Algorithms for the description of molecular sequences

Vis, J.K.

Chapter 5

Extraction of HGVS Variant Description for Short Tandem Repeat Structures

Characterization of complex variants such as short tandem repeats (STRs) is an important tool in forensics. Historically, these variants are difficult to characterize. Here, we propose a method that can automatically construct a description following the philosophy of the standard nomenclature of the Human Genome Variation Society (HGVS). Within these descriptions there is also a way of describing variation in the flanking regions relative to the repeat structure.

We accurately detect repeat units in a (reference) sequence and we can use these repeat units to generate a meaningful reference-based description. We propose an alternative repeat variant that can be added to the existing HGVS nomenclature and finally, we provide an implementation of our proposed methods as part of the Description Extractor package for Mutalyzer: `https://pypi.python.org/pypi/description-extractor`. The Python source code is available at: `https://github.com/mutalyzer/description-extractor`.

5.1 Introduction

A considerable part of the genome consists of repeated regions that occur in many forms. In this chapter we focus on a particular from of repetition in genomic sequences called *Short Tandem Repeats* (STRs) also known as microsatellites or simple sequence repeats. STRs contain small motifs or *units* ranging from 2–5 base pairs which are repeated in *tandem*, i.e., immediately adjacent, typically 5–50 times. The combination of the lengths and different loci of STRs is a powerful tool in population genetics and forensics. The analysis of repeated sections of the genome have always been problematic not only on the sequencing level, but also in automated analysis [Weischenfeldt et al., 2013]. In this chapter we present a method for the automatic construction of descriptions of repeated structures, especially tailored for forensic applications. These descriptions are based on the recommendations of the Human Genome Variation Society (HGVS) [den Dunnen et al., 2000, Taschner and den Dunnen, 2011]. While the HGVS nomenclature currently provides a way to describe one repeat unit in tandem, forensic applications demand a more rigorous construct to describe the repeated structures, possibly containing multiple repeat units, as a whole. In any case, the detection of repeated sequences including the repeat unit is a non-trivial task.

The STR allele descriptions need to maintain compatibility with existing databases including the definition of the repeat units, annotation on the forward or reverse strand, etc. In Section 5.3 of [Gettings et al., 2015] an attempt is made to create a uniform representation of commonly used STR alleles in terms of strands and repeat units. They also envision three possible methods for describing STR alleles:

1. complete sequence strings;
2. a bracketed description;
3. unique identifiers.

A description method based on the use of complete sequence strings entails the direct use of Next Generation Sequencing output. This output is deemed to be

large and complex and therefore not necessary tailored to the intended purpose here. In particular, storage in databases will become more costly and matching techniques is expected to require more computational power. On the other hand, more relevant data can be stored; small variants in the "flanking" region are also present in this approach. The bracketed description is often informally used in reports and is especially useful in describing the actual repeat units and the number of occurrences, e.g., TCTA[9], where the unit TCTA is repeated 9 times. Having this kind of descriptions means that comparison on (allele) length is relatively straightforward. However, combinations with changes in the flanking regions are more difficult to express as no real guidelines exist at the moment. Finally, one could consider unique identifiers for each STR allele. In [Gettings et al., 2015] it is suggested that a somewhat meaningful identifier can be assigned. While based on the number of repeats and its flanking variations these identifiers give no proper description. Unique identifiers will of course be superior when performing database searches. However, this approach requires more administration as these identifiers need to be globally unique. Furthermore, humans (in a laboratory setting) can have difficulties in dealing with these kind of descriptions. In particular, partial searches are impossible.

Other attempts of automatically characterizing STR alleles have been made. Most importantly in [Anvar et al., 2014] where a sequence can be matched against a pre-defined database of STR allele descriptions given as regular expressions. Some small variants in the flanking regions are tolerated, but no part of the description. Amongst other reasons, this method does not result in a full description of the observed sequence. In [Hoogenboom et al., 2016] the results from [Anvar et al., 2014] are used and improved upon by focusing on the reduction of PCR stutter noise. An accurate estimate of repeat lengths can be given and there is some support for the characterization of unknown STR alleles.

