# Evolution

## Genetik für Bioformatiker # 1

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

14. Oktober 2015

# Evolution

Nothing in Biology Makes Sense Except in the Light of Evolution
-Theodosius Dobzhansky, 1973

Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts some of them interesting or curious but making no meaningful picture as a whole.

# Genetik & Bioinformatik

Nothing in Bioinformatics Makes Sense Except in the Light of ~~Biology~~ Genetics
-A scholar, 2014

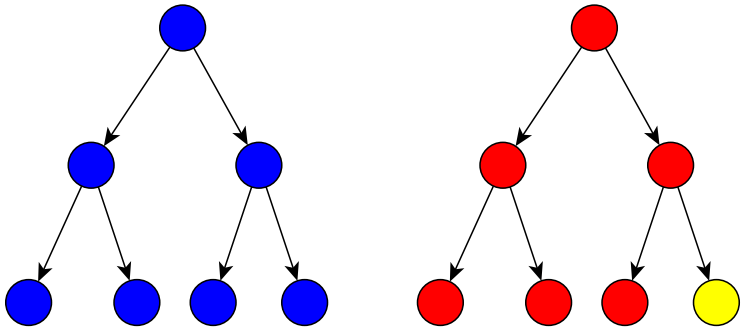

Bild: Zhu et al. (2013) *Scientific Reports* **3**:3084

In diesem Kurs wollen wir die Genetik nicht als auswendig zu lernendes biologisches Allerlei betrachten, sondern aufzeigen wie die Bioinformatik genetische Daten und Algorithmik verbindet um Biologie zu verstehen

# Dieser Kurs

- Kurs Homepage: http://compbio.charite.de/contao/index.php/teaching.html
- Vorlesungen
- Übungen: Aktive Teilnahme
- Klausur
- Sprechstunde: Nach Vereinbarung (peter DOT robinson AT charite DOT de)

# Evolution: Grundlagen



- Organismen vermehren sich
- Heredität: “Gleiches zeugt Gleiches”
- Variation: Ab und zu kommt es zur Ausnahme von der Hereditätsregel
- Selektion: Zumindest ein Teil der Variation beeinflusst das Überleben und Reproduktion der Organismen

Evolution: Multiplikation, Heredität, Variation, Selektion

# Evolution: Grundlagen

Die Theorie der natürlichen Selektion postuliert nicht nur den evolutionären Wandel, sondern sagt auch etwas darüber aus, wie dieser Wandel passiert

- Populationen werden Charakteristika entwickeln, womit sie in ihrer Umgebung besser überleben und reproduzieren können
- Der Wandel erfolgt schrittweise

# Genotyp & Phänotyp

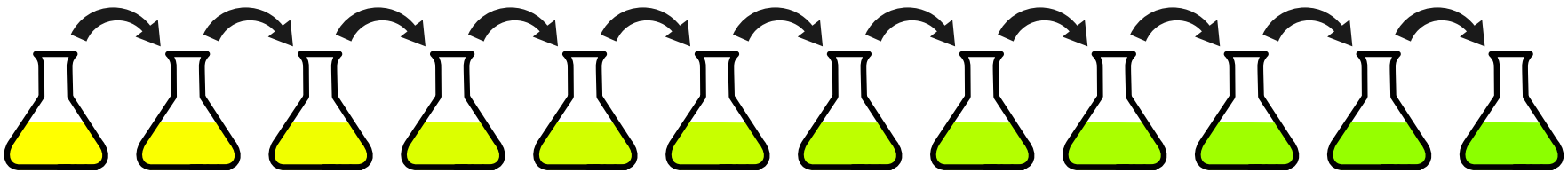
- Phänotyp: Erscheinungsbild, die Menge der beobachtbaren Merkmale eines Organismus
- Genotyp: “Erbbild”, die Menge der Gene und Varianten eines Organismus
- Umweltfaktoren und Genotyp bestimmen Phänotyp

In diesem Kurs wollen wir statistische Modelle und bioinformatische Algorithmen aufzeigen, womit wir das Zwischenspiel von Genotyp, Umwelt und Phänotyp besser verstehen können

# Evolution in der Retorte

Das RNA Virus Q$\beta$ infiziert *Escherichia coli* Bakterien

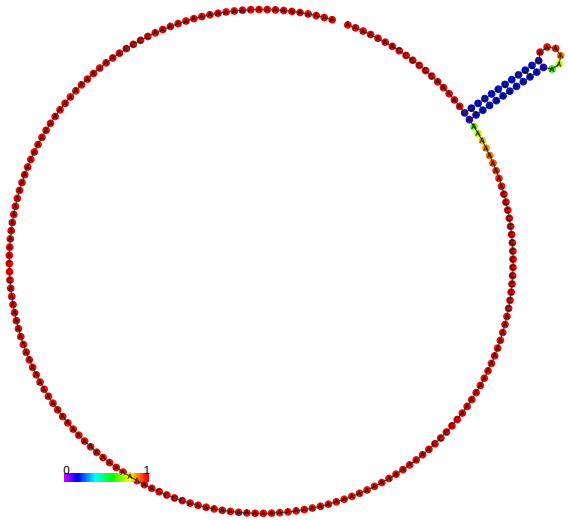
- Das Q$\beta$-Genom kodiert für ein Enzym, das RNA repliziert: Q$\beta$-Replikase
- Q$\beta$-Replikase repliziert fast jedes RNA-Molekül in vitro, solange die vier Nukleotide ATP, CTP, GTP und UTP vorhanden sind

