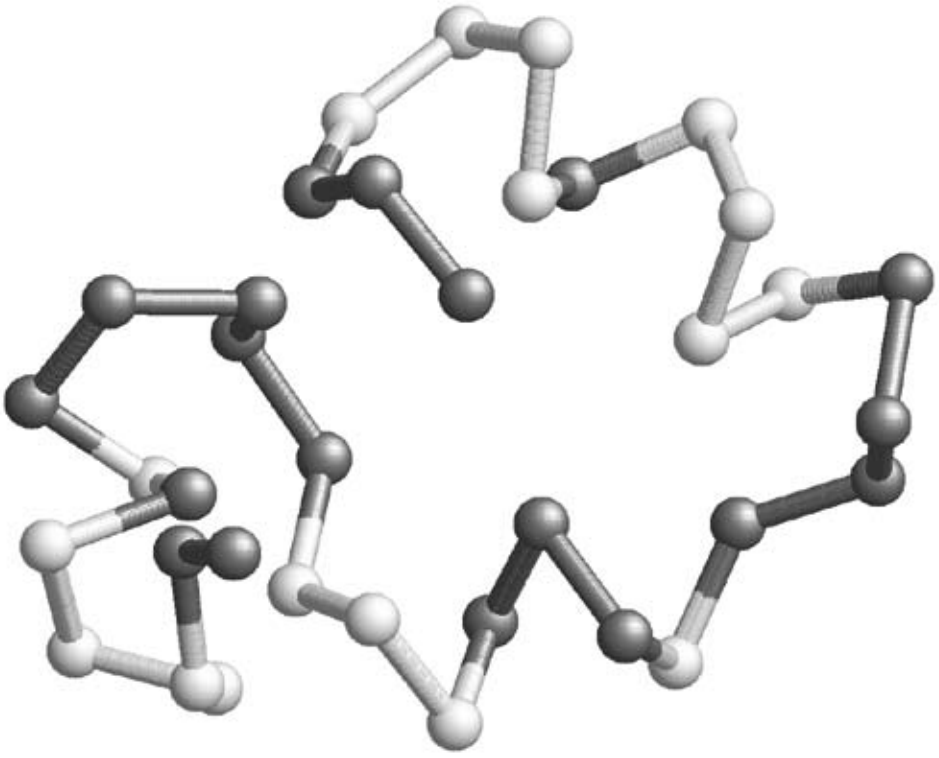
dass die *genetische Barriere* gegen Resistenz hoch ist. Die oben eingeführten mutagenetischen Bäume enthalten keinen direkten Bezug zur zeitlichen Entwicklung von Mutationsmustern. Um Aspekte wie die Langfristigkeit der Resistenzunterdrückung zu bewerten, muss dieser Bezug jedoch hergestellt werden. Eine Möglichkeit dazu ist folgende: Es ist plausibel anzunehmen, dass, je geringer die Wahrscheinlichkeit *p(e)* einer Kante in einem mutagenetischen Baum ist, umso länger die Zeit, die das Virus für die entsprechende Mutation benötigt. Unter Annahme eines Modells (zum Beispiel eines Poisson-Prozesses) für den zeitlichen Verlauf einer Mutation kann man also die *p(e)* umrechnen in mittlere Zeitdauern bis zum Auftreten einzelner Mutationen, und damit lassen sich auch Zeitdauern für beliebige Folgen von Mutationen simulieren. Nachdem der Bezug zwischen dem evolutionären Mischmodell, der Stärke der Resistenz und der Zeit hergestellt ist, kann man nun daran gehen, genetische Barrieren zu quantifizieren. Dazu wird zuerst der Begriff der genetischen Barriere für eine gegebene Virusvariante definiert als Summe der Wahrscheinlichkeiten von Mutationspfaden im Mischmodell, die für eine vorgegebene Therapiedauer zu Änderungen der Resistenzfaktoren führen, die unterhalb der willkürlich gewählten Schwelle bleiben (siehe auch den Kasten zum Thema „*Genetische Barriere*“). Dann werden anhand klinischer Beobachtungen Mischmodelle mutagenetischer Bäume aufgestellt und damit genetische Barrieren für verschiedene Behandlungen berechnet. So wurde beispielsweise gefunden,[10] dass sich die genetische Barriere gegen Zidovudine-Resistenz nach 96 Wochen Behandlung erhöht, wenn gleichzeitig das Medikament Lamivudine gegeben wurde. Andererseits halbiert sich die genetische Barriere bei einer Kombination von Zidovudine und Didanosine, das heißt von einer solchen Therapie wäre abzuraten unter dem Gesichtspunkt der langfristigen Vermeidung von Zidovudine-Resistenz. Dieses Beispiel demonstriert, wie ein mathematisches Modell der viralen Evolution helfen kann, bessere Therapien zu entwickeln.