In this chapter we propose a method for the automatic generation of reference based descriptions for STR alleles. In essence this is closest to the aforementioned bracketed approach. Apart from giving descriptions for the repeat structure, we also incorporate the classical extraction method from Chapter 3 for the description of the flanking, and possibly interspersing, regions

in such a way that the variation within these flanking regions is described relative to the repeat structure, cf. the descriptions of coding regions in the HGVS nomenclature. This approach combines the strengths of both methods while eliminating their respective drawbacks.

The remainder of this chapter is structured as follows. In Section 5.2 we present a method for a reference sequence based description for STR alleles. This method is in principle database-free, but can also be used in combination with a DNA data bank yielding descriptions that are closer to the annotation used in the literature. In Section 5.3 we introduce the data and experiments, followed by a discussion of the results of these experiments in Section 5.4 and the conclusions in Section 5.5.

5.2 Methods

First we focus on the detection (or extraction) of repeat units from a string. Later we combine these results and an adaptation of the HGVS extraction algorithm in [Vis et al., 2015] to generate a reference-based HGVS-like description. Finally, we address the problem of generating descriptions relative to the repeat structure following a similar approach as is used in the HGVS nomenclature when describing variants in the $5'$ and $3'$ UTR relative to the coding region of a transcript.

5.2.1 Finding repeat units

In order to have a database-free method we must be able to find candidate repeat units in a string. For the concrete application at hand, STRs, we can take advantage of the properties of STRs. The *tandem*, i.e., immediately adjacent, property of these repeats is of primary concern. Often suffix trees are used for finding substrings with certain properties. However, as we deal with non-overlapping directly adjacent repeats, a more straightforward method can be used; a variation on classical run-length encoding [Golomb, 1966]. In the classical run-length encoding algorithm, the length of the repeat unit is fixed at 1. The algorithm can easily be extended to arbitrary fixed-length repeat

units. For this particular application we do not know the length of the repeats units for a certain string a-priori. As the repeat units within STRs are of limited length, i.e., 2–5, one can easily try to find repeat units of all of these sizes. Special care has to be taken when dealing with self-similar repeat units, e.g., `TATA`, because an other, smaller unit exists, `TA` describing the same repeated sequence. In these cases we are interested in the smallest unit describing the repeated sequence. Algorithm 1 is an implementation of such a variation on the run-length encoding algorithm.

In Algorithm 1, we iterate over the length of the string (line 3). Every candidate length of the repeat unit (k) is tried for the given position in the string (line 5). Candidates are limited by the optional parameters min_length and max_length. The number of repeats for this particular unit is tracked (line 11). If for the given repeat unit length we obtain a larger covered region in the string, this is considered the "best" candidate repeat unit at this position in the string (line 13). Ultimately, the selected repeat unit for this position is added to the repeat unit set (line 17), and the next possible position for a repeat unit is considered. In the end we can end up with a repeated region at the end of the string, so we need to add that as necessary (line 21). The algorithm returns a partitioning of the string in repeat units.

On the resulting partitioning it is trivial to filter for a specific minimum or maximum number of occurrences, as well as constructing a set of unique repeat units with characterized lengths and counts. Such a set can be used in the second stage of our method.

An implementation in Python of our proposed method is provided at: `https://github.com/mutalyzer/ssr-extractor`.

5.2.2 Reference-based description of the repeat structure

The purpose of this method is to yield a bracketed type of description, similar to the existing HGVS nomenclature for repeated sequences, given a set of repeat units. This set could be generated by the application of Algorithm 1, or alternatively, a set from a database or literature can be used. As further input the method expects a reference string as well as the observed string.