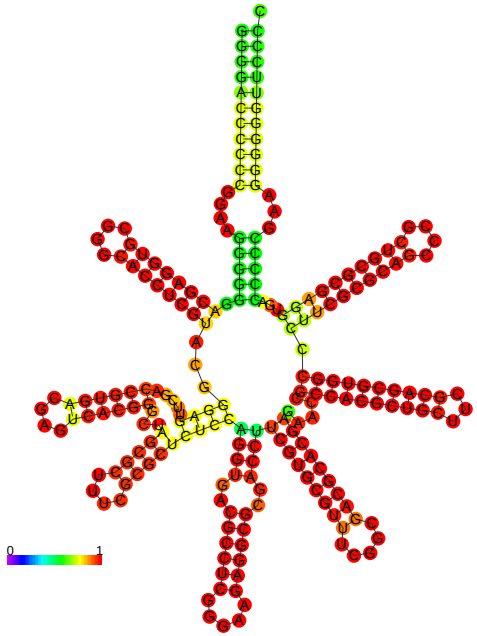


- Unser Versuch: Ein RNA Template wird einer Q$\beta$-Replikase und Nukleotide enthaltenden Retorte hinzugefügt
- Nach 30 Minuten wird ein Aliquot entnommen und der nächsten Retorte hinzugefügt
- Dieses Verfahren wird n-mal wiederholt

# Q$\beta$-Replikase

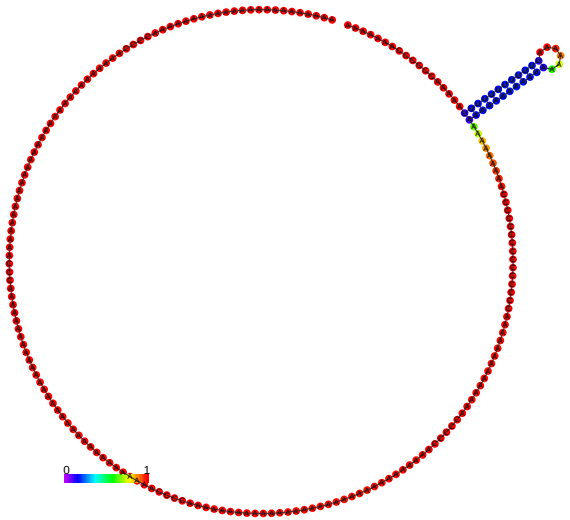
- Falls die Replikation exakt ist, würden wir genau dieselbe RNA-Sequenz in der letzten wie in der ersten Retorte vorfinden
- Aber: die “Fehlerquote” der Q$\beta$-Replikase beträgt etwa 1:10.000 Nukleotide
- Wir haben also Multiplikation und Variation, aber woher kommt die Selektion?

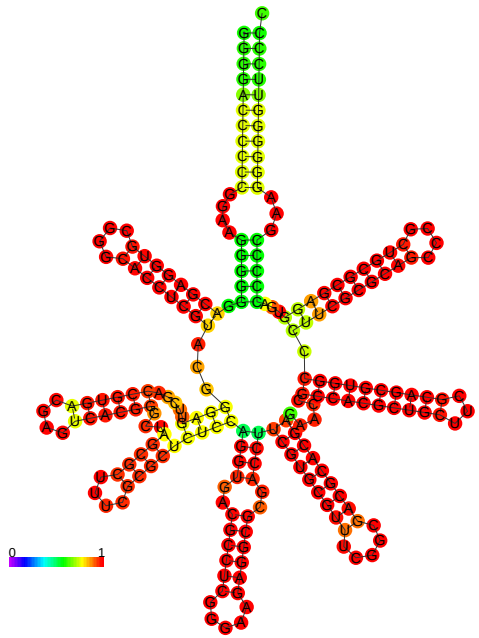




# Q$\beta$-Replikase

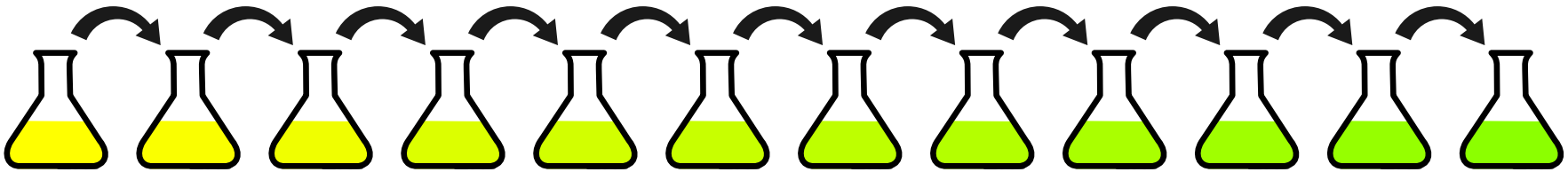
- RNA-Moleküle haben dreidimensionale Strukturen
- Komplementäre Basen bilden Haarnadel-Strukturen
- Die Q$\beta$-Replikase repliziert die rechts unten stehende RNA besonders rapide