(5) Ein Mischmodell aus $K = 3$ mutagenetischen Bäumen. Die Faktoren 0.19, 0.34 und 0.47 entsprechen den $\alpha_k$ in der Gleichung in der Box *Wahrscheinlichkeit eines Mutationsmusters im Mischmodell*. Die Mischung sagt die Häufigkeiten von Mutationsmustern von HIV sehr gut vorher, die bei Behandlung mit dem Medikament Zidovudine auftreten.

## Virtuelle Evolution zur Optimierung von Molekülen

Während im obigen Beispiel ein realer Evolutionsvorgang mathematisch modelliert wurde, soll nun gezeigt werden, wie der Optimierungs-Charakter einer *virtuellen* Evolution genutzt werden kann, um



(6) Peptidkette und Faltung. Gezeigt ist die Faltung der Kette einer Mutante des Peptids Villin Headpiece ($VH^{11}$) aus 36 Aminosäuren. Von jeder Aminosäure ist nur das zentrale Atom als kleine Kugel dargestellt. Tatsächlich ist das Peptid ein kompaktes, aus vielen Atomen bestehendes Gebilde. Die meisten Atome sind jedoch weggelassen um den Kettencharakter des Peptids zu unterstreichen. Die hellen Kugeln gehören zu hydrophilen (gut wasserlöslichen), die dunklen zu hydrophoben (schwer wasserlöslichen) Aminosäuren. Die Abfolge der Aminosäure bestimmt die Faltung.

Genetische Barriere

Die genetische Barriere $B^{d,c}_{R,t}(x)$ für ein Mutationsmuster x wird definiert als

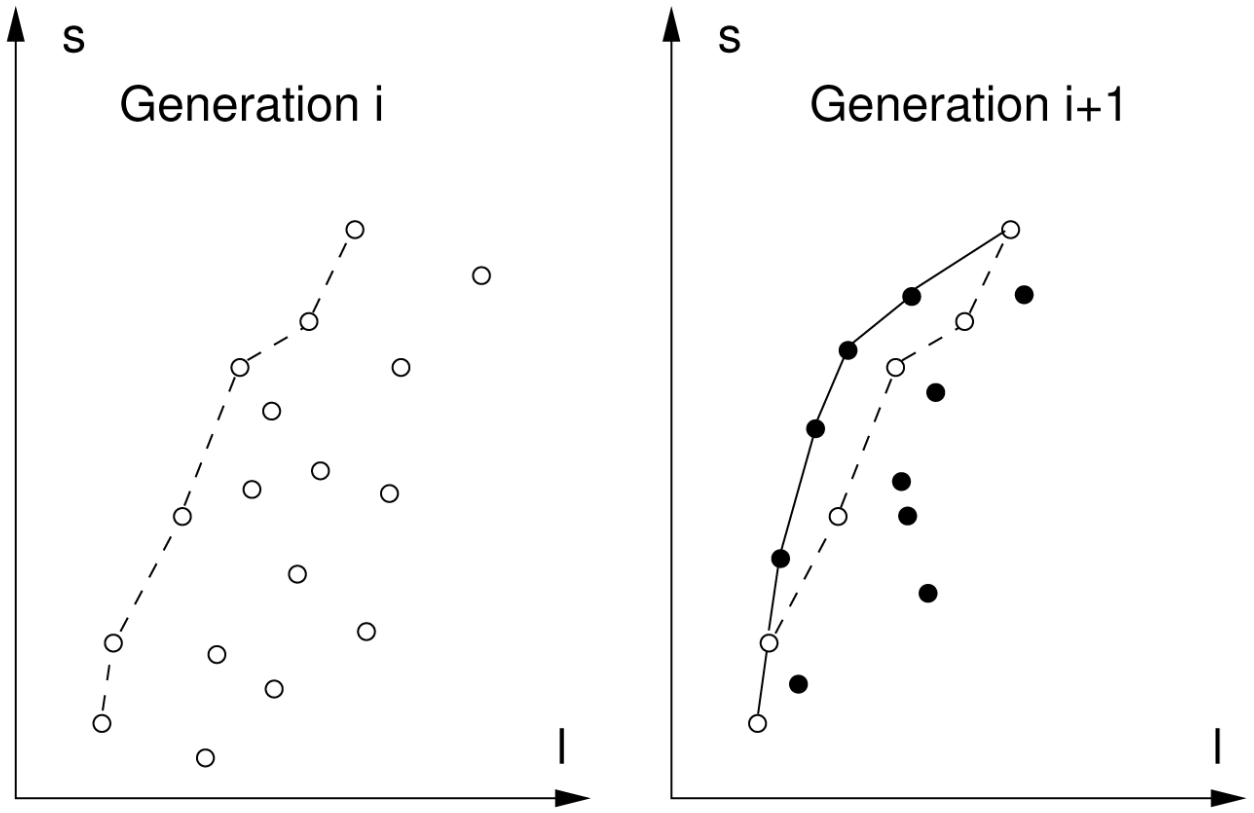
$$B^{d,c}_{R,t}(x) = Pr(FC(Y) < c | X = x)$$

