1. Basic Notions and Definitions

Slides for the Lecture on
Foundations of Sequence Analysis
Winter 2022/2023
Stefan Kurtz

October 16, 2022

Basic Notions and Definitions

- Let S be a set, i.e. a collection of items of the same kind, without duplicates.
- We can have sets of numbers, sets of characters, sets of strings, sets of pairs of numbers, etc.
- $|S|$ denotes the number of elements in S.
- $\mathbb{N}$ denotes the set of positive integers including 0.
- $\mathbb{R}^+$ denotes the set of positive real numbers including 0.
- The symbols $h, i, j, k, l, \ell, m, n, q, r$ refer to integers if not stated otherwise.
- $|i|$ is the absolute value of i and $i \cdot j$ denotes the product of i and j.
- Sometimes we omit the multiplication operator and e.g. write something like $2n + 3m$ instead of $2 \cdot n + 3 \cdot m$.

Notation for sequences

- Let $\mathcal{A}$ be a finite set, the *alphabet*.
- The elements of $\mathcal{A}$ are *characters*.
- Strings are written by juxtaposition of characters, i.e. we write characters composing a sequence without separators like commas or spaces.
- We introduce a special notation for the sequence containing no characters: ε denotes the *empty sequence*.
- Concatenation of sequences u and v means to write u before v without any separator.
- So uv is the concatenation of u and v.
- If the empty sequence is concatenated with a sequence, then we omit the empty sequence.
- That is, $\varepsilon w = w\varepsilon = w$ for all sequences w.

Notation for sequences

The set $\mathcal{A}^*$ of *sequences over* $\mathcal{A}$ is defined by

$$\mathcal{A}^* = \bigcup_{i \geq 0} \mathcal{A}^i = \mathcal{A}^0 \cup \mathcal{A}^1 \cup \mathcal{A}^2 \cup \ldots$$

where $\bigcup$ is the union-operator and

$$\mathcal{A}^i = \begin{cases} \{\varepsilon\} & \text{if } i = 0 \\ \{aw \mid a \in \mathcal{A}, w \in \mathcal{A}^{i-1}\} & \text{if } i > 0 \end{cases}$$

That is, $\mathcal{A}^i$ is the set of sequences of length i. This set is defined recursively:

- The first case (for $i = 0$) states that the only sequence of length zero is the empty sequence.
- The second case (for $i > 0$) states that the sequences of length i are composed by some character from $\mathcal{A}$ prepended to some sequence of length $i - 1$ from $\mathcal{A}^{i-1}$.

Notation for sequences

You all should have seen the notation for sets, in which the elements of a set are specified and enclosed in curly brackets, like in

$$\{1, 3, 5, 7, 9\}$$

specifying the set of all odd numbers smaller than 10.
In many cases, one wants to specify a set by properties satisfied by all elements in the set, for example:

$$\{i \mid i \in \mathbb{N}, i < 10, i \text{ is odd}\} \text{ or shorter } \{i \in \mathbb{N} \mid i < 10, i \text{ is odd}\}$$

- This specifies the same set as above.
- The elements are specified by a variable i and properties referring to i given after the symbol $|$.
- Instead of this symbol, other authors often use a colon $:$ instead.
- The properties to be satisfied are separated by commas.

Formal definitions, like the one for $\mathcal{A}^i$, are very helpful to derive and prove properties of the defined items, as shown in the following lemma.

Lemma 1

For all $i \geq 0$, $|\mathcal{A}^i| = |\mathcal{A}|^i$.

Proof.

For $i = 0$ we have $|\mathcal{A}^i| = |\mathcal{A}^0| = |\{\varepsilon\}| = 1 = |\mathcal{A}|^0$, that is, the lemma holds for $i = 0$. Suppose that $|\mathcal{A}^{i-1}| = |\mathcal{A}|^{i-1}$ holds for a fixed but arbitrary $i > 0$. Then, we can conclude

$$\begin{aligned}
|\mathcal{A}^i| &= |\mathcal{A}| \cdot |\mathcal{A}^{i-1}| && (\textit{by Definition of } \mathcal{A}^i = \{aw \mid a \in \mathcal{A}, w \in \mathcal{A}^{i-1}\}) \\
&= |\mathcal{A}| \cdot |\mathcal{A}|^{i-1} && (\textit{by assumption}) \\
&= |\mathcal{A}|^i && (\textit{by evaluation}).
\end{aligned}$$

By the principle of induction this shows that the lemma holds for any $i > 0$ which completes the proof. $\square$

- $\mathcal{A}^+$ denotes $\mathcal{A}^* \setminus \{\varepsilon\}$, where $\setminus$ means subtraction of sets.
- That is, $\mathcal{A}^+$ is the set of all non-empty sequences over $\mathcal{A}$.
- The symbols a, b, c, d refer to characters and $p, s, t, u, v, w, x, y, z$ to sequences, unless stated otherwise.