# Q$\beta$-Replikase

Unser Versuch kann daher zur Darwinistischen Evolution führen. Mit jedem Transfer ändert sich die Zusammensetzung der RNA-Moleküle, so dass die anfänglichen RNA-Moleküle nach und nachdurch rapide replizierenden RNA-Moleküle ersetzt werden.



- Das in der letzten Folien gezeigte 218 nt lange Molekül ist das Ergebnis eines solchen Versuches
- Es ähnelt einem in der Natur vorkommenden, “Minivariant” genannten Molekül, das in durch das Q$\beta$ Virus infizierten *E coli* Bakterien vorkommt

# Ein Riesenzufall?

Unser in der Retorte gezüchtetes RNA 218 nt langes Molekül gleicht also einem in der Natur vorkommenden RNA-Molekül. Wie genau kommt es dazu?

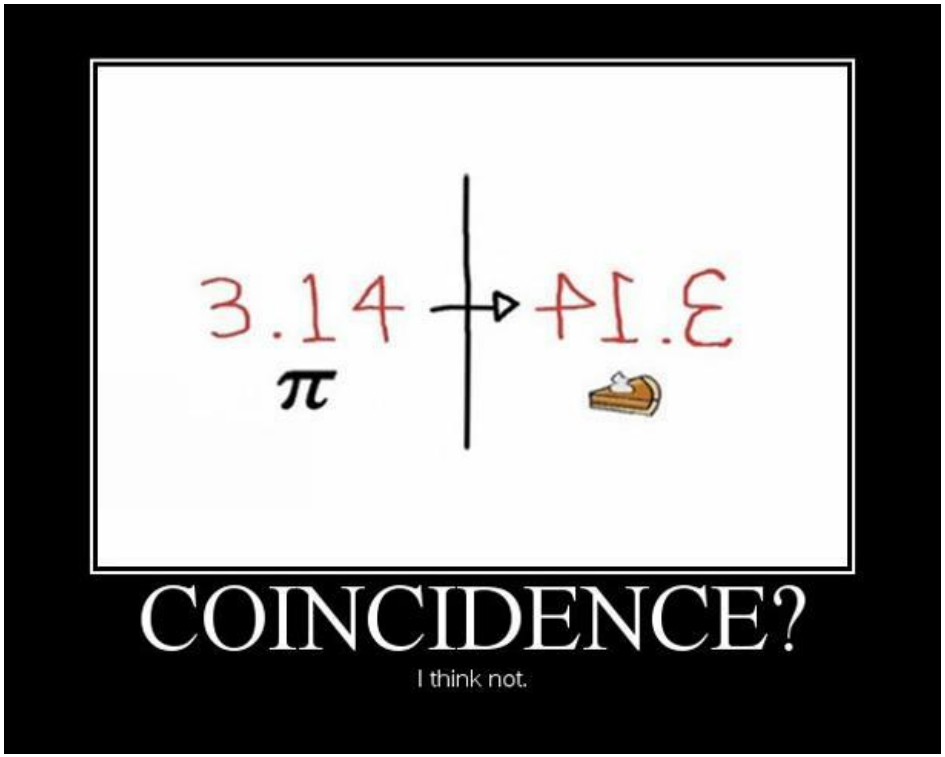



Bild: http://thepoisonedpencil.com

# Ein Riesenzufall?

Wie viele mögliche RNA-Moleküle der Länge 218 Nukleotide gibt es?

# Ein Riesenzufall?

$$4^{218} \approx 10^{132} \quad \text{unterschiedliche Moleküle}$$

- Dies wirft ein Rätsel auf
- In der ersten Retorte befanden sich ca. $10^{16}$ RNA-Moleküle
- In unserem Versuch haben wir 100mal ein Aliquot von einer Retorte in die nächste übertragen
- Das heißt, wir haben allermaximal $10^{18}$ RNA-Moleküle “erprobt”
- Wie haben wir denn ein bestimmtes Molekül, d.h., eins aus $10^{132}$, “gezüchtet”?