Algorithm 1 SHORTTANDEMREPEAT

input: $string, [min\_length = 2, max\_length = 5]$ (optional unit length)
output: $repeats$ (partitioning in repeat units)

while $0 \leq i < |string|$ **do**
  $max\_count = 0, max\_k = 1$
  for $k \in [min\_length, \ldots, max\_length]$ **do**
    $count = 0$
    for $i + k \leq j \leq |string| - k + 1$ **step** k **do**
      if $string_{i \ldots i+k} \neq string_{j \ldots j+k}$ **then**
        break
      end if
      $count \leftarrow count + 1$
    end for
    if $count > 0$ **and** $count \geq max\_count$ **then**
      $max\_count \leftarrow count, max\_k = k$
    end if
  end for
  $repeats$.add($\{i, max\_k, max\_count\}$)
  $i \leftarrow i + max\_k(max\_count + 1)$
end while
if $max\_count > 0$ **then**
  $repeats$.add($\{i, max\_k, max\_count\}$)
end if

return $repeats$

As we try to follow the current nomenclature standard closely, the resulting textual output of these descriptions will be resembling the HGVS nomenclature. However, we introduce a new type of variant, for describing repeated sequences, with slightly stricter defined semantics. As none of this is part of the official HGVS nomenclature we will, for now, use a different notation to avoid confusion. The newly introduced type of variant is a form of the deletion/insertion operator in HGVS. For its notation we list the repeat unit

(in bases) once, followed by the number of its consecutive occurrence in the observed sequence in parentheses, e.g., ATCT(4). This denotes that at a certain position in the reference sequence all tandem occurrences of this particular repeat unit are deleted and the specified number of occurrences is inserted in the observed sequence, for the given example, 4.
Consider the example:

$$\begin{aligned} R &= \texttt{ATCTATCTATCTGGATCT} \\ S &= \texttt{ATCTATCTATCTGGATCTATCT}, \end{aligned}$$

where R is the reference sequence and S the observed sequence and we choose {ATCT} as repeat unit set. The aim is to come to a description that shows that the repeat unit is present 4 times in the observed sequence. The repeat unit is present only 3 times in the reference sequence. The remainder of both sequences is a complete match. We propose to describe this STR allele as: [ATCT(4);13_14;ATCT(2)], giving an unambiguous description from the reference sequence to the observed sequence. Note that the identical (matching) part of the reference sequence is described using our transposition notation defined in Section 3.1.1. Indeed, the whole description should be interpreted as a deletion of the complete reference string with the observed string inserted expressed as transpositions of the reference string and repeat units. Small variants, e.g., single nucleotide substitutions can be described in the classical way. One could remark that the repeat unit itself can also be expressed as a transposition, however, in order to be closer to both the currently used notation in databases as well as the HGVS nomenclature, we prefer the notation in bases for the repeat unit. Furthermore, it might be required to give a description of an STR allele where the repeat unit is not present in the reference sequence, in which case there is no valid transposition.

The extraction of the description follows the method described in Section 3.2 with a slight alteration; we introduce a so-called masking character, i.e., a character not part of any of the molecular alphabets with the added property that this masking character does not match any character (including itself). We denote this character by $. The first step of our method masks all the occurrences of the repeats from the repeat unit set with the masking characters

in both the reference and the observed sequence. In this way we can abstract from the actual sequence without losing the positioning information which is needed to construct the description. Following the aforementioned example R and S are transformed into `GG$$$$$$$$$$$$G` and `GG$$$$$$$$$$$$$$$$G` respectively. These masked strings are used as input for the variant description extraction (VDE) method presented in Chapter 3. In this particular case, the variant extraction does not yield any variants as both prefixes and suffixes are identical, but as mentioned before, small variants can be described in the classical way.

Regular substring matching is used to fill in the masked regions with repeat units from the repeat unit set, e.g., at position 3 of the observed sequence we find 4 occurrences of the unit `ATCT`. The descriptions resulting from the VDE and the masked regions are then merged, where identical regions are transformed into transpositions. When dealing with multiple variants, the VDE will sometimes combine multiple atomic variants into larger deletions/insertions based on the description length of the HGVS nomenclature. In particular instances it may happen that variants are combined while spanning a repeat unit as well. In these cases we have to split the combined variant and interleave it with the repeat variant in order to perform the merging step.

