Der E-Wert

$$E = Kmn\, e^{-\lambda s}$$

von Bettina Hofmann und Kerstin Weilmünster

Der E-Wert

Gliederung:

1. Fragestellungen
2. Vorgehensweise
3. Experimente
4. Ergebnisse
5. Fazit

Der E-Wert

Fragestellungen

- Betrachtung unabhängiger Sequenzen in Bezug auf den E-Wert
- Betrachtung von Sequenzen mit Muster in Bezug auf den E-Wert
- Gibt der E-Wert tatsächlich die Anzahl zu erwartender Alignments mit gegebenem Score an?

Der E-Wert

Beispiel eines Blast-Ergebnisses

Score = 36.2 bits (18), Expect = 0.064
Identities = 18/18 (100%)
Strand = Plus / Minus

```
Query: 23792  cgcgcggacgcgtccgcg 23809
              ||||||||||||||||||
Sbjct:  34774  cgcgcggacgcgtccgcg 34757
```

Der E-Wert

- **Score**
  - Match – Belohnung
    - Mismatch – Bestrafung
    - Gap-Open – Bestrafung
    - Gap – Extention – Bestrafung
- **E-Wert**
  - $E = Kmn\, e^{-\lambda S}$
  - K und λ: Skalierungsparameter für Suchraum und Score-Matrix

Der E-Wert

Vorgehen:

- Erzeugen von Sequenzen
  - zufällig
  - mit Musteranteil
- Anwendung von Blast auf Sequenzen
- Filtern der E- und Score-Werte
- Auswertung mit R

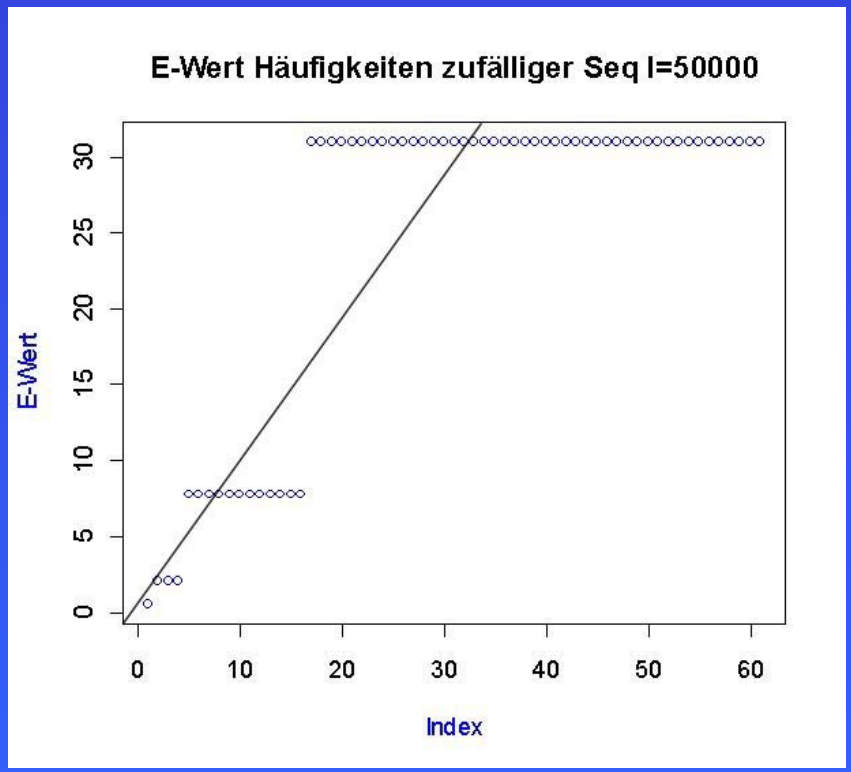
Der E-Wert

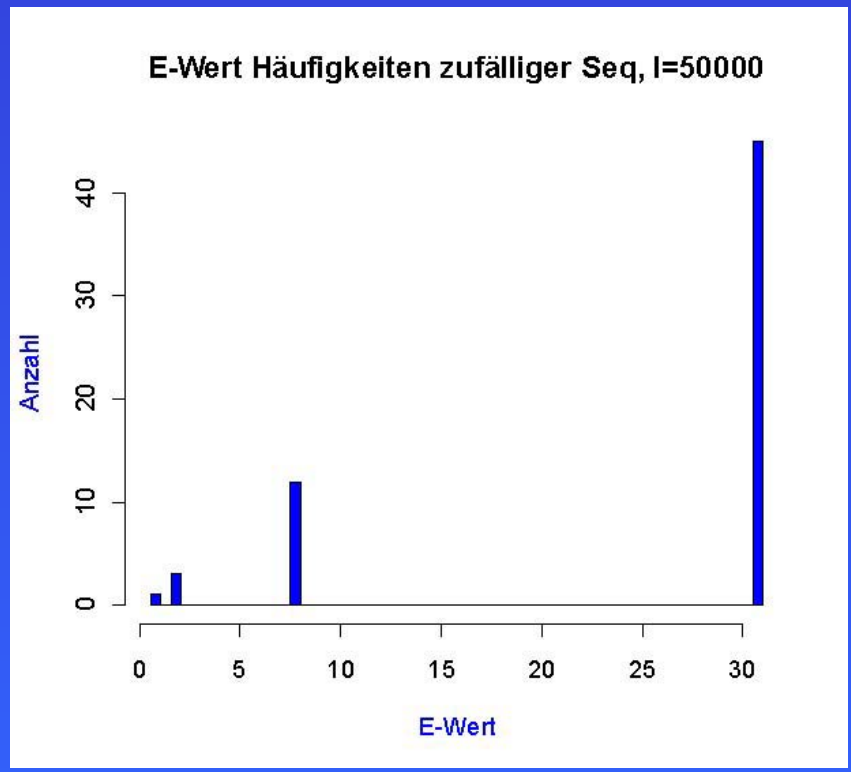
Ergebnisse:

- Zufällige Sequenzen erzeugt
- Erwartung:
  - wenige/kurze Alignments
  - wenn Alignments, dann hoher E-Wert, wenige niedrige E-Werte

Der E-Wert

Zufällige Sequenzen





Der E-Wert

Sequenzen mit Muster

Verwendetes Muster: CpG-Islands

- CpG-Anteil: ca. 50%
- Länge: einige hundert bis wenige tausend Basen

Nicht-CpG-Island-Sequenz enthält nur geringen CG-Anteil

Der E-Wert

Ergebnisse:

- Sequenzen mit Muster
- Erwartung:
  - Hoher Musteranteil – viele Alignments
  - Hoher Musteranteil – viele niedrige E-Werte