# Ein Riesenzufall?

Es ist falsch zu glauben, dass die Evolution alle möglichen Phänotypen “ausprobiert” bis sie durch Zufall den “besten” identifiziert

- Die Evolution kann stattdessen mit einem Prozess des Treppensteigens verglichen werden, wobei der Phänotyp nach jedem Schritt besser ist als der vorherige

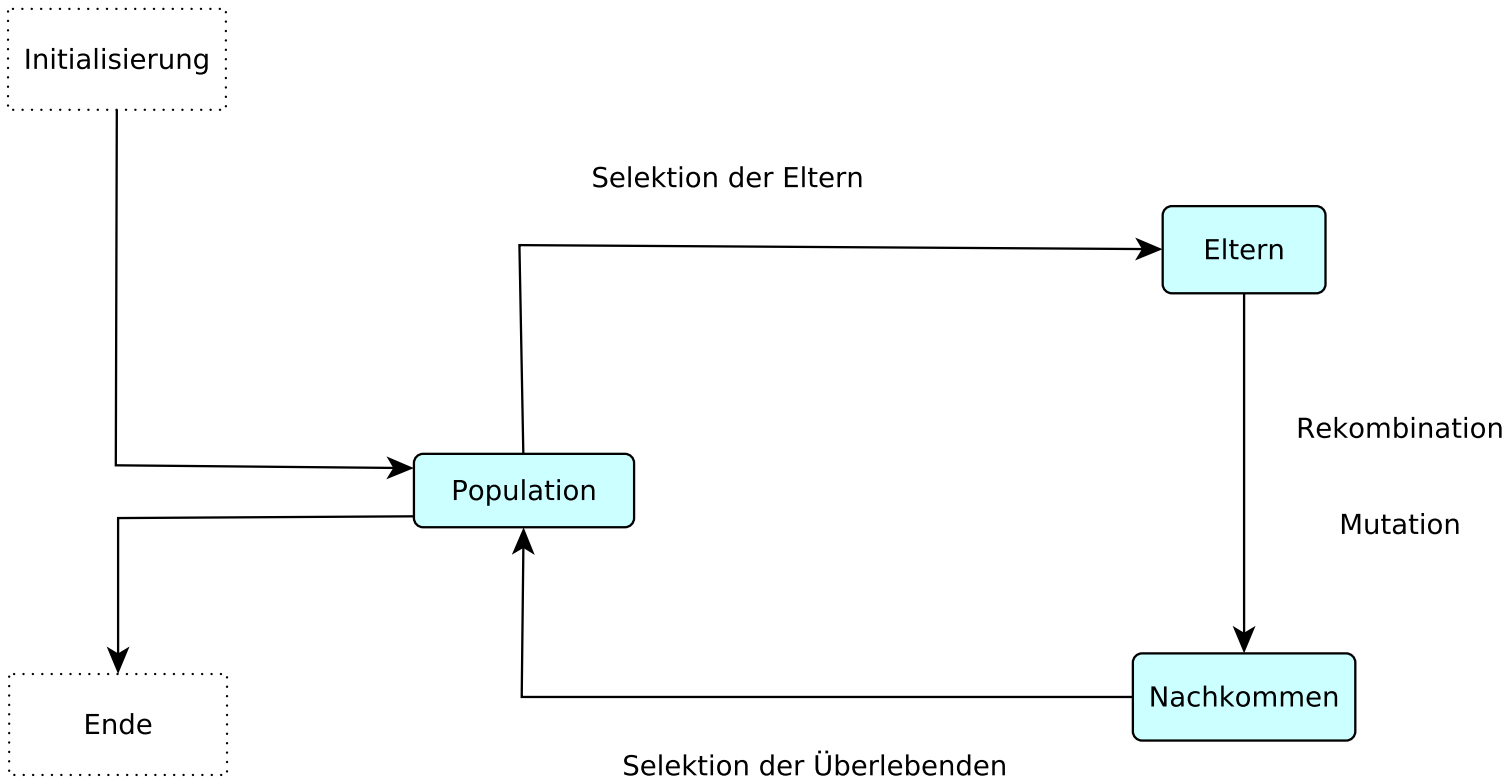
# Optimisierung

**Evolution als “Treppensteigen”**

- Jeder Schritt erfolgt durch eine einzelne Mutation
- Jeder Schritt erhöht die Replikationsgeschwindigkeit
- Die Gesamtzahl der Schritte ist verhältnismäßig klein, auf jeden Fall $\ll 10^{132}$.
- Die Wahrscheinlichkeit, dass das Endergebnis nach einem einzelnen Schritt auftritt ist verschwindend gering

# Evolutionäre Algorithmen

Die natürliche Selektion stand für eine Klasse von Algorithmen Pate: “Genetische Algorithmen” bzw. “Evolutionäre Algorithmen” gennant.



# Genetische Algorithmen

> Genetic Algorithms are good at taking large, potentially huge search spaces and navigating them looking for optimal combinations of things, solutions you might not otherwise find in a lifetime.
>
> -Salvatore Mangano, Computer Design, 1995

- Evolutionäre Algorithmen basieren auf Algorithmen, die eine gewisse Ähnlichkeit zur Evolution, Genetik und der natürlichen Selektion aufweisen
- Zahlreiche Variationen der Grundidee, die wir hier nicht vorstellen wollen
- Wir werden hier eine stark vereinfachte Version eines EA vorstellen um das Problem des reisenden Kaufmanns (TSP) zu lösen

# Genetische Algorithmen

Die “Zutaten” eines genetischen Algorithmus umfassen ein Problem und . . .

