Introducció

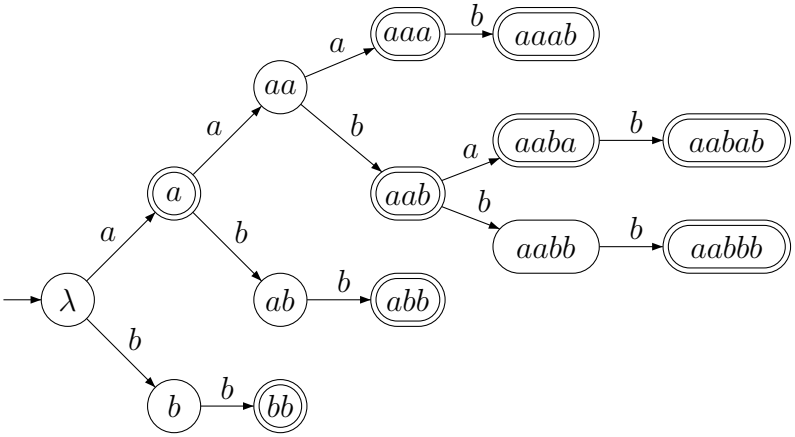
El problema de detectar en quines posicions apareix una o un conjunt de distintes cadenes (usualment denominades patrons) en una cadena més llarga x (text) es coneix com *String Matching* o *Pattern Matching*. Aquest problema és de gran interés algorísmic i té utilitat pràctica en camps com, per exemple, la Biologia Molecular o la Genètica, ja que permet el processament ràpid de seqüències biològiques.

Una primera aproximació (*naive*) al problema comporta buscar cada patró en la seqüència cosa que suposa un cost de $\mathcal{O}(n \cdot |p| \cdot |x|)$, on n és el nombre de patrons a localitzar, $|p|$ és la longitud del patró més llarg i $|x|$ és la longitud del text.

Donat un conjunt M de cadenes sobre determinat alfabet Σ, l'*arbre acceptor de prefixos* per a M ($AAP(M)$) és un autòmat determinista que accepta exclusivament M. Per exemple, donat el conjunto:

$$M = \{a, bb, aaa, aab, abb, aaab, aaba, aabab, aabbb\}$$

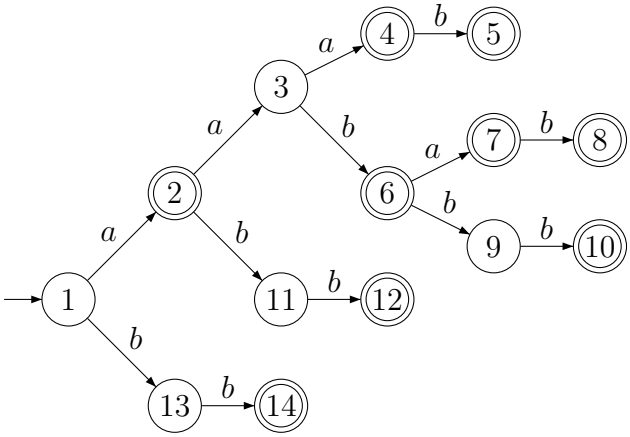
l'$AAP(M)$ seria el següent:



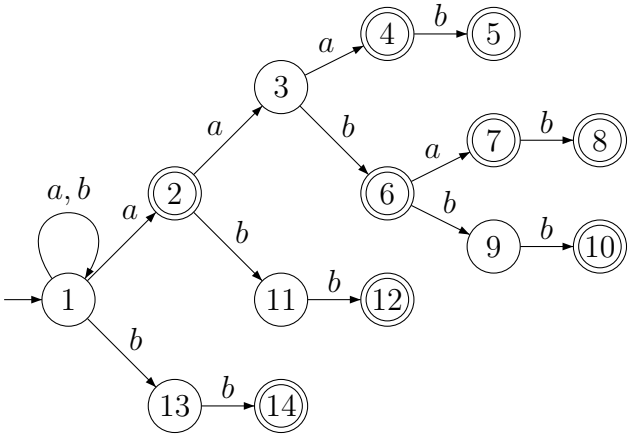
Formalment, l'$AAP(M)$ es defineix com l'autòmat $A = (Q, \Sigma, \delta, q_0, F)$ on:

- $Q = \{x \in \Sigma \ : \ x \in Pref(M)\}$
- $q_0 = \lambda$
- $F = M$
- $\delta(x, a) = xa$ si $xa \in Q$, estant indefinida en cas contrari.