mit Wirkstoff d, Resistenzschwelle c, Therapie R (besteht im Allgemeinen aus mehreren Medikamenten), Behandlungsdauer t. Y ist eine Mutante von x mit einem $\log_{10}$-"fold change" der Resistenz FC(Y). Die genetische Barriere ist die Summe aller Wahrscheinlichkeiten (Pr) für Mutanten, die aus x entstehen und deren Resistenz bei der gegebenen Behandlung nach der Zeit t unterhalb der Schwelle c bleibt.

Eigenschaften von Peptid-Molekülen im Computer zu optimieren. Von ihrer chemischen Natur her sind Peptide Proteinmoleküle (wie die oben beschriebene HIV-Protease), also Kettenmoleküle aus zwanzig verschiedenen Aminosäure-Bausteinen. Der einzige Unterschied zu Proteinen besteht in der Länge der Ketten: Peptide sind kurze Ketten, von bis zu fünfzig Aminosäuren, während die Ketten von Proteinen länger sind. Die Reihenfolge der verschiedenen Aminosäuren in der Kette legt die Eigenschaften des Moleküls fest, vergleichbar dem Genom-Text, der Eigenschaften des Organismus bestimmt. Und genauso wie das Genom im Laufe der Evolution über viele Zyklen mutiert, repliziert und anhand der Fitness des Phänotyps selektiert wird, so kann man auch die Sequenz der Aminosäuren im Computer über viele Zyklen mutieren, replizieren und anhand einer errechneten Fitness des Peptids selektieren. Wie bei der natürlichen Evolution, sollte auch hier die Fitness der Peptide über die Zeit wachsen – wir haben ein so genanntes Evolutionäres Optimierungsverfahren. Eine herausragende Eigenschaft der Peptidkette ist ihre dreidimensionale Struktur, die so genannte Faltung, da sie bestimmt, wie das Peptid im biologischen Kontext wirkt. Auch die *Faltung* wird determiniert durch die Abfolge der Aminosäuren in der Kette, wobei sehr kurze Peptidketten keine Faltung ausbilden sondern flexibel sind. Viele etwas längere Ketten von typischerweise zwanzig bis fünfzig Aminosäuren falten sich dagegen auf wohldefinierte Weise (Abb. 6). Diese faltenden Peptide sollen uns im Folgenden beschäftigen.