| “Zutat” | “Biologie” |
|---|---|
| Kodierung | Gen, Chromosome |
| Initialisierung | Erschaffung |
| Evaluierung | Selektion |
| Selektion der Eltern | Reproduktion |
| Genetische Operatoren | Mutation, Rekombination |

# Genetische Algorithmen

**Algorithm 1** Simple genetic algorithm

```
1: initialize population
2: evaluate population
3: while Termination Criteria Not Satisfied do
4:   select parents for reproduction
5:   perform recombination
6:   perform mutation
7:   evaluate population
8: end while
```

- Grundideen eines genetischen Algorithmus

# Genetische Algorithmen: Reproduktion

Genetische Algorithmen bedienen sich einer Population von Individuen um in jeder "Generation" (d.h. Iteration) "Nachkommen" zu erzeugen

- Rekombination: Die genetischen Information ("Chromosome") zweier Eltern werden miteinander kombiniert
- Die "Chromosome" werden häufig als Bitstrings repräsentiert (00101001110111110101010...)
- Nach der Rekombination werden die Nachkommen einer Mutation unterzogen.
- Zum Beispiel könnte eine Mutation den Bitwert an einer bestimmten Position des Bitstrings verändern ($0 \rightarrow 1$ oder $1 \rightarrow 0$)
- Die Eltern werden jenach Fitnesswert ausgesucht

# Genetische Algorithmen: Mutation

| | |
|---|---|
| Vorher | (1 0 1 1 0 1 1 0) |
| Nachher | (1 0 1 0 0 1 1 0) |

- Die Variablen der Nachkommen werden durch kleine Störungen verändert
- “Schrittweise” Evolution
- Hier wird ein Bitstring (Chromosom) an der vierten Position von 1 nach 0 verändert

Auch andere Datentypen können verwendet werden...

| | |
|---|---|
| Vorher | (1.38 -69.4 326.44 0.1) |
| Nachher | (1.38 -67.5 326.44 0.1) |

- Die Folge ist eine Bewegung innerhalb des Suchraums

# Genetische Algorithmen: Rekombination

| Schritt | Chromosom 1 | Chromosome 2 |
|---|---|---|
| Vorher | (0 1 1 0 0 1 0 0) | (1 1 0 1 1 0 1 0) |
| Nachher | (0 1 0 1 1 0 1 0) | (1 1 1 0 0 1 0 0) |

- Die Rekombination ("Crossover") ist ein kritischer Bestandteil von genetischen Algorithmen
- Sie beschleunigt die Suche insbesondere in der frühen "Evolution" einer Population
- Sie erlaubt die wirksame Kombination von "Teillösungen" in unterschiedlichen Bereichen einzelner Chromosomen

# Genetische Algorithmen: Evaluation

- Nach erfolgter Rekombination und Mutation erfolgt eine Evaluation der einzelnen Nachkommen
- Die Evaluationsfunktion kodiert sozusagen die Logik des spezifischen Problems; die “Maschinerie” des genetischen Algorithmus ist von den Spezifika des Problems entkoppelt
- Verschiedene Varianten des GA
  - die ganze Population wird nach jeder Generation ersetzt
  - Die Population wird nur teilweise ersetzt, und zwar durch Nachkommen mit besseren Fitnesswerten

# Problem des Handlungsreisenden

Travelling salesman problem (TSP), ein NP-schweres Problem

- Finde eine Reihenfolge für den Besuch mehrerer Orte so dass...
  - jeder Ort nur einmal besucht wird
  - die gesamte Reisestrecke möglichst kurz ist

# Problem des Handlungsreisenden: GA

Die Repräsentation der Daten als Chromosom ist eine geordnete Liste der Städte

| | | |
|---|---|---|
| 1) Darmstadt | 2) Berlin | 3) Regensburg |
| 4) Essen | 5) Neukirchen | 6) Hamburg |
| 7) Kiel | 8) Bielefeld | ... |

| | |
|---|---|
| Liste 1 | (3 5 7 2 1 6 4 8) |
| Liste 2 | (2 5 7 6 8 1 3 4) |

# Problem des Handlungsreisenden: GA

Nachkommen können nun durch Rekombination und Mutation erzeugt werden

| | |
|---|---|
| Elternteil 1 | (3 5 7 2 1 6 4 8) |
| Elternteil 2 | (2 5 7 6 8 1 3 4) |
| Nach Rekombination | (3 5 7 2 1 6 8 4 ) |