Der E-Wert

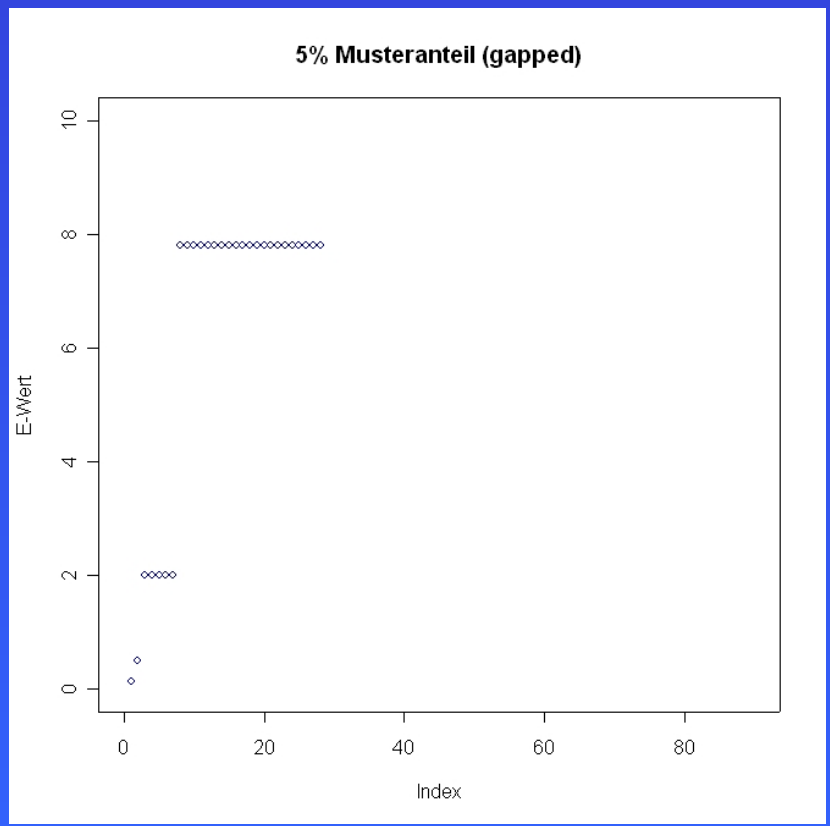
Sequenzen mit Muster

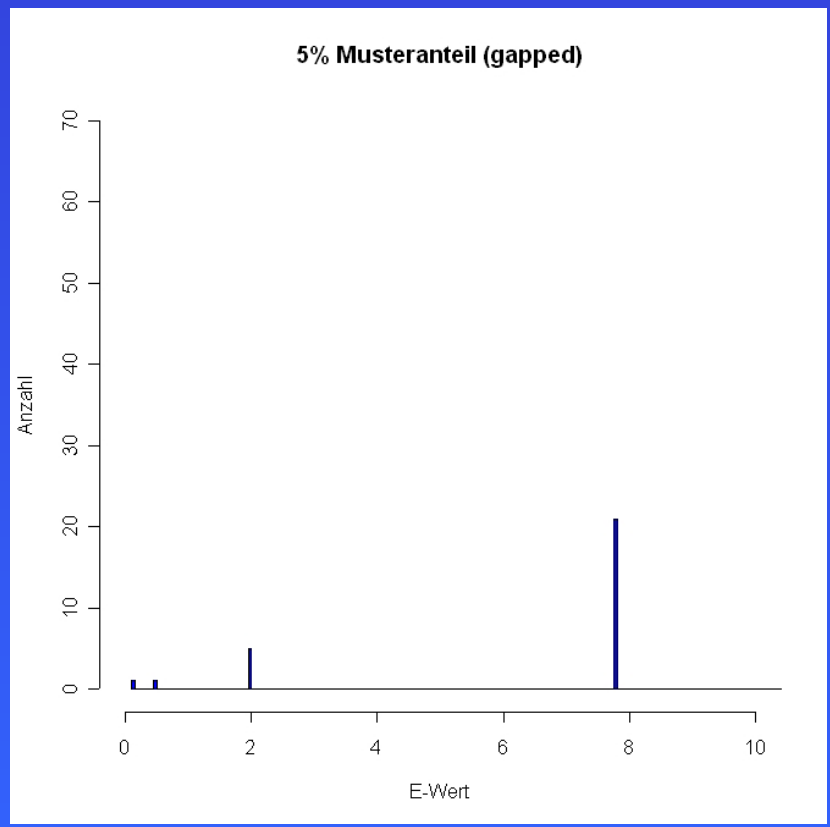
Überprüfter Musteranteil:

- 5 %
- 15 %
- 20 %
- 30 %

Der E-Wert

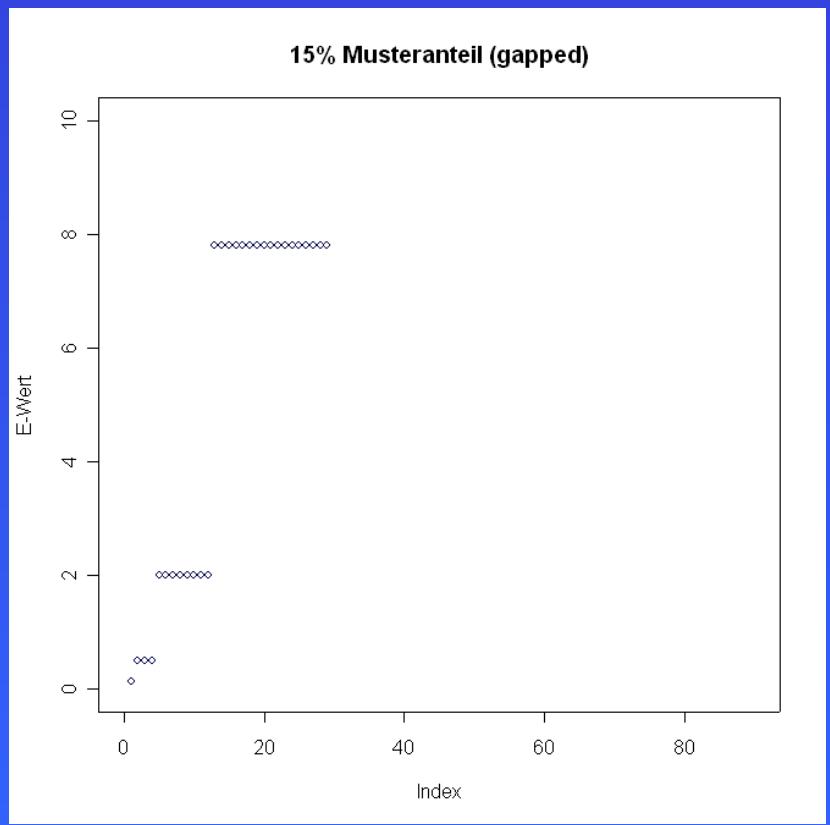
Sequenzen mit 5 % Musteranteil

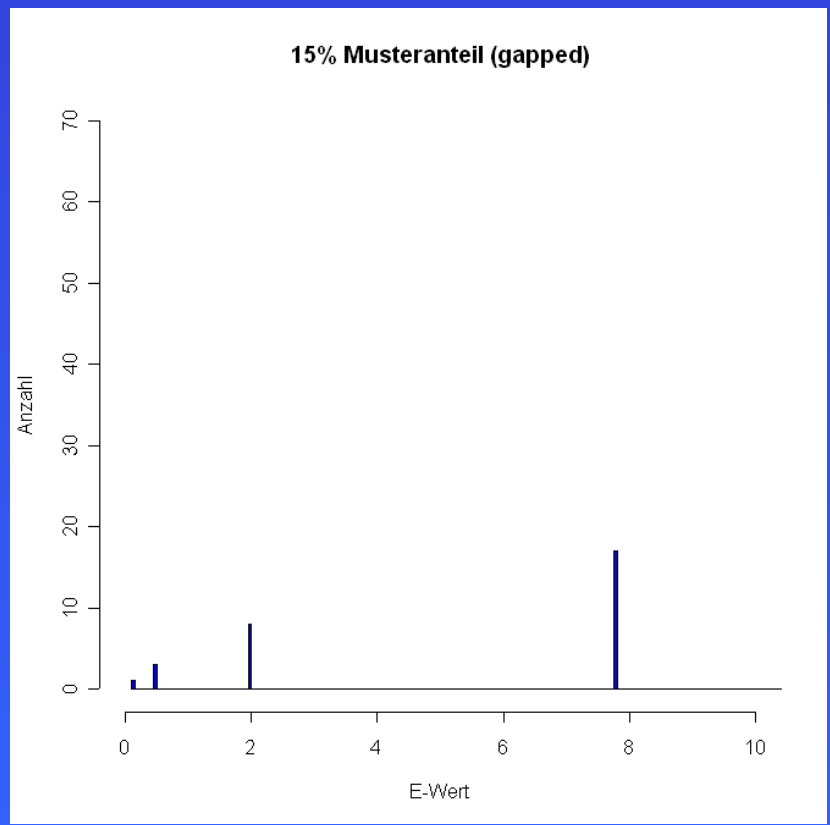




Der E-Wert

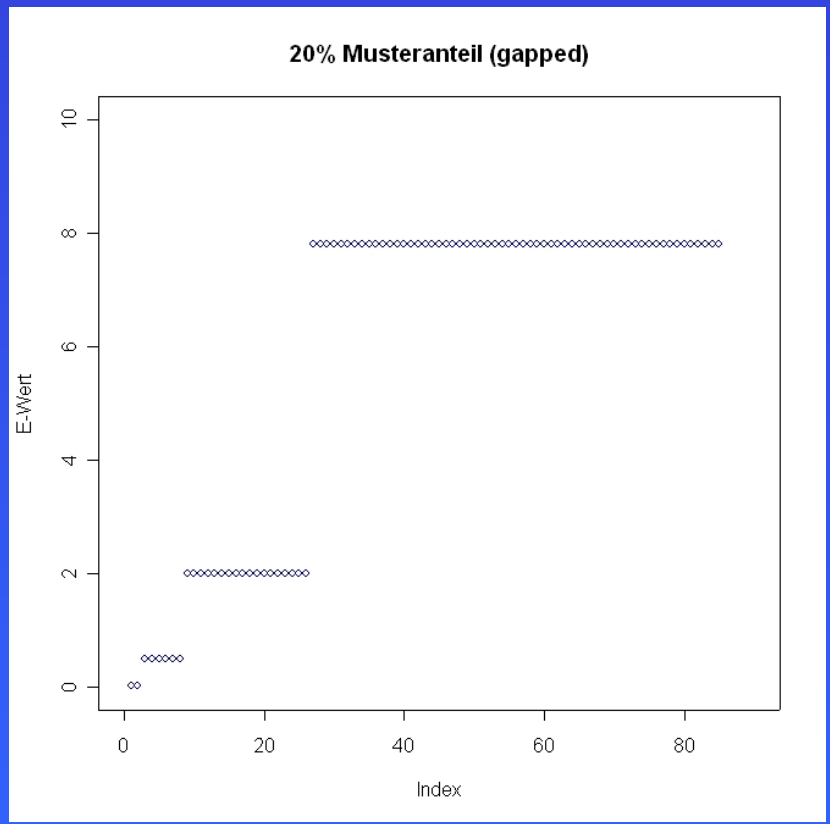
Sequenzen mit 15 % Musteranteil

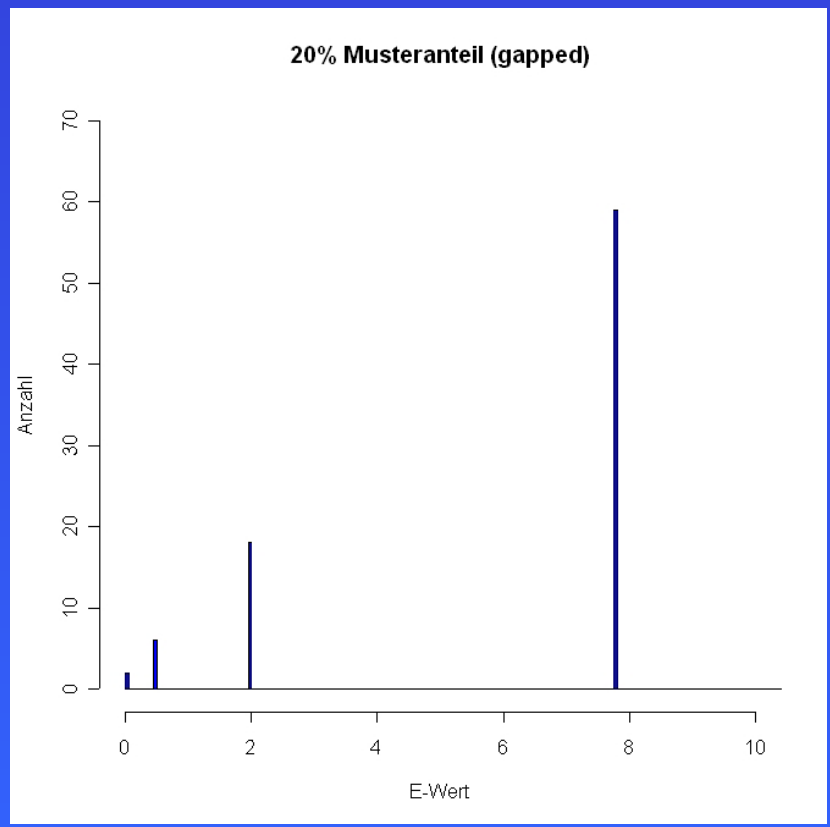




Der E-Wert

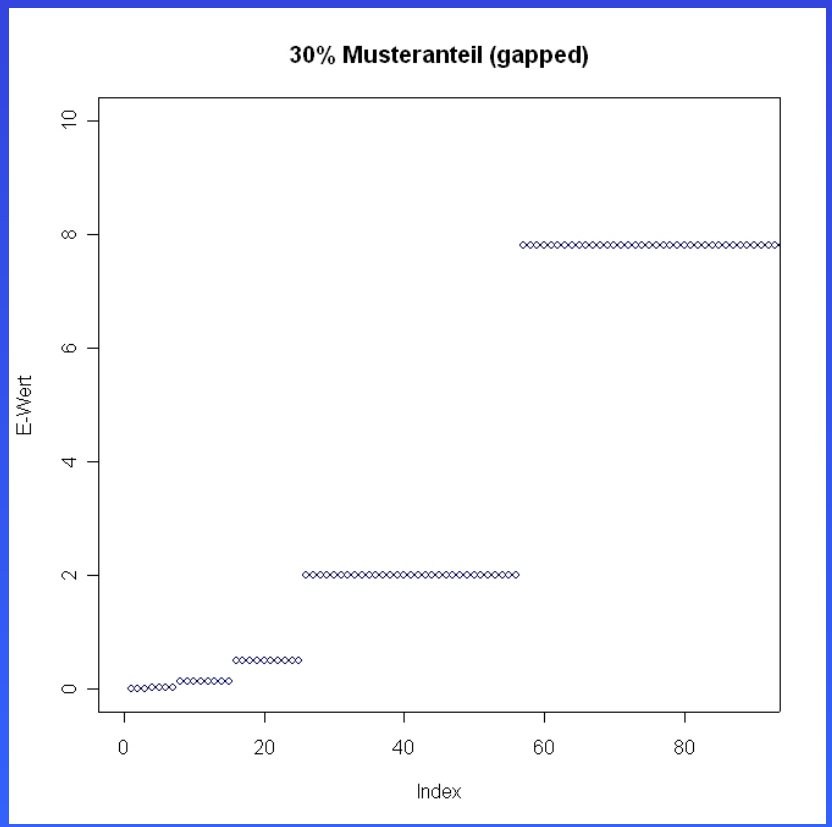
Sequenzen mit 20 % Musteranteil