5.2.3 Relative description of the flanking regions

The final part of our method deals with the automatic generation of relative descriptions, i.e., positioning relative to some part of the reference sequence, for describing variation in the flanking regions. These relative descriptions are similar to the coding region oriented descriptions used in HGVS for the description of $3'$ and $5'$ UTR variants, e.g., `c.-56C>T` and `c.*32G>A`. As the variation in the flanking regions can be characteristic for certain STR alleles (in combination with the repeat structure) this is an important element of the STR allele description. A naive approach could be generating a description that is a composite deletion/insertion, containing transpositions and insertions, of the whole reference sequence, including the flanking regions. However, the resulting descriptions can become quite complex and difficult to interpret,

especially when variable length flanking regions are used. Furthermore, these descriptions distract the attention from the actual repeat structure, making it more difficult to identify. In our method we propose to solve this problem by using a relative description, i.e., we describe the variation in the flanking regions relative to a predefined region. In practice, we use the repeat structure, but in general this can be any substring of the reference sequence. A uniform way of defining the repeat structure would be from the first occurrence of any of the repeat units in the repeat unit set to the last occurrence of any of the repeat units. This results, when applying the steps in Section 5.2.2, in descriptions that look like `[-15A>T;ATCT(4);*4_5insAT]`, cf. the HGVS nomenclature by using the minus operator to describe variation before the repeat structure and the star operator to describe variation after the repeat structure.

The Mutalyzer tool suite [Wildeman et al., 2008] module that deals with the generation of relative descriptions is called the Crossmapper. This module performs the non-trivial conversion between the different positioning schemes in HGVS. In particular it is used to convert variant descriptions on different transcripts, i.e., from genome to transcript and vice versa. We can make use of this module for our purpose by introducing an artificial coding region that represents the repeat structure. After generating a naive (genomic) description of the complete reference sequence, we can apply the Crossmapper, given this artificial transcript, providing us with a view relative to the repeat structure, e.g., `TPOX(122_142):[-50G>T;TGAA(6)]`, where the selected transcript is made explicit in the notation of the reference sequence. In this example, the repeat structure is selected from position 122 to 142. In the flanking region there is a single nucleotide substitution 50 bases before the start of the repeat structure. The flanking region after the repeat structure contains no variants in this case.

5.3 Experiments

In this section we describe the experiments designed to illustrate the performance of our proposed method for the generation of STR allele descriptions. We focus on forensic use cases. In Table 5.1 we introduce a set of 24 STR loci that are commonly used with their repeat structure annotation as found

in [Gettings et al., 2015]. For each locus only one STR allele is annotated, which can serve as the reference sequence for that locus. We have a dataset containing 1,631 STR alleles from the loci in Table 5.1, including the reference sequence. In Table 5.2 we give a characterization of the dataset.

Table 5.1: A set of 24 frequently used STR loci with their respective repeat structure. All repeat structure characterizations are normalized following [Gettings et al., 2015], but do not include any flanking regions.

STR locus	Repeat structure
Amel[†]	none
CSF1P0	AGAT(13)
D10S1248	GGAA(13)
D12S391	AGAT(11) AGAC(7) AGAT
D13S317	TATC(11) AATC(2)
D16S539	GATA(11)
D18S51	AGAA(18)
D19S433	AAGG AAAG AAGG TAGG AAGG(12)
D1S1656	TAGA(16) TAGG TG(5)
D21S11	TCTA(4) TCTG(6) TCTA(3) TCA TCAT(2) TCCATA TCTA(11)
D22S1045	ATT(14) ACT ATT(2)
D2S1338	TGCC(7) TTCC(13) GTCC TTCC(2)
D2S441	TCTA(12)
D3S1358	TCTA TCTG TCTA(14)
D5S818	AGAT(11)
D7S820	GATA(13)
D8S1179	TCTA TCTG TCTA(11)
DYS391	TCTA(11)
FGA	TTTC(3) TTTT TTCT CTTT(14) CTCC TTCC(2)
PentaD	AAAGA(13)
PentaE	AAAGA(5)
TH01	AATG(7)
TPOX	AATG(8)
vWA	TCTA TCTG(5) TCTA(11) TCCA TCTA

[†] Amelogenin (Amel) is not an STR, but often included as a way to determine sex.

The automatic extraction of the repeat unit set in Table 5.2 is performed by applying Algorithm 1 on each of the STR alleles belonging to a certain STR locus. An aggregated repeat unit set is then created for each locus by only selecting the repeat units that are present in all separate repeat unit sets. In addition, the total count for each repeat unit should be more than 5. This selection is rather arbitrary, but in practice it yields reasonable repeat unit sets. Although a comprehensive method of finding the "best" repeat unit set given a set of STR alleles is outside the scope of this chapter, some improvements are discussed in Section 5.4.

