A little SeqFeatR tutorial

Bettina Budeus and Daniel Hoffmann,
Bioinformatics, University of Duisburg-Essen,
Essen, Germany

February 10, 2015

Contents

This tutorial gives you some of the technical background underlying SeqFeatR that should enable you to understand what SeqFeatR does and also how to use it and how to interpret the output. If you are solely interested in a HowTo, you may find the following two video tutorials helpful:

- For the GUI in the SeqFeatR R-package: https://www.youtube.com/watch?v=-CYidGPE6dw
- For the SeqFeatR web-server: https://www.youtube.com/watch?v=3z4Smk3mI18

1 SeqFeatR discovers statistically significant feature - sequence associations

Imagine the following alignment[1] of amino acid sequences in FASTA format, taken from 14 patients that either have a certain feature ("f") or do not have that feature ("n"). The feature may for instance be an HLA type[2], a genetic disease, etc. In the following FASTA formatted input for `SeqFeatR` you can see a letter (f or n) indicating feature or not-feature at the end of each comment line:

```
>P01_HLA_A01_00_B01_02;f
LPDIQGNENMGYQPSWIFCGMETNGSQCLEEMFHCCWINC
>P02_HLA_A01_00_B01_02;f
MPDWNQKWGNDHLASINLD-WLKTIQQPGIEKHLRFYENW
>P03_HLA_A01_02_B01_02;f
VPDASGKHGIIGMDVTSSMERRHGMVQLPWPAMVWGRPHW
>P04_HLA_A01_00_B01_02;f
MPDVRGVGCARRDCLIVHRFCMPFNNQVYCKVWIVYWTYK
>P05_HLA_A01_00_B01_02;f
QPDTPKITRKEATAIHKCGIHWQTNCQKLSTVHPFHHQVD
>P06_HLA_A01_02_B01_02;f
SWDDFSDFTMVHQWYAQGTLGPYKAMQLKMIFQGVSIMEV
>P07_HLA_A01_02_B01_02;f
IPDEPCYCCVKNKILTVEIGVHHAKSQVRRNIDNIRRKTE
>P08_HLA_A04_03_B04_03;n
HFST-ICPYIWKMYFTWMGQKLVIQKVNGRTPPHCDECNQ
>P09_HLA_A04_03_B04_03;n
SNFT-TTKLRDQHNLYPAGLQEIEHKVDHQILGIYGQIWY
>P10_HLA_A04_03_B04_03;n
ETSTALRTQDQTFMLALRANYMVMLKVLDCISVKLFICWR
>P11_HLA_A04_03_B04_00;n
DSSTMDAECSTLQRFIWWHAHYAWIRVAKKPYCLDCPYAV
>P12_HLA_A04_03_B04_03;n
KKSTLGIARGIQRSHGWYWRQTHCVMVLTPSQHKMGEKSW
>P13_HLA_A04_03_B04_00;n
ICSTELCGCLINWPPMQWIVFAHMDDVNDSQTNTCDMRSQ
>P14_HLA_A04_03_B04_03;n
GPSTNARTMGGQDCAYMTHTLTKHIWVILAFDPIMIVHKP
```

Can you discover statistically significant associations of the feature with the presence or absence of certain amino acids at certain sequence[3] positions? It is difficult to spot such associations with the naked eye, but they are there:

[1] The "alignment" shown here does not look like a good alignment. Usually, alignments show much more columns where many sequences have the same amino acids, and there may also be gaps, indicated by "-" in some sequences. This bad alignment was chosen for one reason only: it demonstrates that it can be difficult to spot relationships between features and sequence positions.

[2] Human leukocyte antigen (`https://en.wikipedia.org/wiki/Human_leukocyte_antigen`), a classification scheme of human immune systems.

[3] Strictly, we are not dealing with *sequence* positions but with *sequence alignment* positions.

- There is a strong association of feature f with amino acid P at the second position, though patient 14 is an exception as she is n and still has P at second position.
- There is a strong association of n (i.e. not having feature f) with amino acid T at fourth position, though patient 5 is an exception as he has f and still has a T at fourth position.

In its basic application, `SeqFeatR` tests *all* sequence positions and quickly identifies the second and fourth positions as being statistically significantly associated with the feature (f) or its absence (n). `SeqFeatR` shows these associations graphically in two ways, as Manhattan plot and as odds ratio (OR) plot.

2 The core of `SeqFeatR`: Fisher's exact test

2.1 An example: association of a feature with sequence

We have mentioned that in the above alignment there is seemingly a strong association of the occurrence of amino acid P at position 2 with the feature f. The probability and strength of this association can be quantified, respectively, by a p-value computed with Fisher's exact test, a well-known statistical test for association, and by an odds ratio (OR). At its core, `SeqFeatR` does exactly this.