**Was bedeutet *Fitness* bei Peptiden?**

Während in der natürlichen Evolution zumindest intuitiv klar ist, was Fitness bedeutet – die Fähigkeit




(7) Mehrkomponentige Fitness, Dominanz und Pareto-Front in der Evolutionären Optimierung. Links: Peptide (Kreise) im Koordinatensystem der Eigenschaften Stabilität s und Länge *l* in der Generation *i* einer Evolutionären Optimierung. Die Pareto-Front, gebildet von den nicht dominierten Peptiden ist durch eine gestrichelte Linie hervorgehoben. Rechts: Eine Generation später in der Evolution. Die besten Peptide aus der vorherigen Generation wurden übernommen und konnten mutierte Nachkommen erzeugen (gefüllte Kreise). Die Pareto-Front hat sich etwas zu höheren *s* und kleineren *l* verschoben (durchgezogene Linie).

unter den gegebenen Umweltbedingungen erfolgreich zu replizieren – ist die Fitness in der virtuellen Evolution von Peptiden im Computer willkürlich gewählt. Sie orientiert sich an den Eigenschaften, die der Anwender der Evolutionären Optimierung optimieren will, und die berechenbar sind. Beispielsweise könnte der Anwender daran interessiert sein Peptide zu erhalten, die eine möglichst stabile Faltung haben. In diesem Falle müsste er ein Peptidmodell verwenden, das die Berechnung der Stabilität erlaubt und dann im Rahmen der Selektion die jeweils stabilsten Peptide auswählen, replizieren lassen, mutieren, etc. Was aber tun, wenn der Anwender zwei oder mehr Eigenschaften der Peptide gleichzeitig optimieren will? Er könnte zum Beispiel an möglichst stabilen Peptiden mit möglichst kurzen Ketten interessiert sein. Eine häufig verwendete Methode zur Formulierung der Fitness ist in solchen Fällen die *ad hoc* Kombination der verschiedenen Optimierungskriterien in einer einzigen Fitnessfunktion (s. Kasten zum Thema „*Ein-Ziel-Fitness*“). Leider ist dieser Ansatz unbefriedigend, da man so Eigenschaften verrechnet, die nicht direkt etwas miteinander zu tun haben, ja die nicht einmal vergleichbar sind, wie im Falle der Stabilität und Kettenlänge von Peptiden.

## Mehrziel-Optimierung und Pareto-Dominanz

Eine Alternative zur *ad hoc* Kombination ist die so genannte Mehrziel-Optimierung, bei der alle Komponenten der Fitness auf gleiche Weise berücksichtigt werden, ohne dabei ihren eigenständigen

### Ein-Ziel-Fitness

In vielen Anwendungen der Evolutionären Optimierung sollen mehrere Ziele gleichzeitig angestrebt werden, zum Beispiel die Vergrößerung der Stabilität s einer Peptidfaltung und die Verkürzung der Peptidlänge l. Diese Zielfunktionen werden häufig zu einer einzigen Fitnessfunktion F kombiniert, zum Beispiel linear in der Form

$$F(s, l) = ß_s \cdot s - ß_l \cdot l,$$

wobei $ß_s$, $ß_l$ hier zwei positive reelle Konstanten sind. Gesamtziel der Optimierung wäre hier die Maximierung von F(s, l). Nachteil dieser Kombinationen ist die völlige Willkürlichkeit, mit der sie mehrere, nicht miteinander vergleichbare Eigenschaften zur einer Funktion verbinden. Ein Zeichen dafür ist, dass es im Allgemeinen keine „richtige" Möglichkeit gibt, die Konstanten $ß_s$, $ß_l$ zu bestimmen.

