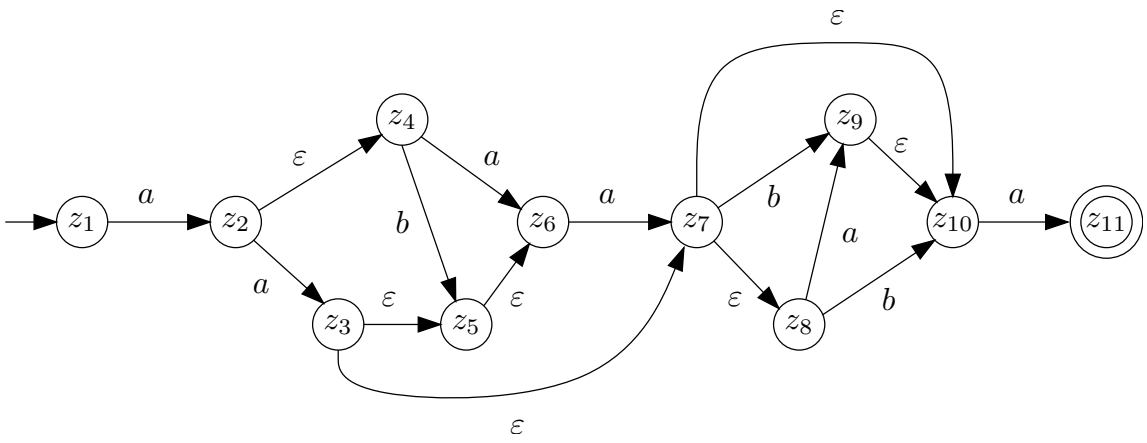
1. Aufgabe 6 Punkte [NFA / RE / PROSITE]

In dieser Aufgabe betrachten wir modifizierte Prosite-Patterns, bei denen das Aminosäurenalphabet durch das Alphabet $\Sigma = \{a, b\}$ ersetzt wird. – Geben Sie (in diesem Sinn) ein Prosite-Pattern an, welches die Sprache beschreibt, die von dem unten stehenden NFA mit spontanen (ε-) Übergängen akzeptiert wird. Erklären Sie Ihren Lösungsweg.



Lösung: a-x(0,1)-a-x(0,1)-a. Findet man, indem man die Knoten der Reihe nach mit den Wörtern beschriftet, die zu ihnen führen.

Oder wie in der Übungsaufgabe, durch Elimination der ε-Übergänge. Zum Beispiel die Knoten z_7, z_8, z_9, z_{10} kann man durch x(0,1) beschreiben. Der Übergang von z_4 nach z_5 kann nach z_6 umgebogen werden, dann kann man z_5 einsparen. Indem man einfach alle möglichen Pfade betrachtet, sieht man, dass das Teilstück von z_2 bis z_7 durch x(0,1)-a ausgedrückt werden kann. Etc.

---

2. Aufgabe 7 Punkte [Aho-Corasick]

Für einen Text $T[1 \,..\, n]$ und eine Menge von Patterns $P = \{P_1, \ldots, P_p\}$ bezeichnen wir die Menge der Matches mit $M := \{\, (i,j) \mid T[i \,..\, i+|P_j|-1] = P_j \,\}$. Die Gesamtlänge aller Patterns sei mit $N := \sum_{j=1}^{p} |P_j|$ bezeichnet. – Beschreiben Sie eine Konstruktion von T und P, in Abhängigkeit von n, so dass $n+N = o(|M|)$, d.h. $\frac{|M|}{n+N} \to \infty$ gilt (für $n \to \infty$). Die Patternanzahl p kann ebenfalls von n abhängen. (Die Konstruktion muss nicht für jedes n gelten, aber für unbeschränkt große.)