Example 1

1. ASCII: 8-bit characters, encoding as defined by the ASCII standard
2. $\{\mathtt{A}, \ldots, \mathtt{Z}, \mathtt{a}, \ldots, \mathtt{z}, \mathtt{0}, \ldots, \mathtt{9}, \}$: alphanumeric subset of the ASCII-set
3. $\{\mathtt{A}, \ldots, \mathtt{Z}\} \setminus \{\mathtt{B}, \mathtt{J}, \mathtt{O}, \mathtt{U}, \mathtt{X}, \mathtt{Z}\}$: letter code for 20 amino acids
4. $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$: DNA alphabet (Adenine, Cytosine, Guanine, Thymine)
5. $\{\mathtt{R}, \mathtt{Y}\}$: purine $(\mathtt{a}, \mathtt{g})$/pyrimidine $(\mathtt{c}, \mathtt{t})$-alphabet
6. $\{\mathtt{I}, \mathtt{O}\}$: hydrophilic/hydrophobic nucleotides/amino acids
7. $\{+, -\}$: positive/negative electrical charge $\square$

Notation for sequences

- While the context usually allows to distinguish variables for characters and strings from concrete characters and strings, we use different fonts for these.
- In particular, we use typewriter fonts for concrete characters and sequences, like `a`, `c`, `ac`, `tataa`.
- Variables denoting a character or a string, like a and w below are written in italic fonts.
- The *length* of a sequence s, denoted by $|s|$, is the number of characters in s.
- We make no distinction between a character and a sequence of length one.

Example 2

Let $\mathcal{A} = \{\mathtt{b}, \mathtt{c}\}$. Then ε is a sequence of length 0, and $\mathtt{bccb}$ is a sequence of length 4. $\mathtt{b}$ and $\mathtt{c}$ are characters in $\mathcal{A}$ but also sequences of length 1. We can determine the set $\mathcal{A}^2$ by applying the above definitions as follows:

$$
\begin{aligned}
\mathcal{A}^1 &= \{aw \mid a \in \mathcal{A}, w \in \mathcal{A}^0\} \\
&= \{aw \mid a \in \{\mathtt{b}, \mathtt{c}\}, w \in \{\varepsilon\}\} \\
&= \{a\varepsilon \mid a \in \{\mathtt{b}, \mathtt{c}\}\} \\
&= \{\mathtt{b}\varepsilon, \mathtt{c}\varepsilon\} \\
&= \{\mathtt{b}, \mathtt{c}\} \\
\mathcal{A}^2 &= \{aw \mid a \in \mathcal{A}, w \in \mathcal{A}^1\} \\
&= \{aw \mid a \in \{\mathtt{b}, \mathtt{c}\}, w \in \{\mathtt{b}, \mathtt{c}\}\} \\
&= \{\mathtt{b}w \mid w \in \{\mathtt{b}, \mathtt{c}\}\} \cup \{\mathtt{c}w \mid w \in \{\mathtt{b}, \mathtt{c}\}\} \\
&= \{\mathtt{bb}, \mathtt{bc}\} \cup \{\mathtt{cb}, \mathtt{cc}\} \\
&= \{\mathtt{bb}, \mathtt{bc}, \mathtt{cb}, \mathtt{cc}\}
\end{aligned}
$$

If $s = uvw$ for some (possibly empty) sequences u, v and w, then

- u is a *prefix* of s,
- v is a *substring* of s, and
- w is a *suffix* of s.

Example 3

Let $s = \mathtt{acca}$. The suffixes of s are $\mathtt{acca}$, $\mathtt{cca}$, $\mathtt{ca}$, $\mathtt{a}$, and ε. The prefixes of s are ε, $\mathtt{a}$, $\mathtt{ac}$, $\mathtt{acc}$ and $\mathtt{acca}$. The only substrings of s which are not prefixes and not suffixes are $\mathtt{c}$ and $\mathtt{cc}$.

Notation for sequences

- $s[i]$ is the ith character of s.
- That is, if $|s| = n$, then $s = s[1]s[2] \dots s[n]$ where $s[i] \in \mathcal{A}$.
- $s[n]s[n-1] \dots s[1]$, denoted by s^{-1}, is the *reverse* of $s = s[1]s[2] \dots s[n]$.
- If $i \leq j$, then $s[i \dots j]$ is the substring of s beginning with the ith character and ending with the jth character.
- If $i > j$, then $s[i \dots j]$ is the empty sequence.
- A sequence w begins at position i and ends at position j in s if $s[i \dots j] = w$.