Charakter zu verlieren. Man stelle sich vor, Stabilität s und Länge *l* der Peptide werden auf zwei Achsen eines Koordinatensystems aufgetragen, wie in Abbildung (7) gezeigt. Das Ziel der Optimierung seien Peptide, die möglichst stabil und kurz sind. Bessere Peptide liegen weiter links und weiter oben als sub-optimale Peptide. Hat man in einem Schritt der Evolutionären Optimierung eine Population von Peptiden mit bestimmten Werten von *s* und *l* (kleine Kreise in Abb. 7), so wird man für die Replikation und Mutation solche Peptide selektieren, die möglichst weit links und möglichst weit oben liegen. Die in diesem Sinne besten Peptide zeichnen sich dadurch aus, dass sie *nicht dominiert sind.* Zu jedem *dominierten* Peptid gibt es mindestens ein anderes Peptid, das in beiden Disziplinen *s* und *l* besser ist, und folglich gibt es zu jedem *nicht dominierten* Peptid kein anderes Peptid, das in beiden Disziplinen besser ist. Die nicht dominierten Peptide bilden also eine Front, die die Menge der bisher generierten Peptide zu hohen Stabilitäten und kurzen Längen hin begrenzt (Linien in Abb. 7). Im Rahmen der Mehrziel-Optimierung versucht man diese Front – sie trägt den Namen *Pareto-Front*[12] nach dem Volkswirt Vilfredo Pareto – immer weiter nach links und oben zu verschieben. Es ist klar, dass damit *s* und *l* nicht auf schwer begründbare Weise verquickt werden wie es bei der *ad hoc* Kombination der Fall war. Andererseits hat diese Mehrziel-Optimierung einen Nachteil: sie liefert nicht ein einziges Optimum wie die Ein-Ziel-Optimierung, sondern alle Peptide auf der Pareto-Front bilden das Optimum. Es bleibt dann die Aufgabe des Anwenders aus diesen Pareto-optimalen Peptiden eine Auswahl zu treffen, zum Beispiel anhand weiterer Kriterien, die nicht Gegenstand der Optimierung waren, wie etwa der leichten Herstellbarkeit. Das Buch von Deb[7] gibt einen Überblick über Evolutionäre Mehrziel-Optimierung.

**Ein Computer-Experiment**

Wie gut funktioniert Evolutionäre Mehrziel-Optimierung in der Praxis? Die Methode findet gerade erst ihren Weg in die Optimierung von Biomolekülen, sodass diese Frage noch nicht abschließend beantwortet werden kann. Wir haben diese Methode erstmals auf das Design von Peptiden angewendet. Zum Testen des Verfahrens haben wir folgendes Computerexperiment durchgeführt. Als Vorbereitung wurde das stabile Peptid Villin Headpiece (VH) so mutiert, dass es instabil wurde. Dann wurde das so geänderte Peptid einem Evolutionären Optimierungsverfahren unterworfen, wobei auf zwei Ziele hin optimiert werden sollte: Erstens sollte die Stabilität des Peptids wieder hergestellt oder womöglich noch weiter verbessert werden, zum zweiten sollte ein bestimmter Teil des Peptids eine Form annehmen wie im ursprünglichen VH. Die Erwartung, die wir mit der Optimierung verbanden, war, dass unter den Peptiden, die das Verfahren generieren würde, auch das ursprüngliche VH sein würde, von dem wir ja wussten, dass es stabil ist und die gewünschte Form hat. Es wurden zwei unabhängige Optimierungsläufe unternommen, jeder bestehend aus 15 Generationen zu je acht Peptiden, so dass insgesamt 240 Peptide im Rechner erzeugt wurden. Das Verfahren verbraucht übrigens sehr viel Rechenzeit, da das Verhalten jedes der Peptide in wässriger Umgebung über eine gewisse Zeit simuliert werden musste, um die Stabilität beurteilen zu können. Insgesamt belief sich die reine Rechenzeit auf ein Jahr. Da sich Evolutionäre Algorithmen gut parallelisieren lassen, konnte dieses Jahr auf mehrere Rechner verteilt werden, sodass die Simulationen ungefähr einen Monat dauerten.[11] Unter den 240 Peptiden, die das Evolutionäre Optimierungsverfahren erzeugte, war erstaunlicherweise nicht das ursprüngliche VH, sondern neun Peptide, die – im Rahmen des berechneten Modells – stabiler waren als dieses Peptid. Eines der vorhergesagten Peptide wurde im Labor synthetisiert und seine Stabilität und Struktur gemessen. Es stellte sich heraus, dass das Peptid tatsächlich stabiler war als das VH-Peptid, während seine Faltung fast identisch mit der des VH-Peptids war. Weitere vorhergesagte Peptide werden derzeit untersucht.