In the above example of the association of P at position 2 with feature f, `SeqFeatR` internally would first collect occurrences in a frequency table and then compute from that frequency table p-value and OR:

- 6 sequences with feature f *and* P at position 2
- 1 sequence with feature f *and not* P at position 2
- 1 sequence with feature n (= *not* f) *and* P at position 2
- 6 sequences with feature n *and not* P at position 2

`SeqFeatR` collects these data in a frequency table:

		Proline	
		+	-
feature f	+	6	1
	-	1	6

Submitting this table to Fisher's exact test yields a p-value of 0.0291. At a significance level of 0.05 we therefore *reject* the null hypothesis (= *no* association of f and P at position 2) and rather assume an association of f and P at position 2.

The *strength* of the association is quantified by the odds ratio (OR) that is computed from the elements of the above contingency table:

$$OR = \frac{N_{f,P2}/N_{f,not\,P2}}{N_{not\,f,P2}/N_{not\,f,not\,P2}} = \frac{6/1}{1/6} = 36.$$

(Note: there are several methods to estimate the odds ratio. The simple one shown here is called Wald's method. The one used by `SeqFeatR` yields about 23.5.)

An OR much greater than 1 ($OR \gg 1$) as we have it here ($OR = 36$) means that we have a *strong positive association* of the feature f with $P2$: f and $P2$ occur much more often together than expected if we had no association.

If we have *a weak or no association*, the OR lies around 1. Then f and $P2$ would occur together and not together in the same ratios.

If we have a *strong negative association*, $0 < OR \ll 1$. In case of f and $P2$, a negative association means that f and $P2$ occur *less* often together than expected if we had no association.

2.2 Another example: association of HLA type and sequence

In the above set of FASTA formatted sequences, we had ended each sequence header with the name of a feature, either f or n, separated from the rest of the header line by a semicolon. A specific type of feature that is often used in `SeqFeatR` analyses is the *HLA type*. (For the HLA type there is an optional way of telling `SeqFeatR` about this specific feature by giving the positions of the HLA information in the FASTA header, see section 4 of this tutorial and first tab of `SeqFeatR` graphical user interface.) For instance, `SeqFeatR` will automatically discover in the sequences above a significant association between HLA*B01 and amino acid D at third sequence position:

- 7 sequences with HLA*B01 *and* D3
- 0 sequences with *not* HLA*B01 *and* with D3
- 0 sequences with HLA*B01 *and not* D3
- 7 sequences with *not* HLA*B01 *and not* D3

Thus, we obtain the following contingency table:

		D	
		+	-
HLA B*01	+	7	0
	-	0	7

Fisher's exact test yields a p-value < 0.001 and we have an OR of infinity. Thus we have a *significant and strongly positive* association of HLA*B01 and D3.

3 Graphical output

3.1 $-\log_{10}$ p-value plot ("Manhattan plot")

The so-called Manhattan plot, i.e. a plot of $-\log_{10}$ p-values along the sequence, is a convenient means to discover significant associations of sequence alignment positions with features. `SeqFeatR` produces Manhattan plots consisting of two separate plots (Figure 1): The top half of the plot focuses on complete epitopes or putative epitopes comprising "windows" of several sequence positions (e.g. windows of 9 positions), while the bottom half gives a more detailed picture of the same data at the level of single sequence positions. The x-axis for both plots is the same, namely the positions in the input sequence alignment.

Let us start the discussion with the more fundamental bottom part of Figure 1, the simple Manhattan plot. In this plot, `SeqFeatR` can mark a significance level α (here: $\alpha = 0.01$) with a horizontal line. Associations with $-\log_{10} p$-values above that line (i.e. p-values $< \alpha$) are shown with a special symbol (here: red stars) and considered significant. To ease the visual localization of the highly significant positions, they are additionally marked with vertical lines that hit the sequence axis at the corresponding positions. (The resolution of the x-axis is usually to coarse to show single positions, but sufficient to localize significantly associated positions in the fully resolved csv-file which is given as second output file.)

Now the top part of the Manhattan plot of Figure 1, focusing not on single positions but on complete putative epitopes. If the features in your `SeqFeatR` input have been HLA types, and if you then see several sequence positions in close proximity showing up with high $-\log_{10} p$ values in the Manhattan plot, you may have found a HLA epitope. The top part of the Manhattan plot highlights such position clusters. It can potentially show three different curves, a black, a red, and a yellow curve, each indicating potential epitopes. The red and yellow curves are optional.

The black curve directly relates to the bottom part of the plot: it shows the number of sequence positions with significant feature association in a window of 9 amino acids (or any other window length given by the user), divided by the window width (e.g. 9). The default window width of 9 amino acids corresponds to a typical length of a MHC I binding peptide. This window is shifted over the whole sequence and the fraction of significantly associated position computed for each window position and plotted as y-value at that position. An HLA epitope will show up as bump of the black curve to high values, similar to the bump in the top part of Figure 1.