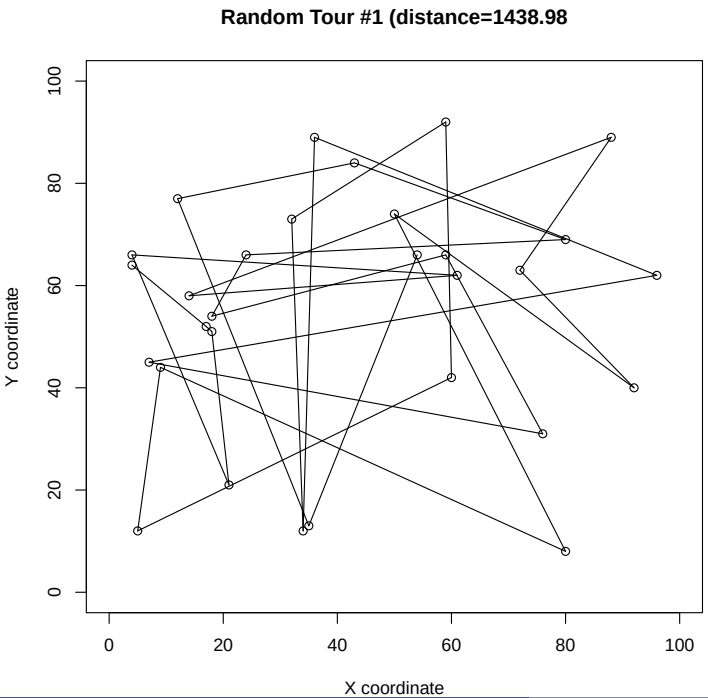
- Crossover kombiniert Inversion und Rekombination, damit das Chromosom des Nachkommens valide ist
- Mutation beinhaltet einen Austausch zweier Elemente

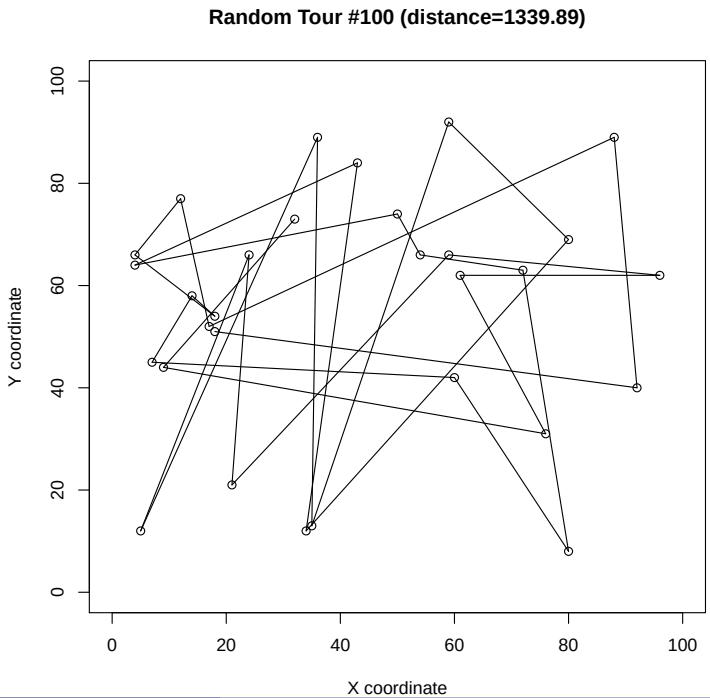
| | |
|---|---|
| Vorher | (3 5 7 2 1 6 8 4 ) |
| Nachher | (3 5 6 2 1 7 8 4 ) |

- Die Evaluation ist einfach die Berechnung der euklidschen Distanz der Städte entlang des Pfades

# Problem des Handlungsreisenden: GA

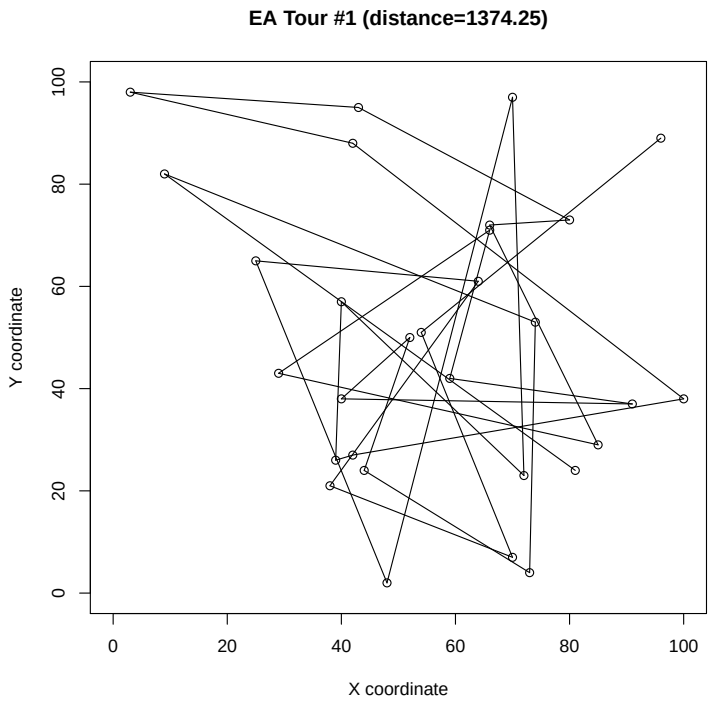
- Hier probieren wir zum Vergleich eine (zufällige) Ordnung der Städte nacheinander
- Falls eine Ordnung besser ist als das bisherige beste Ergebnis, ersetzen wir das beste Ergebnis dadurch
- Wir beobachten eine recht langsame Verbesserung des Ergebnisses