As can be observed from Table 5.2, the majority of the repeat units is of length 4. Two repeat units are of length 5, one of length 2 and one of length 3. The automatically extracted repeat units correspond in most cases to the ones that are described in literature, especially when taking into account the possible rotations of the pattern (i.e., `ATCT` can also be described as `TCTA`, `CTAT` and `TATC`). In multiple cases, i.e, CSF1P0, D13S317, D21S11, D8S1179 and FGA, two rotations of the same repeat unit are automatically extracted (usually `ATCT`), this phenomenon usually arises when the same repeat unit occurs in multiple, separate, repeated regions, where the surrounding bases have an influence on the selection of the particular rotation. In some cases, i.e., D13S317, D16S539 and D7S820, a novel repeat unit is extracted. This unit is present in the STR allele, but it may not occur in a variable number over the STRs, i.e., they occur always a fixed number of times. This is not a property that can be intrinsically detected by Algorithm 1. For one locus, TH01, our dataset clearly contains the opposite strand from the one that is commonly used in literature.

As a second experiment we applied the full description generator method on the full dataset from Table 5.2, once using the automatically extracted repeat units and once using the repeat units taken from [Gettings et al., 2015] (where we used the reverse complement for TH01). As reference sequence we selected the STR allele with the repeat structure from Table 5.1. Note that even when we extract a STR description for the reference sequence itself, our method gives a description in terms of repeat units, where, on the other hand, the VDE would result in the idem description. The repeat structure location

Table 5.2: Characterization of the dataset used for the experiments. For each STR locus the number of variant alleles is given as well as their annotated repeat unit set (from [Gettings et al., 2015]) and the automatic extracted repeat unit set.

STR locus	Count	Repeat unit set (literature)	Repeat unit set (extracted)
Amel	3	none	none
CSF1P0	22	{AGAT}	{TAGA, AGAT}
D10S1248	16	{GGAA}	{GGAA}
D12S391	107	{AGAT, AGAC}	{CAGA, TAGA}
D13S317	81	{TATC}	{AATC, ATCT, TATC}
D16S539	43	{GATA}	{ACAG, GATA}
D18S51	41	{GAAA}	{AGAA}
D19S433	56	{AAGG}	{AGGA}
D1S1656	76	{TAGA, TG}	{GT, TAGA}
D21S11	323	{TCTA, TCTG}	{TCTG, TATC, TCTA}
D22S1045	34	{ATT}	{ATT}
D2S1338	176	{TGCC, TTCC}	{TGCC, TTCC}
D2S441	28	{TCTA}	{TCTA}
D3S1358	48	{TCTA}	{TGTC, TATC}
D5S818	60	{AGAT}	{AGAT}
D7S820	112	{GATA}	{GATA, TTT}
D8S1179	66	{TCTA}	{TCTA, TATC}
DYS391	13	{TCTA}	{TATC}
FGA	69	{TTTC, CTTT, TTCC}	{TCCT, TTTC, TCTT}
PentaD	49	{AAAGA}	{AAGAA}
PentaE	36	{AAAGA}	{AAAGA}
TH01	32	{AATG}	{GTCT, ATCT, ATCC}†
TPOX	26	{AATG}	{TGAA}
vWA	114	{TCTG, TCTA}	{TATC}

† Our sequence data contains the opposite strand from the one that is used in literature.

is automatically designated by the method described in Section 5.2.3. In the general case the start and end positions of the repeat structure could be chosen

based on literature, however, currently there is no consensus on these positions.