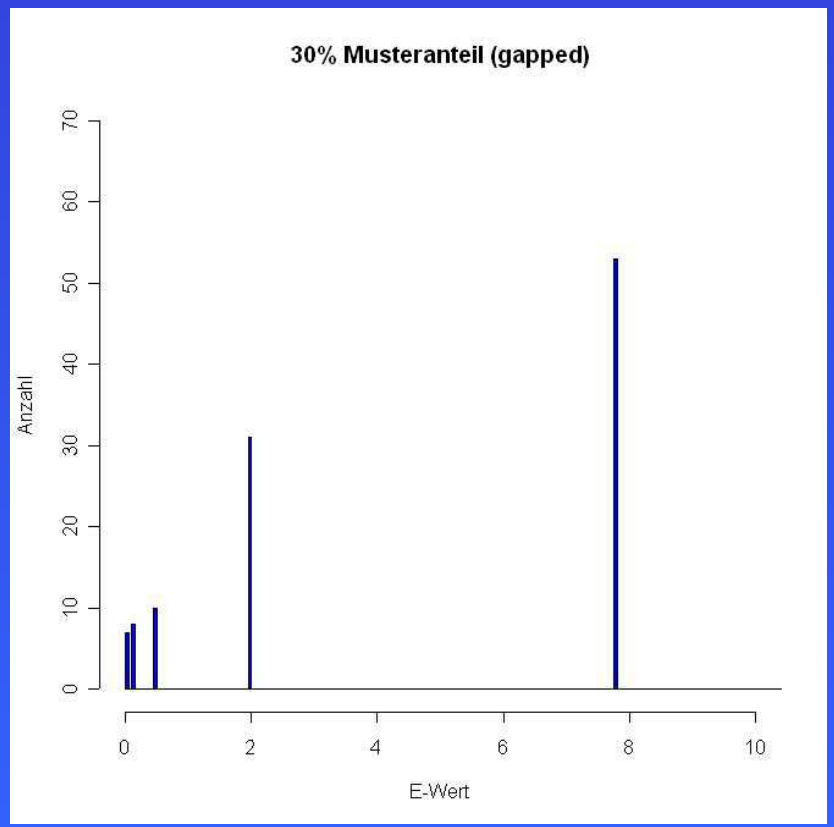




Der E-Wert

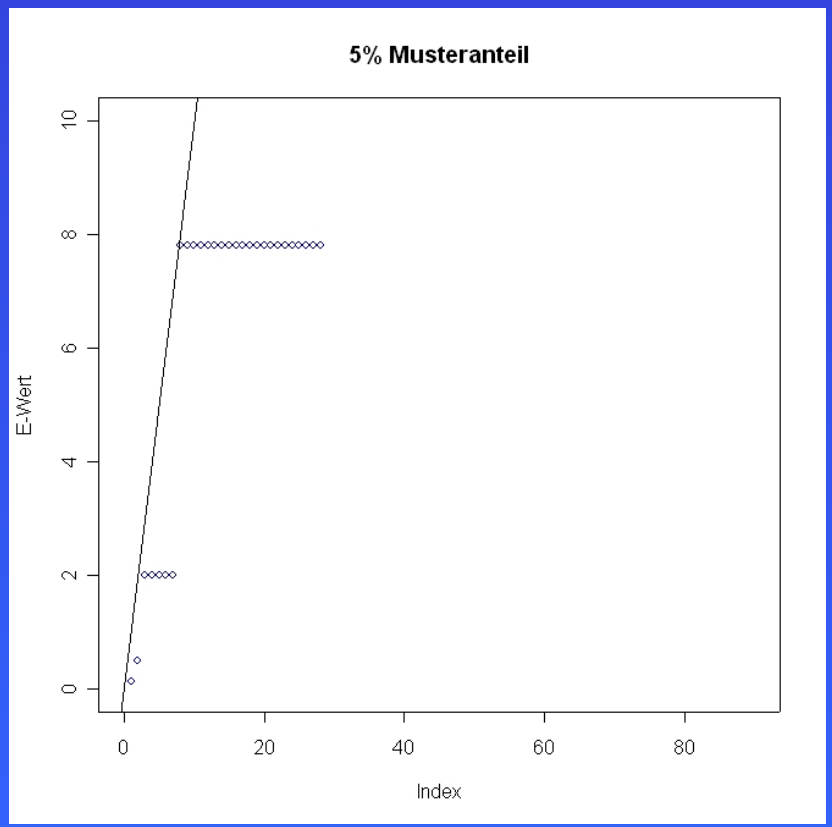
Sequenzen mit 30 % Musteranteil

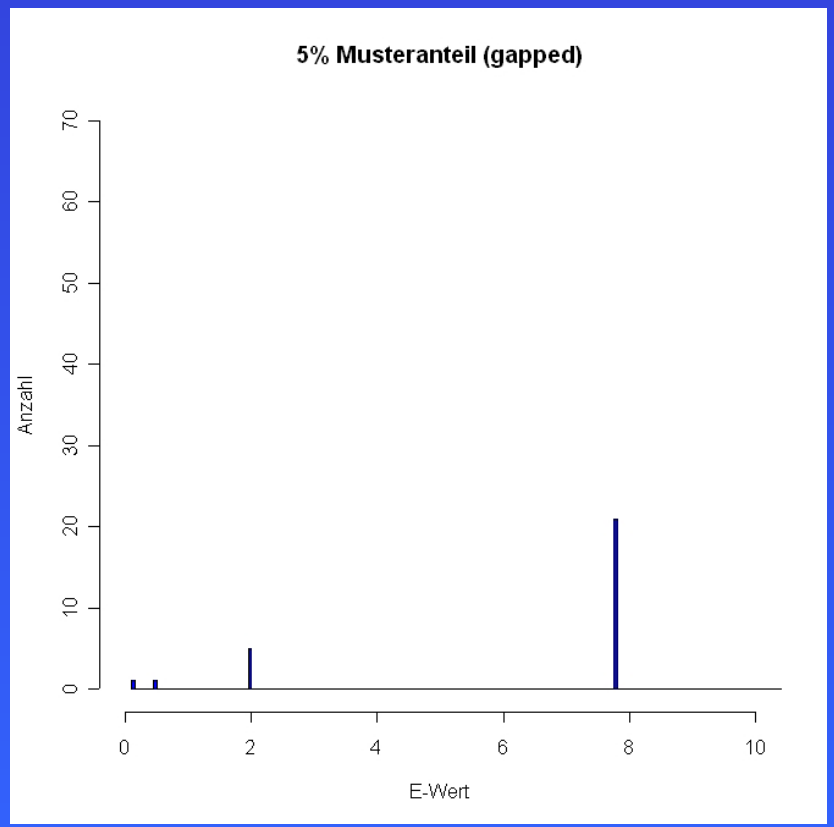




Der E-Wert

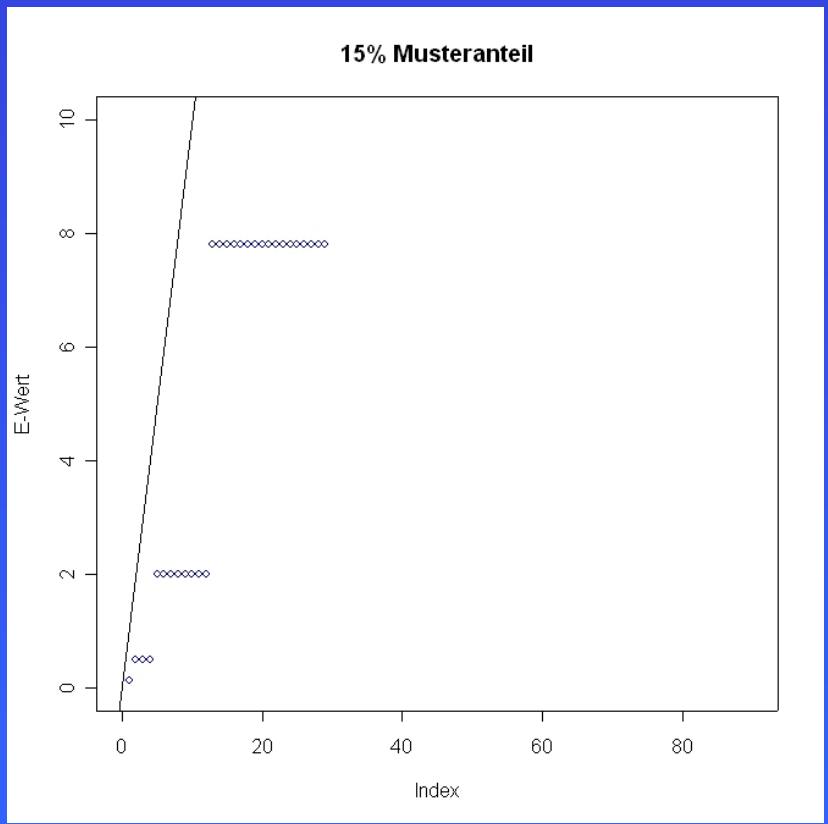
Sequenzen mit 5 % Musteranteil

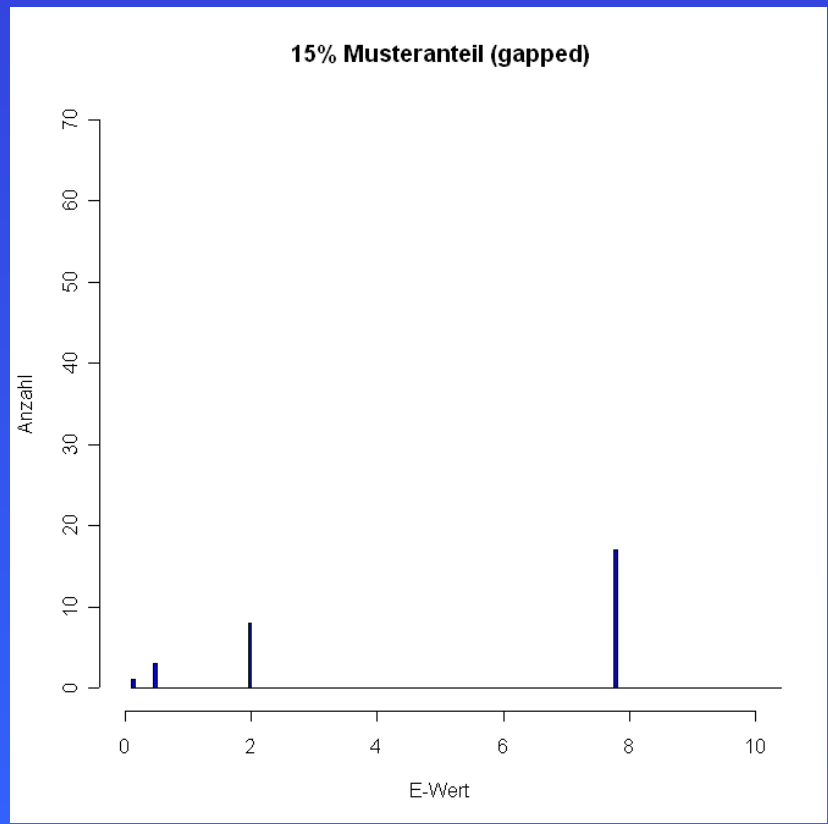




Der E-Wert

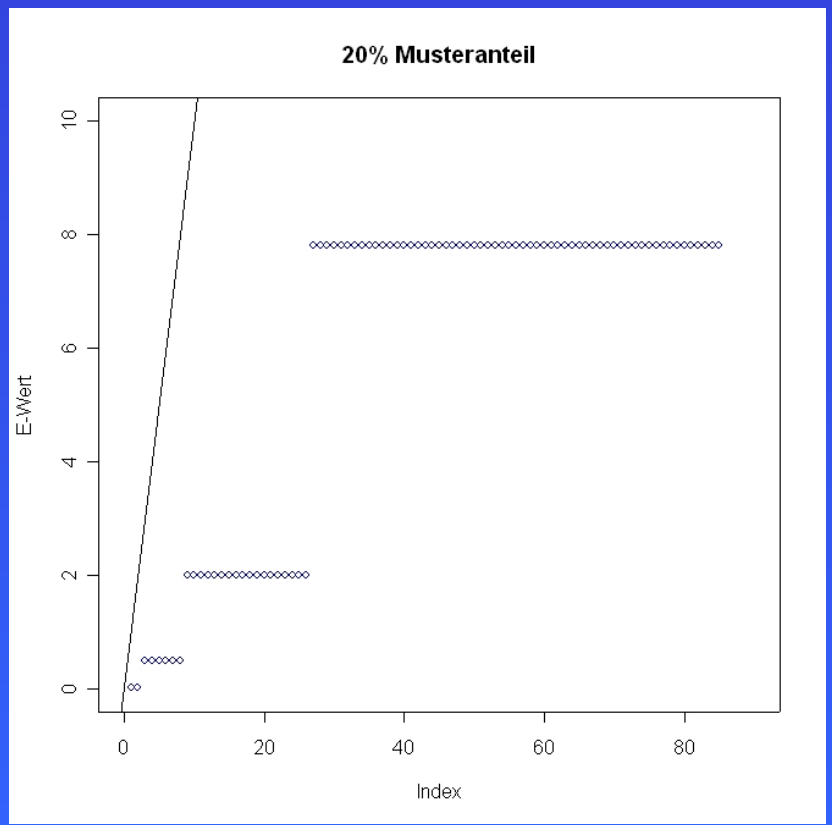