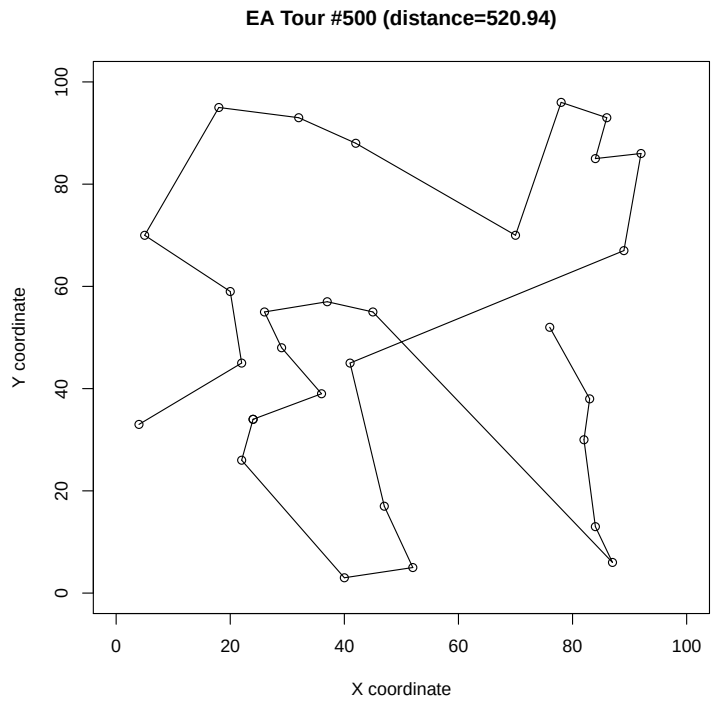


# Problem des Handlungsreisenden: GA

- Hier zeigen wir eine Implementierung eines simplen EAs
- Wesentliches schnellere Konvergenz zu einer guten Lösung



# Evolutionäre Algorithmen

Evolutionäre Algorithmen bilden die Prinzipien der natürlichen Evolution im Computer nach. Sie stellen eine wichtige Klasse von Optimierungsalgorithmus dar, und bieten eine Reihe von interessanten Parallelen zur biologischen Evolution.

| "Zutat" | "Biologie" |
|---|---|
| Kodierung | Gen, Chromosome |
| Initialisierung | Erschaffung |
| Evaluierung | Selektion |
| Selektion der Eltern | Reproduktion |
| Genetische Operatoren | Mutation, Rekombination |

# Evolution & die Genauigkeit der Replikation

- Kommen wir nun zu unserer Population von replizierenden RNA-Molekülen zurück
- Sei $\mathscr{S}$ eine RNA-Sequenz, die schneller als andere RNA-Moleküle Kopien von sich erzeugt
- Sei $R$ die Replikationsrate von Sequenz vom Typ $\mathscr{S}$
- Alle anderen RNA-Sequenzen erzeugen Kopien mit der Replikationsrate $r < R$.
- Wir interessieren uns nun für die Bedeutung der Genauigkeit der Replikation. Sei $Q$ die **Wahrscheinlichkeit, dass eine Sequenz eine genaue Kopie von sich erzeugt**

# Evolution & die Genauigkeit der Replikation

- Sei $x$ die Anzahl von RNA-Molekülen vom Typ $\mathscr{S}$
- Sei $y$ die Anzahl von RNA-Molekülen, die nicht vom Typ $\mathscr{S}$ sind
- Dann kann die Rate der Veränderung von $x$ gegeben werden durch

$$\frac{\mathrm{d}x}{\mathrm{d}t} = RQx - Dx \tag{1}$$

- Die Geschwindigkeit des "Todes" der RNA-Molekülen vom Typ $\mathscr{S}$ wird durch $Dx$ wiedergegeben, wobei $D$ die Rate des Todes pro Zeit darstellt
- die Rate der Veränderung von $y$ kann (unter der Annahme, dass $y$ nicht zu $x$ "rückmutiert") gegeben werden durch

$$\frac{\mathrm{d}y}{\mathrm{d}t} = ry + R(1-Q)x - Dy \tag{2}$$

# Evolution & die Genauigkeit der Replikation

Wir wollen wissen, ob die “optimalen” RNA-Moleküle vom Typ $\mathscr{S}$ überleben können bei einer bestimmten Replikationsgenauigkeit $Q$?

