TD 1

Stéphane Vialette

1 Nussinov

2 RNA design

Definition 1. *An alphabet is a finite set Σ of symbols, together with a symmetric binary relation $M \subseteq \Sigma \times \Sigma$ such that:*

- *every symbol $a \in \Sigma$ is related to at most one symbol $b \in \Sigma$;*
- *if a is related to b, then $a \neq b$.*

As an immediate consequence of the definition, an alphabet Σ contains a certain number m of couples of related symbols (we will also speak of *matching symbols* or *paired symbols*) and a certain number ℓ of symbols without a mate. Clearly, $2m + \ell = |\Sigma|$ and we say that Σ is an alphabet of type (m, ℓ).

We will always assume that all the alphabets have $m \geq 1$ pairs of matching symbols (otherwise the problems we are going to consider are not interesting).

Definition 2. *The standard alphabet is the alphabet $\{A, C, G, U\}$, where A is paired with U and C is paired with G. So the standard alphabet is of type $(2, 0)$.*

Definition 3. *Given an alphabet Σ, an alphabet automorphism of Σ is a bijective function $\phi : \Sigma \to \Sigma$ such that $M(x, y) \Leftrightarrow M(\phi(x), \pi(y))$. In other words, ϕ must map pairs of matching symbols to pairs of matching symbols. For instance, it is easy to check that the standard alphabet has 8 automorphisms.*

Definition 4. *A sequence over an alphabet Σ is a string of symbols of Σ. For instance, $AUAGGC$ is a sequence over the standard alphabet.*

Definition 5. *A structure is a valid expression made of dots and parenthesis. For instance, $().(.)$ and $.(()(.()))$ are structures, whereas $.).()($ is not. The length of a structure is the number of symbols it consists of.*

Definition 6. *The rank of a structure S is the number of pairs of parenthesis in S. For instance, the rank of$().(.)$ is 2.*

Definition 7. *A sequence X and a structure S are compatible if they are of the same length and, for every pair of matching parenthesis in S, the corresponding symbols in X match. For instance, the sequence $AUAGGC$ and the structure $().(.)$ are compatible.*

Given a sequence X, one is interested in the structures compatible with X with the maximum rank. Such structures are called *folds* of X.

Definition 8. *A sequence X is called a* realization *of a structure S if S is the unique fold of X.*

Definition 9. *Let Σ be an alphabet. A structure S is said to be* Σ-designable *(or* Σ-allowed*) if it has a realization over the alphabet Σ. A structure S is* Σ-forbedden *if it is not* Σ-designable *(*i.e. *if there is no sequence over the alphabet Σ that has S as the unique fold). If the used alphabet is clear by the context, we simply speak of designable (allowed) or forbidden structures, without specifying the alphabet.*

We adopt the standard alphabet $\Sigma = \{A, C, G, U\}$.

- **How to determe if a given sequence X has a unique fold (and, if so, finding it)?**
- **How to determine if a given structure S is designable?**

3 Longest common subsequence

Definition 10. *The longest common subsequence (LCS) problem is the problem of finding the longest subsequence common to all sequences in a set of sequences.*

It differs from problems of finding common substrings: unlike substrings, subsequences are not required to occupy consecutive positions within the original sequences.

- **How to compute the LCS of 2 sequences?**
- **How to compute the LCS of k sequences?**

Notice that the LCS is not necessarily unique; for example the LCS of AGC and ACG is both AG and AC. Indeed, the LCS problem is often defined to be finding all common subsequences of a maximum length. This problem inherently has higher complexity, as the number of such subsequences is exponential in the worst case, even for only two input strings.

- **Illustrate the above remark.**