Sequenzen mit 15 % Musteranteil

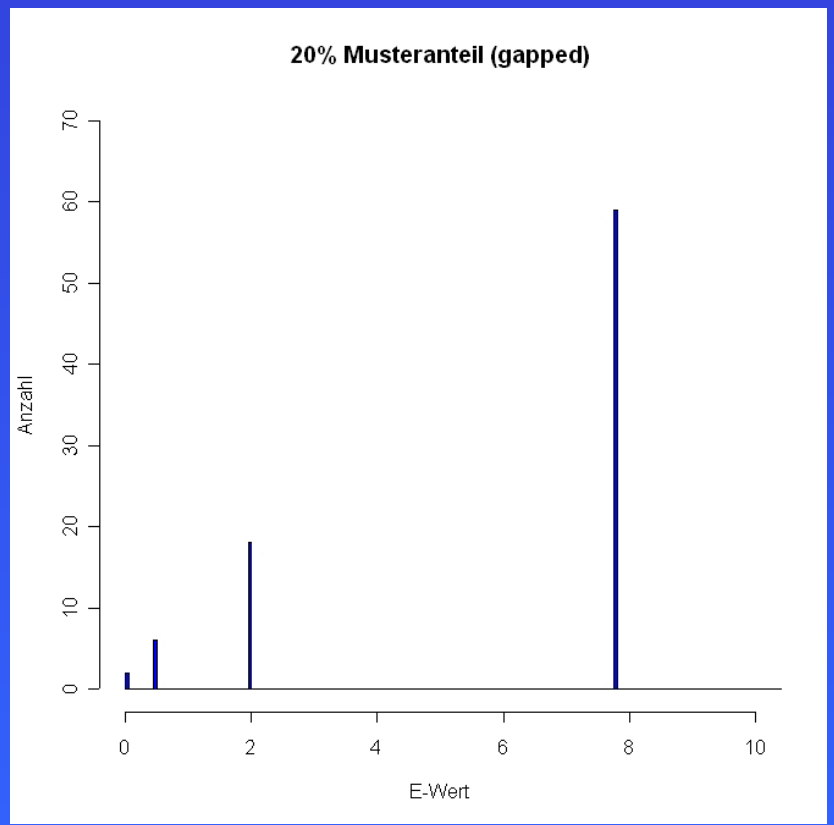




Der E-Wert

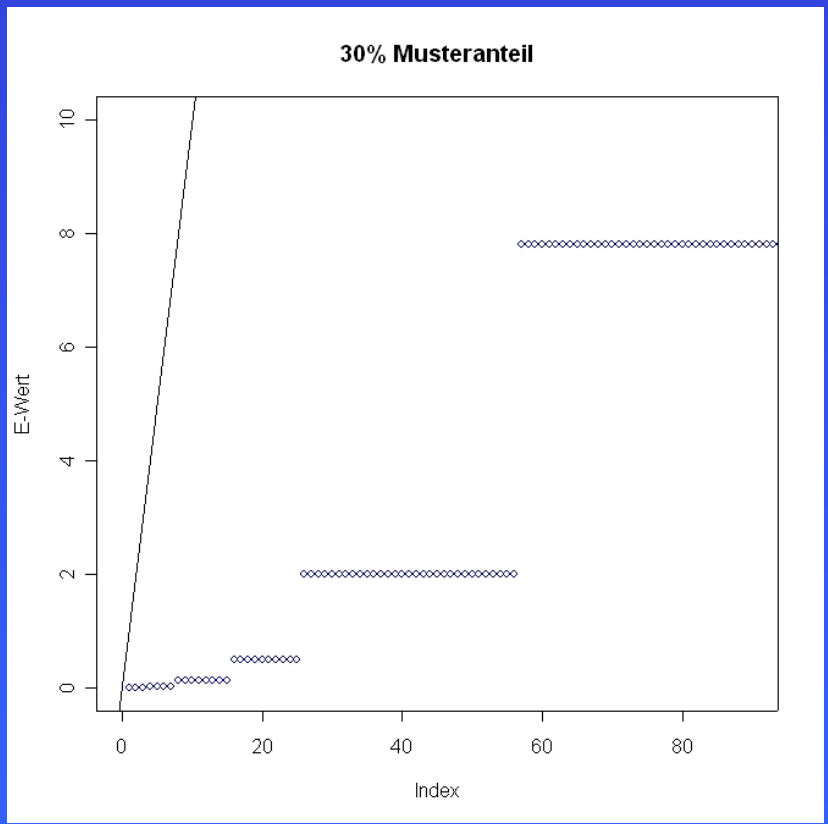
Sequenzen mit 20 % Musteranteil

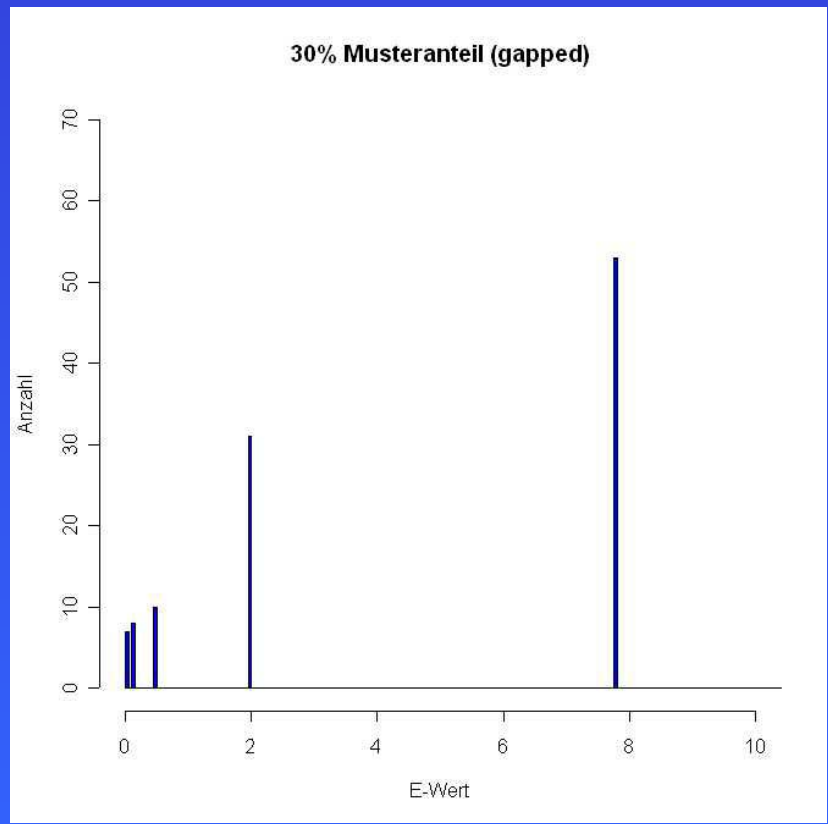




Der E-Wert

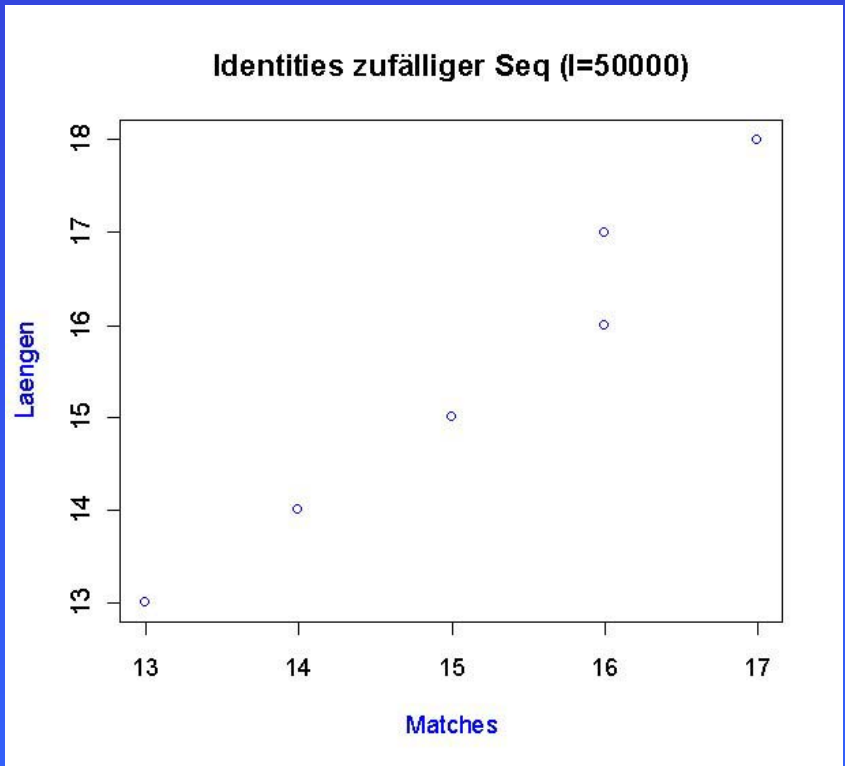
Sequenzen mit 30 % Musteranteil





Der E-Wert

Gegenüberstellung Alignmentlänge - Matches:



Score = 32.2 bits (16), Identities= 16/16

Score = 30.2 bits (15), Identities= 15/15

Score = 28.2 bits (14), Identities= 14/14

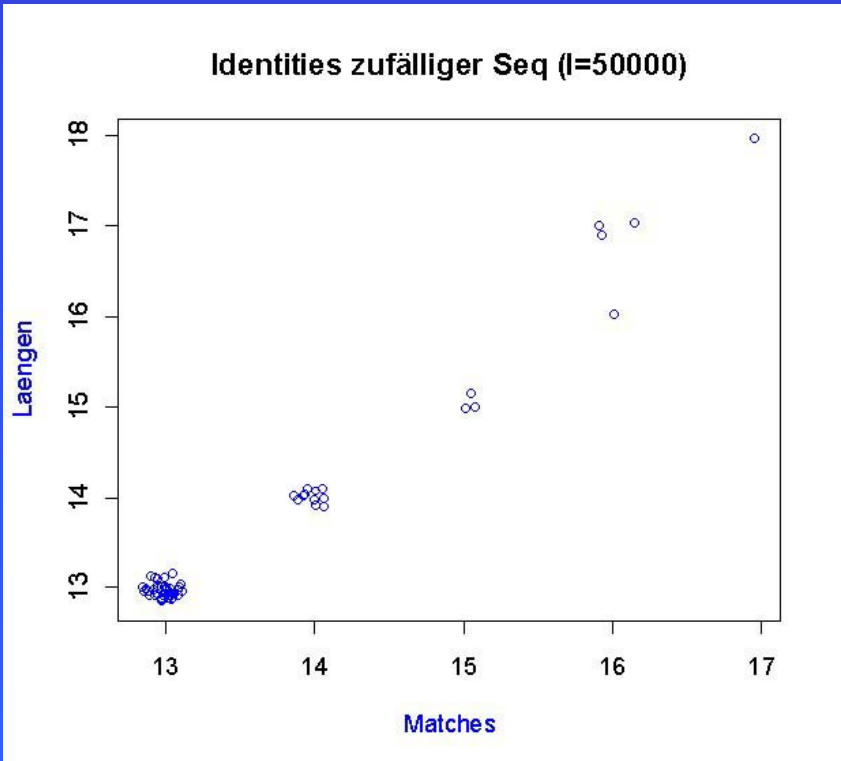
Score = 28.2 bits (14), Identities= 17/18

Score = 26.3 bits (13), Identities= 13/13

Score = 26.3 bits (13), Identities= 16/17

Der E-Wert

Gegenüberstellung Alignmentlänge - Matches:

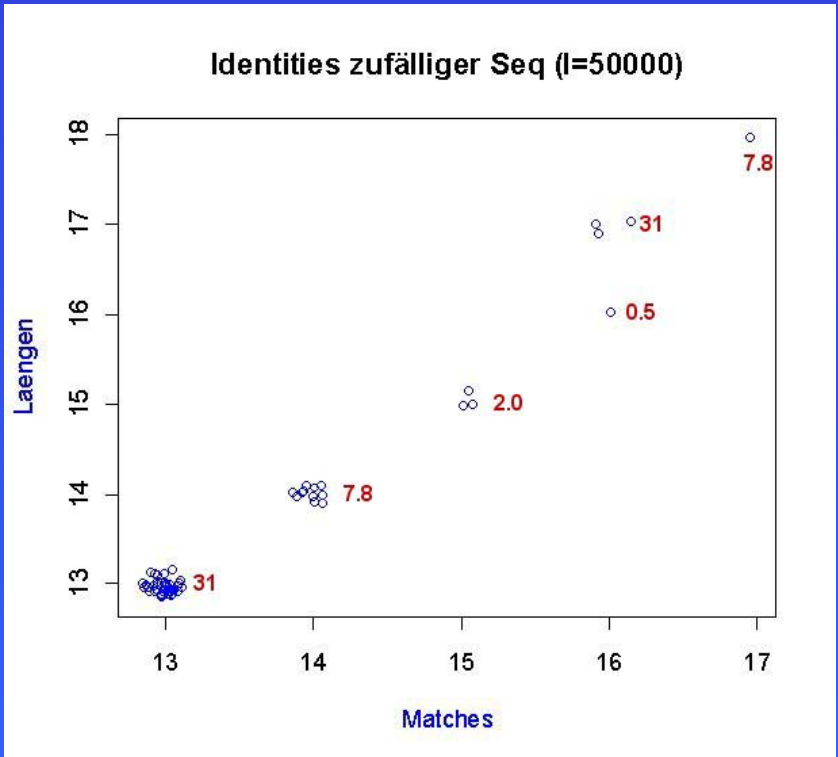


Jittern: streut Punkte zufällig nach Normalverteilung

plot(a+rnorm(60,0,0.1),
b+rnorm(60,0,0.1))

Der E-Wert

Gegenüberstellung Alignmentlänge - Matches:

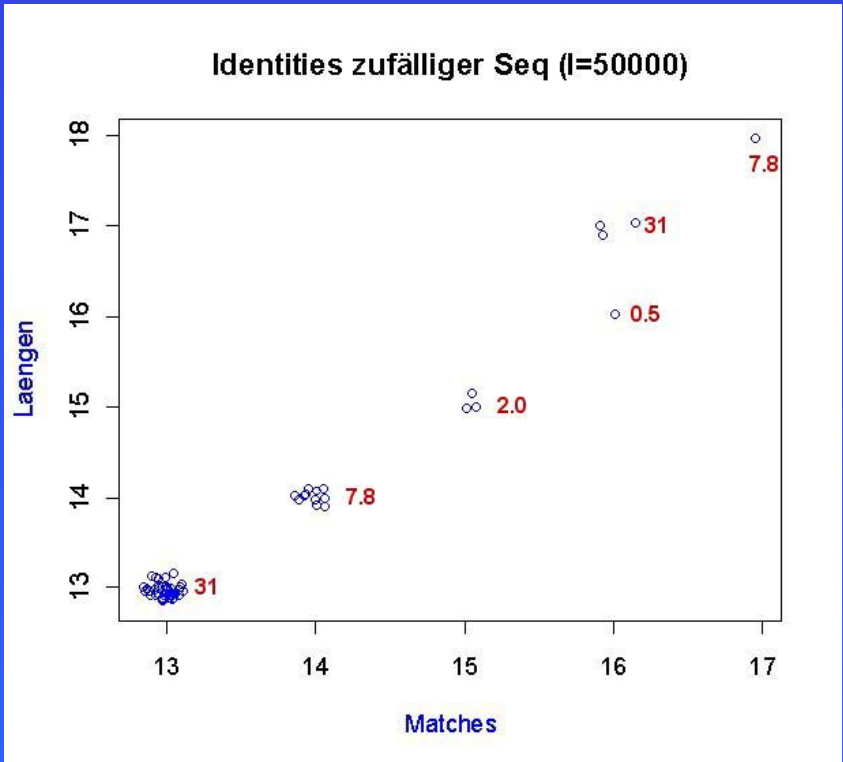


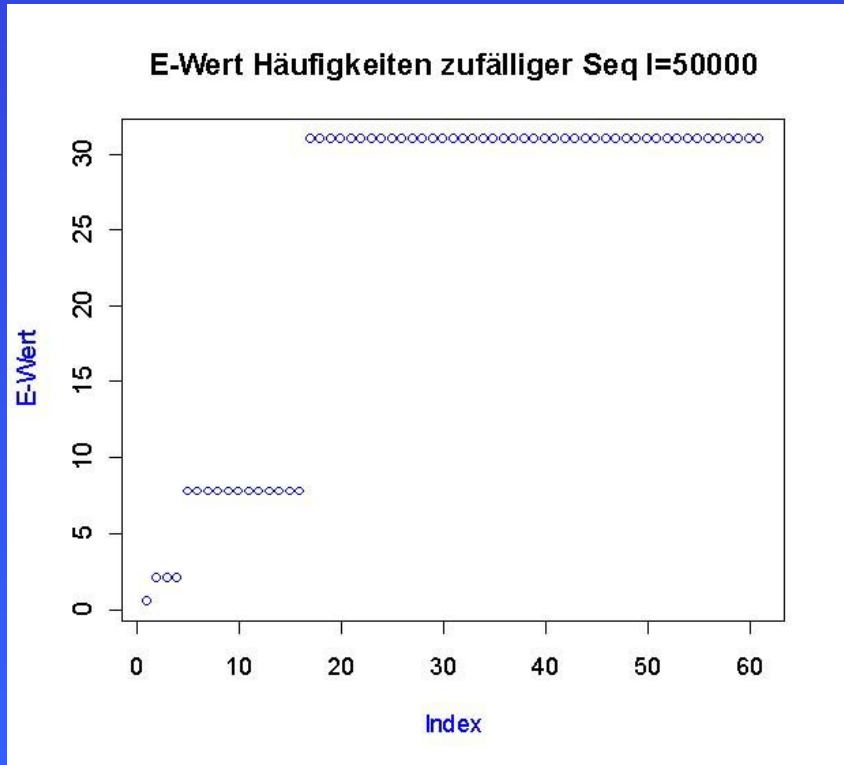
Jittern: streut Punkte zufällig nach Normalverteilung

```
plot(a+rnorm(60,0,0.1),
     b+rnorm(60,0,0.1))
```