- Wir können nun die beiden Gleichungen addieren

$$\begin{aligned}\frac{\mathrm{d}x+y}{\mathrm{d}t} &= RQx - Dx + ry + R(1-Q)x - Dy \\ &= Rx + ry - D(x+y)\end{aligned}$$

# Evolution & die Genauigkeit der Replikation

- Nachdem ein Gleichgewicht erreicht worden ist, d..h., “Geburten” gleichen “Todesfällen”, wie es langfristig der Fall sein muss, gilt $\frac{\mathrm{d}x+y}{\mathrm{d}t} = 0$.
- Sei $p = x/(x+y)$, d.h., der Anteil von “optimalen” RNA-Molekűlen in der Population

$$\begin{aligned} 0 &= Rx + ry - D(x+y) \\ D &= \frac{Rx+ry}{x+y} \\ &= Rp + r(1-p) \end{aligned}$$

# Evolution & die Genauigkeit der Replikation

- Kommen wir nun zur Gleichung für die Anzahl von RNA-Molekülen vom Typ $\mathscr{S}$ zurück und verwenden wir den Ausdruck für $D$

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= RQx - Dx \\ &= RQx - \{Rp + r(1-p)\} \cdot x\end{aligned}$$

- Beim Gleichgewicht gilt

$$\begin{aligned}0 &= RQx - \{Rp + r(1-p)\} \cdot x \\ RQx &= \{Rp + r(1-p)\} \cdot x \\ Q &= p + \frac{r}{R}(1-p)\end{aligned}$$

# Evolution & die Genauigkeit der Replikation

Der Anteil von RNA-Molekülen vom Typ $\mathscr{S}$ (d.h., $p$) sinkt als die Genauigkeit der Replikation $Q$ sinkt. Falls $p \to 0$, dann

$$\begin{aligned} \lim_{p\to 0} Q &= \lim_{p\to 0} p + \frac{r}{R}(1-p) \\ &\approx \frac{r}{R} \end{aligned}$$

- Im Schnitt muss jedes RNA-Molekül vom Typ $\mathscr{S}$ mindestens eine perfekte Kopie von sich erzeugen
- Falls jedes RNA-Molekül im Schnitt $R$ Kopien produziert, dann ist es erforderlich, dass $Q > 1/R$.
- Es erscheint plausibel, dass $Q \geq 1/2$.

# Genauigkeit der Replikation & Genomgröße

- Betrachten wir nun ein Genom mit $n$ Nukleotiden
- Sei $u$ die Fehlerwahrscheinlichkeit für die Replikation eines einzelnen Nukleotids
- Ein RNA-Molekül erzeugt eine perfekte Kopie von sich, wenn es zu keinem einzelnen Fehler bei der Replikation von $n$ Nukleotiden kommt

$$Q = (1-u)^n \approx e^{-nu} \tag{3}$$

- Damit $Q \geq 1/2$, muss (ungefähr) $nu < 1$ sein[1]

[1] $e^{-1} \approx 0.3678794$.

# Genauigkeit der Replikation & Genomgröße

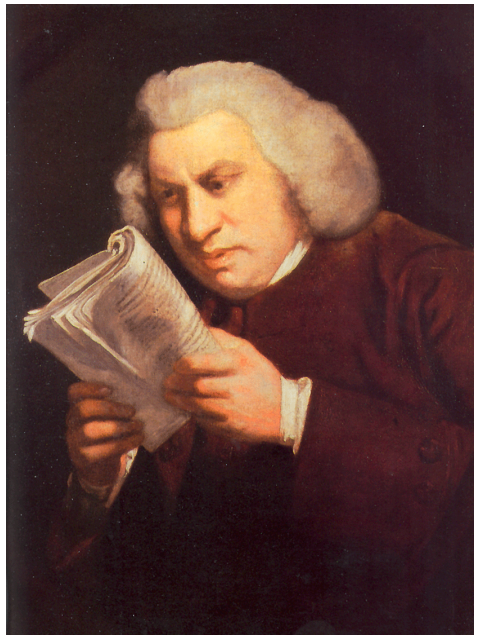
| | Fehlerrate ($u$) | Genomgröße ($1/u$) |
|---|---|---|
| Nicht enzymatische RNA-Replikation | 1/10 – 1/100 | ? |
| RNA-Replikation | $10^{-3}$ bis $10^{-4}$ | ca. $10^4$ |
| DNA-Replikation | $10^{-9}$ bis $10^{-10}$ | ca. $10^9$ |

- Dies entspricht in etwa den Größen der bekannten RNA Viren
- Dies erklärt weshalb die Genomgröße der höheren Eukaryonten nicht wesentlich über eine Gigabase hinausgeht
- Wirft Fragen über den Ursprung des Lebens auf: Vor dem ersten Enzym hätte demnach kein Genom mit über 100 Nukleotiden sich stabil vererben können, aber 100 Nukleotide reichen kaum aus um eine RNA-Replikase zu kodieren...

# Evolution

# The End of the Lecture as We Know It

- Kontakt:
  peter.robinson@charite.de
- Aufgabe zu Übung 1



Lectures were once useful; but now, when all can read, and books are so numerous, lectures are unnecessary. If your attention fails, and you miss a part of a lecture, it is lost; you cannot go back as you do upon a book... People have nowadays got a strange opinion that everything should be taught by lectures. Now, I cannot see that lectures can do as much good as reading the books from which the lectures are taken. I know nothing that can be best taught by lectures, except where experiments are to be shown. You may teach chymistry by lectures. You might teach making shoes by lectures!

Samuel Johnson, quoted in Boswell's Life of Johnson (1791).