When comparing the generated descriptions using both versions of the repeat unit set, we observe that in the majority of cases either the descriptions are identical (when the repeat unit set was identical) or highly similar, especially when considering the length of the description, or the total number of repeat units covered in repeated sequences, e.g.:

```
D19S433(37_75)  :  [AAGG(1);5_6;AAGG(1);11_14;AAGG(6)]
D19S433(38_83)  :  [AGGA(1);5_10;AGGA(7);39_41;AGGA(1)],
```

where the same sequence is described by two different rotations of the same repeat unit: For the top description `AAGG` from literature is used, while `AGGA` is automatically extracted using Algorithm 1 and is used for the bottom description. The absolute length of both descriptions is very similar; the top one is one character shorter. However, the number of repeat units covered by the respective descriptions is in favor of the bottom description, where 9 units are covered as opposed to 8 in the top description. Both descriptions obviously yield the same observed sequence, but this might not be immediately clear. The location of the repeat structure is slightly different for both and so are the descriptions of this structure, not only the repeat unit itself (which is clear), but, more importantly, the interleaving transpositions. When storing these descriptions in databases unambiguous descriptions might be preferred.

5.4 Discussion

In this section we discuss some of the issues of the application of our method for the automatic generation of descriptions for STR alleles. In particular the usefulness of these descriptions in forensics.

5.4.1 Reference sequence

When the goal is to have easily comparable and meaningful descriptions without the need of resorting to tools, there must be a fixed reference sequence for each STR locus. As this kind of information is stored for relatively long times,

these reference sequences have to be invariant over time. General reference sequences deal with the whole genome, whole chromosomes or genes. None of these is particularly useful; the interesting regions are very small compared to chromosomes and usually not situated inside or near genes. We advocate the use of so-called Locus Reference Genomic (LRG) reference files [MacArthur et al., 2014]. LRGs are fixed references that do not change over time, furthermore, they can contain transcript annotation, which can be used for the specification of the location of the repeat structure. This has the additional advantage that the annotation of the repeat structure is explicitly present in the reference sequence. LRGs can be easily requested via the webpage: `http://www.lrg-sequence.org/`.

5.4.2 Repeat unit set per STR locus

After choosing a suitable reference sequence, there should also be a consensus on the repeat unit set that is used for a specific STR locus. Note that Algorithm 1 does not necessarily give the same repeat unit set for each allele of the same STR locus. This can be regarded as a potential problem for having unambiguous descriptions. Unfortunately, it is impossible to explicitly include this information in an LRG reference sequence. This means that either an additional database has to be curated, or an extension of the LRG sequences should be proposed containing the repeat unit set for each STR locus.

As we observed from the experiments in Section 5.3, repeat units can be rotated. The chosen rotation in literature seems to be, to a certain extent, arbitrary, e.g., the "best" fitting rotation for a certain (reference) allele. In the case of the FGA locus, the literature gives two rotations of the same repeat unit as part of the repeat unit set: `TTTC` and `CTTT`. Rotated versions of the same repeat unit occur more frequently when considering multiple STR loci. One could argue that, at least per locus, only one rotation of the same repeat unit can be present. On the other hand, considering that "best" fitting is rather ill-defined, an arbitrary selection could be made. e.g., the lexicographical smallest. This provides as an advantage that repeat structures over different loci can be easily compared.

The automatic selection of the repeat unit set used in Section 5.3 could be improved upon by disallowing rotations of the same repeat unit, as well as, disregarding any of the static repeat units, i.e., repeats that always occur a fixed number of times. Considering these improvements, the automatic extraction of the repeat unit set would result in the same set as is currently used in literature except for its rotations. Given the need for backwards compatibility we expect that no universal decision can be reached on the selection of the rotations. For this reason our approach has only a loose coupling between the automatic extraction of repeat units and the generation of the corresponding descriptions.

5.5 Conclusions

We introduced a method for the automatic generation of reference-based descriptions for STR alleles. Our method consists of two parts: 1) the automatic detection of a repeat unit set given a sequence and 2) the automatic generation of a reference based description given a repeat unit set, a reference sequence and an observed sequence. The automatic detection of repeat units uses a variation of the well-known run-length encoding algorithm and performs well on STR alleles with many small repeats. The results from this method can be used as input for the second part. Alternatively, repeat unit set defined in literature can be used.

The resulting descriptions are easily interpretable and can be made unambiguous for use in databases. We have discussed the application of our proposed method for a forensic use case with special attention regarding the unambiguity. We advocate the creation and use of specialized LRG reference sequence files.

Our proposed repeat variant, including its semantics and notation, should be proposed and added to the HGVS nomenclature.