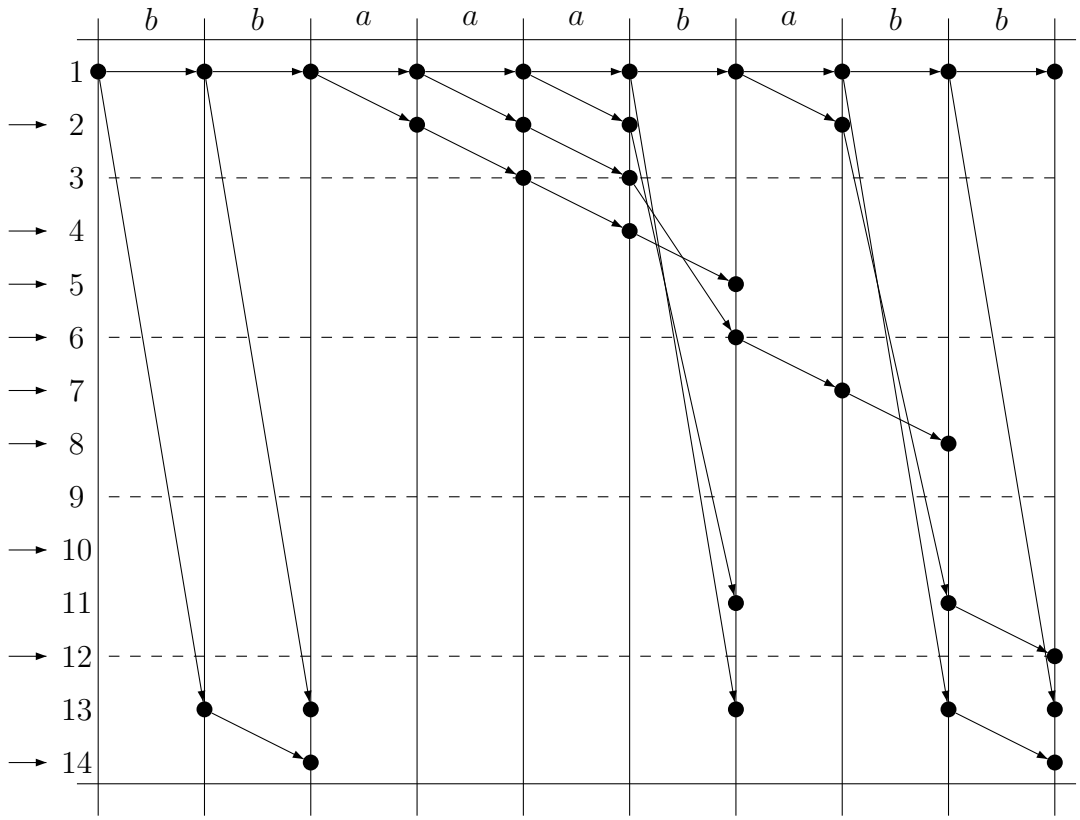
Si enumerem els estats queda:



Aquest autòmat pot modificar-se fàcilment per a obtenir un AFN que accepte $\Sigma^* M$. Note's que per a açò prou afegir un bucle sobre l'estat inicial amb tots els símbols de l'alfabet. A continuació es mostra l'AFN obtingut a partir de l'exemple anterior.



A partir d'aquest autòmat, és possible detectar totes les posicions on apareix una cadena de M en un text x. Per a açò prou realitzar una anàlisi no determinista modificada lleugerament. Aquesta modificació consisteix a detectar cada vegada que s'aconsegueix un estat final en el conjunt d'estats actius. Per exemple, considerant el text $x = bbaaababb$, l'anàlisi no determinista pot representar-se com segueix:



En aquest diagrama es pot vore que després d'analitzar el segon símbol s'arriba a l'estat 14, que en ser final indica que s'ha detectat un patró del conjunto M (el patró bb). De la mateixa manera, per exemple: després d'analitzar bba i $bbaa$ s'arriba a l'estat 2 la qual cosa indica que s'ha detectat el patró a; quan s'analitza $bbaaa$ s'arriba als estats finals 2 i 4 la qual cosa indica que s'han detectat els patrons a i aaa, y així successivament.

Exercicis

Exercici 1

Es demana implementar un mòdul Mathematica que, prenent un conjunt de cadenes M com entrada, torne el conjunt de prefixos de M.

Exercici 2

Es demana implementar un mòdul Mathematica que, prenent un conjunt de cadenes M com entrada, torne el conjunt de sufixos de M.

Exercici 3

Es demana implementar un mòdul Mathematica que, prenent un conjunt de cadenes M com entrada, torne l'arbre acceptor de prefixos del conjunt.

Ajuda:

- La utilització dels distints prefixes del conjunt M com identificadors dels estats de l'autòmat facilita la implementació del mòdul.

Exercici 4

Es demana implementar un mòdul Mathemática que, prenent un conjunt de cadenes M com entrada, torne un AFN que accepte el llenguatge $\Sigma^* M$.

Exercici 5

Es demana implementar un mòdul Mathematica que, donats un AFN A i una cadena x, determine si $x \in L(A)$.

Exercici 6

Es demana implementar un mòdul Mathematica que, donats un conjunt de patrons M i un text x, torne el conjunt de posicions de x en les quals apareix un element de M.

Bibliografia

Maxime Crochemore, Christophe Hancart and Thierry Lecroq ALGORITHMS ON STRINGS. *Cambridge University Press*, 2007.