**Summary**

Evolution is both a concept in biology and an abstract optimization principle. In biology, evolution has for the first time rigorously explained how new species emerge from predecessor species. The process is iterative and involves the steps of mutation, selection, and replication of individuals. In the course of the process, species change towards an optimal adaptation to their environment. In this text, two current examples are given of application of evolutionary concepts. First, it is demonstrated how the adaptation of human immunodeficiency virus (HIV) to its host and to antiviral therapies can be modelled with evolutionary concepts such as the mutagenetic tree, a kind of phylogenetic tree of the viral genome focused on mutations relevant to resistance to antiviral therapies. It is shown how mutagenetic trees and mixtures thereof are used to quantify the longterm effectiveness of therapies in view of the possibility of the virus to develop resistance mutations. The second topic of the text is how virtual evolution in the computer can be used to optimize biomolecules. The concept of multi-objective optimization based on Pareto-dominance as fitness criterion is introduced as it promises to provide solutions to

real-world optimization problems with multiple independent fitness requirements.

## *Literatur*

1) C. Darwin. The Autobiography of Charles Darwin and Selected Letters. Dover, New York, 1958.
2) C. Darwin. On the Origin of Species. J. Murray, London, 1859.
3) G. H. Hardy. Mendelian Proportions in a Mixed Population. Science, 28:49–50, 1908.
4) W. Weinberg. Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg, 64:368–382, 1908.
5) J. D.Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids. Nature, 737–8, 1953.
6) M. A. Nowak. Evolutionary Dynamics. Harvard University Press, 2006.
7) K. Deb. Multi-Objective Optimization using Evolutionary Algorithms. Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Ltd., Baffins Lane, Chichester, West Sussex, England, 2001.
8) V. Stoll, W. Qin, K. D. Stewart, C. Jakob, C. Park, K. Walter, R. L. Simmer, R. Helfrich, D. Bussiere, J. Kao, D. Kempf, H. L. Sham, and D. W. Norbeck. X-ray crystallographic structure of ABT-378 (lopinavir) bound to HIV-1 protease. Bioorg. Med. Chem., 10:2803–6, 2002.
9) N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer. Learning multiple evolutionary pathways from cross-sectional data. J. Comput. Biol. RECOMB 2004, 12:584–598, 2005.
10) N. Beerenwinkel, M. Däumer, T. Sing, J. Rahnenführer, T. Lengauer, J. Selbig, D. Hoffmann, and R. Kaiser. Estimating hiv evolutionary pathways and the genetic barrier to drug resistance. J. Infect. Dis., 191:1953–1960, 2005.
11) W. Gronwald, T. Hohm, and D. Hoffmann. Evolutionary Paretooptimization of stably folding peptides. BMC Bioinformatics, 9, 2008. doi:10.1186/1471-2105-9-109.
12) V. Pareto. Cours d'économie politique, volume 2. Droz, Genève, 1897.

## *Der Autor*

Daniel Hoffmann hat an der Universität Heidelberg Physik studiert und das Studium abgeschlossen mit einer Diplomarbeit aus der theoretischen Molekülphysik am Max-Plank-Institut für Medizinische Forschung unter Anleitung von Ulrich Haeberlen. Er ging dann als wissenschaftlicher Mitarbeiter in die Arbeitsgruppe Macromolecular Modelling (Ernst-Walter Knapp) an das Institut für Kristallographie am Fachbereich Chemie der Freien Universität Berlin und promovierte dort 1996 über Methoden zur Simulation der Langzeitdynamik von Proteinen. Es schloss sich ein Postdoc-Aufenthalt in der Bioinformatik-Abteilung von Thomas Lengauer am GMD-Forschungszentrum Informationstechnik GmbH an. Im Jahr 2000 ging Daniel Hoffmann an das Centre of Advanced European Studies and Research in Bonn, um dort eine Gruppe aufzubauen, die Methoden entwickelte für die Strukturaufklärung von Proteinen und für das Design funktionaler Peptide. In 2005 war er dort Research Fellow und gleichzeitig Professor für Bioinformatik an der Fachhochschule Bingen. Seit 2006 ist Daniel Hoffmann Professor für Bioinformatik an der Universität Duisburg-Essen.

**DOI:** 10.17185/duepublico/73805
**URN:** urn:nbn:de:hbz:464-20210204-154849-2