Der E-Wert

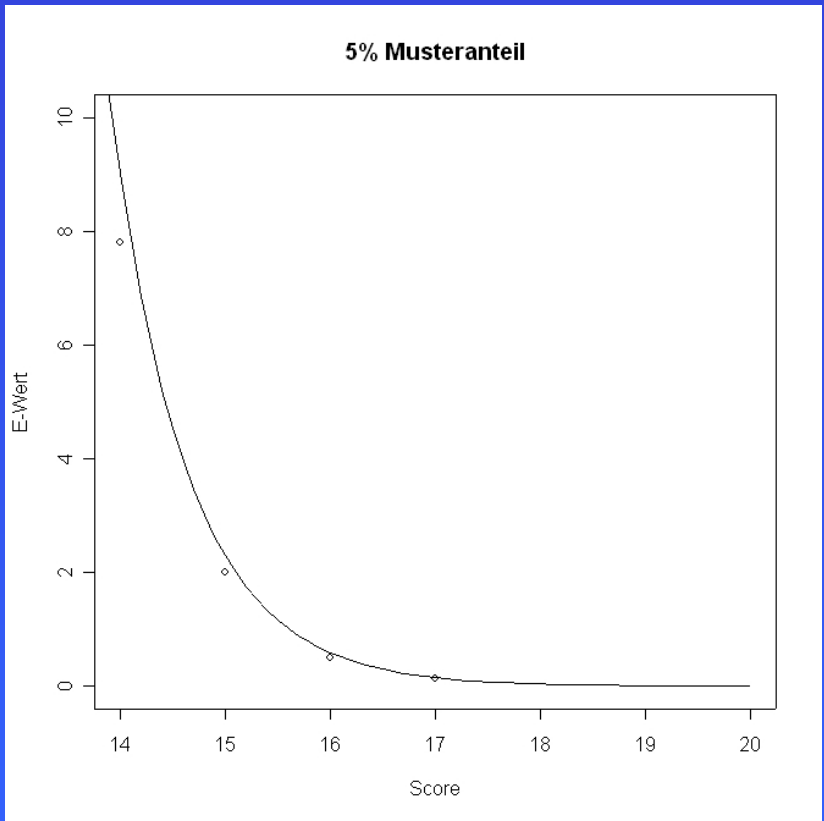
Gegenüberstellung Identität – E-Wert:

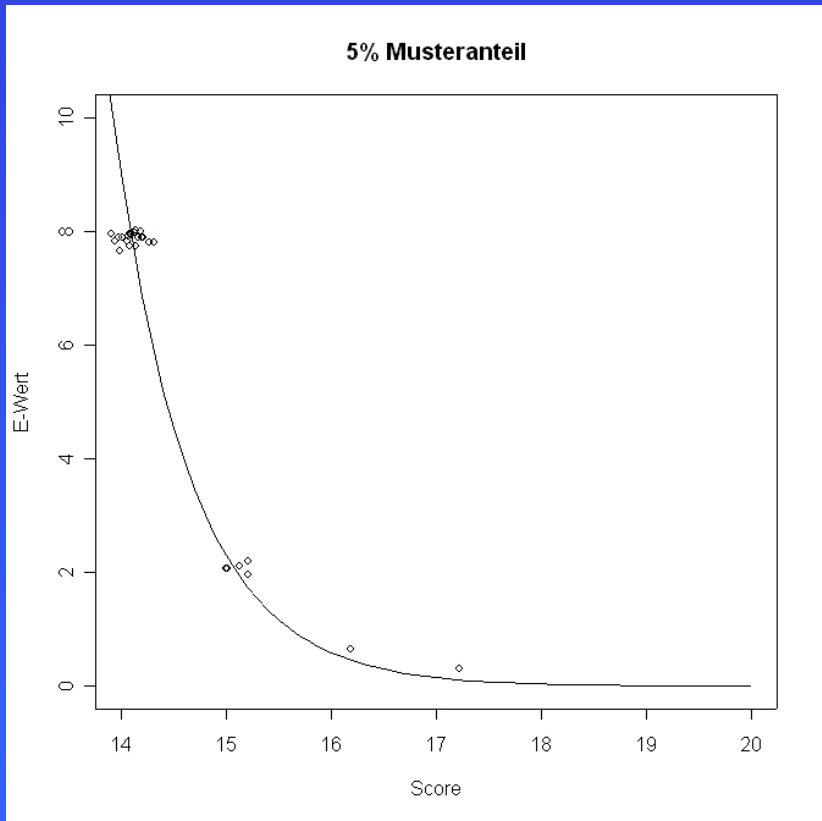




Der E-Wert

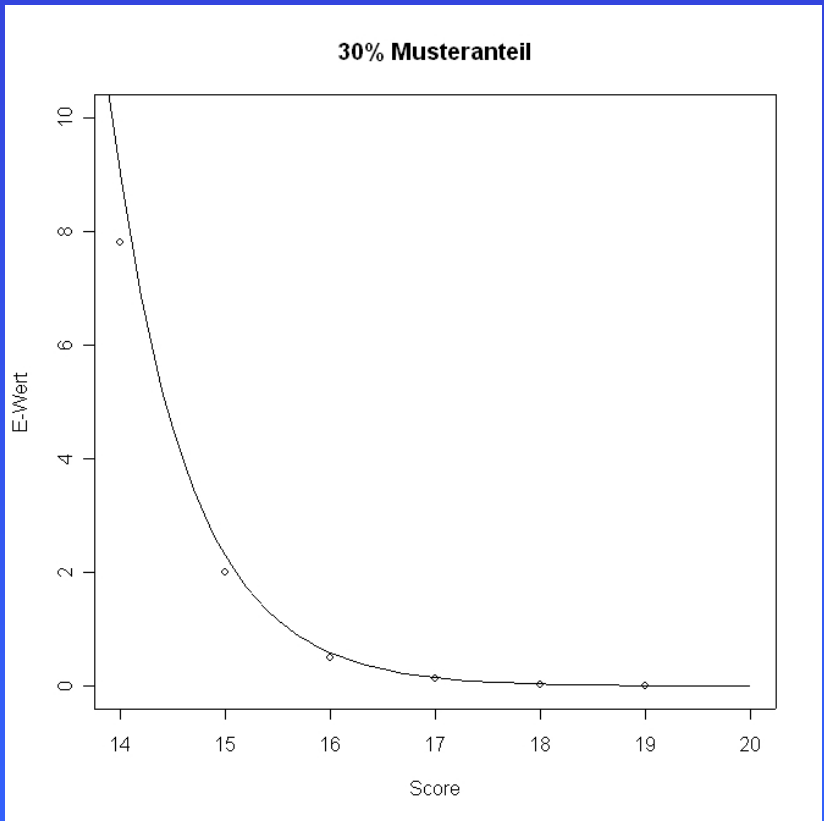
Gegenüberstellung Score – E-Wert: 5 %

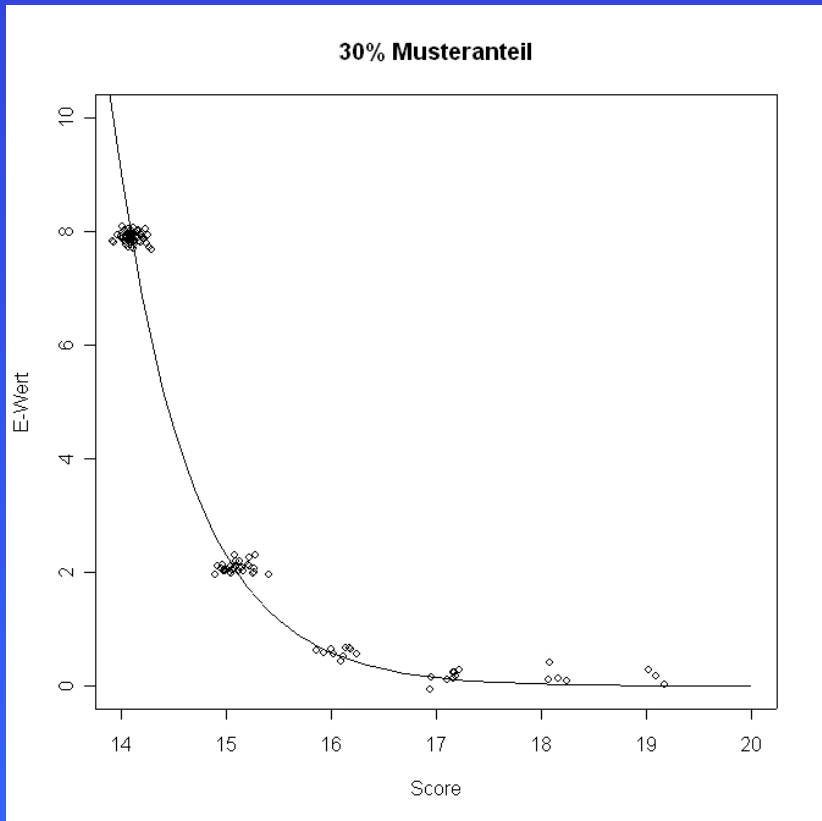




Der E-Wert

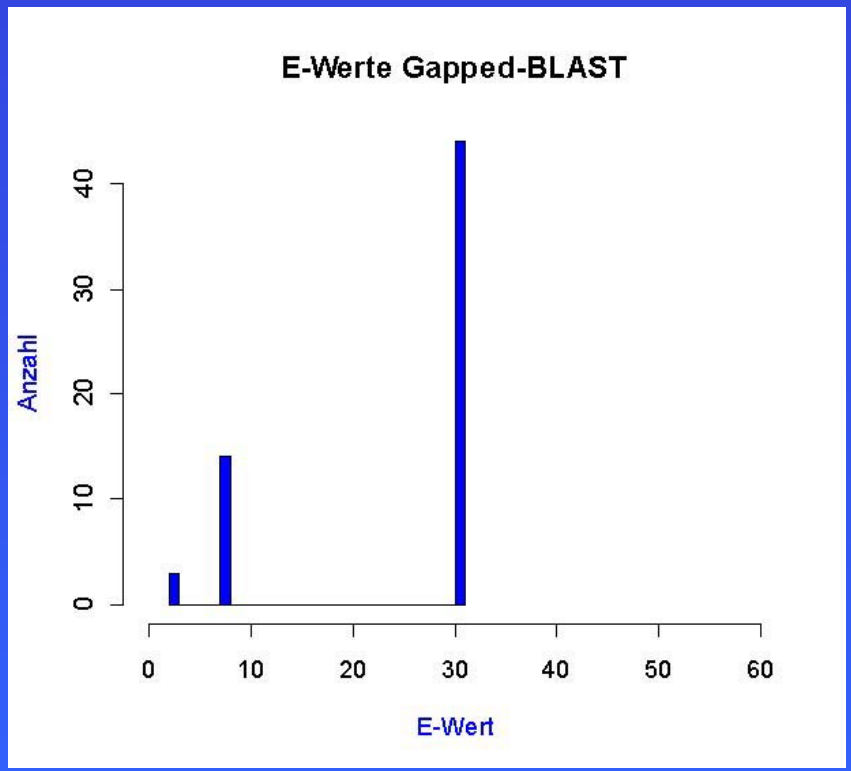
Gegenüberstellung Score – E-Wert: 30 %

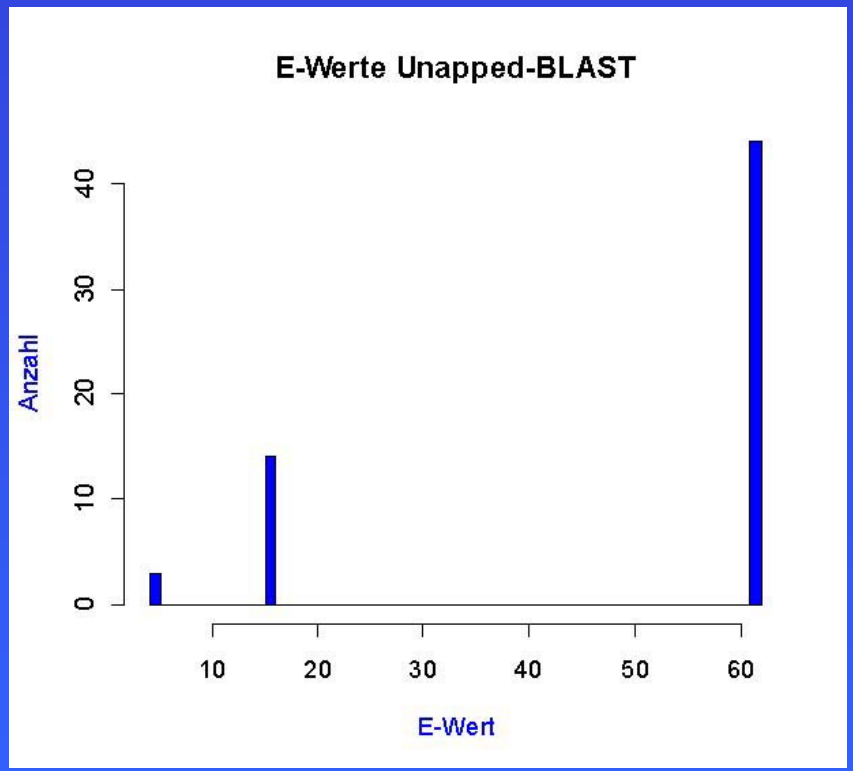




Der E-Wert

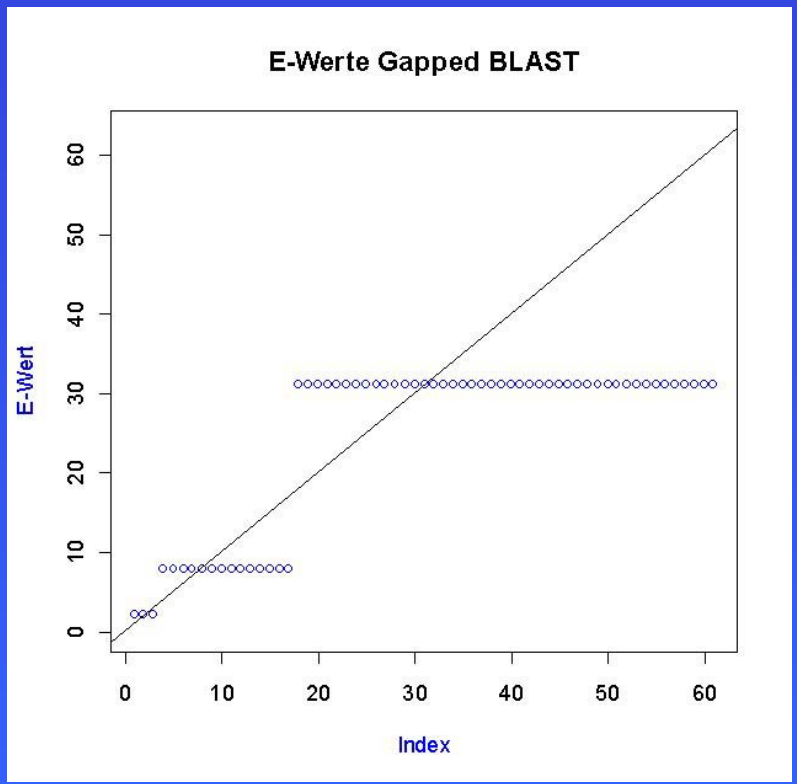
Unterschiede Gapped und gaploser Blast

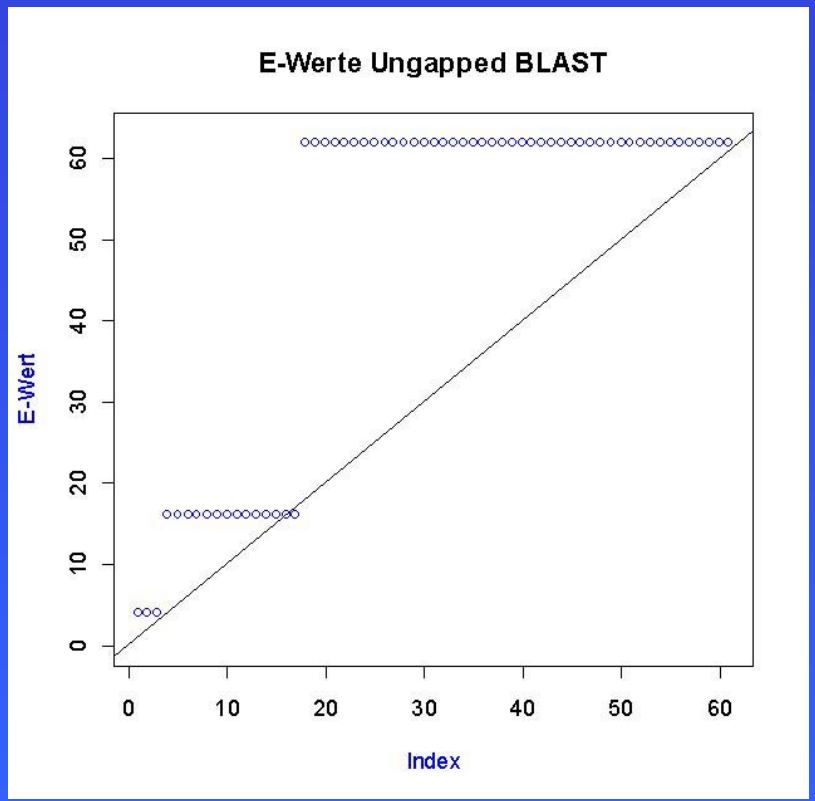




Der E-Wert

Unterschiede Gapped und gaploser Blast





Der E-Wert

Weitere Betrachtungen:

- **Inverse Sequenzen**
- **Verschiedene Sequenzlängen**

Sequenzlänge	E-Wert
1000	0.19
50000	484

- **Mikrosatelliten/repetitive Sequenzen:**
  - Low Complexity Filter, DUST

Der E-Wert

Fazit:

- **Beim Vergleich unabhängiger Sequenzen findet man wenig niedrige und viele hohe E-Werte**
- **Beim Vergleich von Sequenzen mit Mustern nimmt die Anzahl der gefundenen Alignments insgesamt zu und man findet bessere (niedrigere) E-Werte, je höher der Musteranteil ist**
- **Die Aussage, dass der E-Wert die Anzahl zu erwartender Alignments mit gegebenem oder besserem Score in zufälligen Sequenzen angibt, trifft tendenziell zu**