Lösung: Wir verwenden das einelementige Alphabet $\Sigma = \{a\}$. Sei $T := a^{\ell^2}$ und $P := \{P_k \mid k = 1, \ldots, p\}$ wobei $P_k := a^k$ und $p := \ell$. Dann ist

$$|M| = |\{(i,j) \mid i \geq j\}| \geq (\ell^2 - l)\ell = \ell^3 - \ell^2\,,$$

denn spätestens ab der ℓ-ten Position des Textes passen alle Patterns. Die Textlänge ist $n = \ell^2$. Die Gesamtlänge aller Patterns ist

$$N = \sum_{k=1}^{\ell} |P_k| = \sum_{k=1}^{\ell} k = \frac{(\ell+1)\ell}{2} = \frac{\ell^2}{2} + \frac{\ell}{2}\,.$$

Für große ℓ ist $\ell^3 - \ell^2 \geq \ell^3/2$ und $\ell^2/2 + \ell/2 \leq \ell^2$. Also ist

$$\frac{|M|}{n+N} \geq \frac{\ell^3 - \ell^2}{\ell^2/2 + \ell/2} \geq \frac{\ell^3/2}{\ell^2} = \frac{\ell}{2} \to \infty\,.$$

---

3. Aufgabe 8 Punkte [BLAST]

Gegeben sei ein Wort x der Länge 12 über dem Alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Jemand möchte eine Variante von BLAST ausprobieren, bei der als seeds alle Worte der Länge 12 verwendet werden, die sich in höchstens 3 Positionen von x unterscheiden. – Bestimmen Sie die Größe der Menge

$$S := \left\{ \; y \in \Sigma^{12} \;\middle|\; \left|\{i \mid y[i] \neq x[i]\}\right| \leq 3 \; \right\},$$

also die Anzahl der seeds, die für x betrachtet werden.

Lösung: Für k Fehler gibt es $\binom{12}{k}$ Möglichkeiten, ihre Positionen zu verteilen. An jeder Fehlerposition gibt es 3 falsche Symbole, die dort stehen können. Bei k Fehlerpositionen sind es 3^k. Wir rechnen also:

$$1 + 12 \cdot 3 + \binom{12}{2}3^2 + \binom{12}{3}3^3 = 1 + 36 + 594 + 5940 = 6571\,.$$

---

4. Aufgabe 3+3+3 Punkte [Markoff-Kette]

Wir betrachten zwei Markoffketten $M = (Q, A)$ und $N := (Q, B)$ mit den Zuständen $Q = \{0, X, Y\}$ und den Transitionswahrscheinlichkeiten

$A :=$

	0	X	Y
0	0.1	0.3	0.6
X	0.1	0.8	0.1
Y	0.5	0.3	0.2

bzw. $B :=$

	0	X	Y
0	0.3	0.1	0.6
X	0.7	0.0	0.3
Y	0.7	0.2	0.1

Hierbei ist 0 ist der Start- und Endzustand.

(a) Sei $s := XYYX$. Welche der beiden Markoffketten generiert s mit höherer Wahrscheinlichkeit? – Berechnen Sie $\Pr_A(s)$ und $\Pr_B(s)$.

Lösung:
$\Pr_A(s) = a_{0X}a_{XY}a_{YY}a_{YX}a_{Y0} = 0.3 * 0.1 * 0.2 * 0.3 * 0.1 = 0.00018$
$\Pr_B(s) = b_{0X}b_{XY}b_{YY}b_{YX}b_{Y0} = 0.1 * 0.3 * 0.1 * 0.2 * 0.7 = 0.00042$

(b) Was ist die Wahrscheinlichkeit, dass die Markoffkette M als zweites Zeichen ein Y generiert (und insbesondere überhaupt ein zweites Zeichen generiert)?

Lösung:
$\Pr(XY*) + \Pr(YY*) = a_{0X}a_{XY} + a_{0Y}a_{YY} = 0.3 * 0.1 + 0.6 * 0.2 = 0.15$

(c) Jemand erfährt nachträglich, dass die generierte Sequenz genau zwei Zeichen lang war. Mit welcher Wahrscheinlichkeit kann er unter dieser Bedingung annehmen, dass die Markoffkette M als zweites Zeichen ein Y generiert hatte?

Lösung:
$\Pr(XY) + \Pr(YY) = 0.3 * 0.1 * 0.5 + 0.6 * 0.2 * 0.5 = 0.075$.
$\Pr(XX) + \Pr(YX) = 0.3 * 0.8 * 0.1 + 0.6 * 0.3 * 0.1 = 0.042$. Die gesuchte bedingte Wahrscheinlichkeit ist also

$$\frac{\Pr(XY) + \Pr(YY)}{\Pr(XY) + \Pr(YY) + \Pr(XX) + \Pr(YX)} = 75/(75 + 42) = 75/117\,.$$

[Anm.: Also deutlich mehr als bei (c)!]