The red curve (optional) allows the user to mark known epitopes, e.g. published in the literature or in a database. This can be helpful for comparisons. You can enter data for the red curve in an extra csv file ("Known epitopes" in `SeqFeatR` GUI and web interface). Here an example of what you could put into such a csv file[4]:

[4] A similar example is part of the R-package `SeqFeatR` and there called `Example_epitopes_aa.csv`.

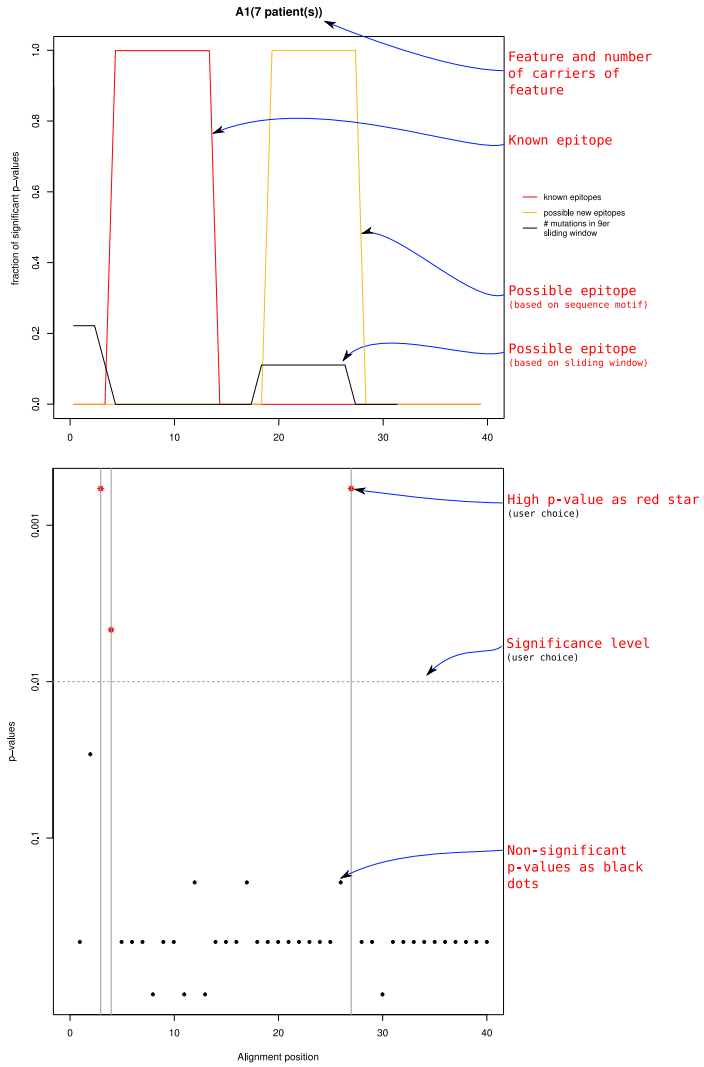

Figure 1: Manhattan plot (p-values along sequence). The y-axis scales with the $-\log_{10}$(p-value), i.e. the higher the point, the more significant the association. Top half of figure combines three different ways of showing possible epitopes: (1) possible epitopes from a window-wise statistical analysis of your data (black line), (2) known epitopes, e.g. from the literature (red line), (3) pieces of sequence alignment that conform with certain sequence patterns (yellow line). Bottom half of figure: Manhattan plot with the p-values for each sequence position. There are additional annotations in this graphic to explain what you see in red. Those are *not* in the real output from `SeqFeatR`.

```
4;12;A1
2;9;A3
```

This example marks two known epitopes with two lines of the form `EpitopeStart; EpitopeStop; HLAtype`. The first line (`4;12;A1`) corresponds to the epitope shown as bump of the red line in the top of Figure 1. The level of the bump is always at a value of 1.0 for these known epitopes. The bump is only shown if the HLA type in the csv file matches the HLA type in the sequence alignment.

With the yellow curve (optional) the user can mark alignment regions that are conforming with given sequence patterns. Such patterns can be defined in another csv file ("Known binding motifs" in `SeqFeatR` GUI and web interface) following this format[5].:

```
Genotype;Motif;Reference
A*01;x[PV]xxxx[DENQ]EN;SYFPEITHI
```

The header line (`Genotype; Motif; Reference`) describes the structure of the following lines. All elements in one row are separated by semicolons. The first element is the HLA type (here: `A*01`), the second is the actual definition of the motif (here: `x[PV]xxxx[DENQ]EN`), and the third gives a reference to the origin of this motif [6].

The definition of the motif requires a bit of explanation. The motif shown here covers nine amino acid positions. Here the nine amino acid positions are shown as indices:

$$\mathtt{x}_1 \underbrace{[PV]}_{2} \mathtt{x}_3\mathtt{x}_4\mathtt{x}_5\mathtt{x}_6 \underbrace{[DENQ]}_{7} \mathtt{E}_8\mathtt{N}_9$$

The letter `x` stands for: "this could be any amino acid". The two square brackets at positions 2 and 7 show which amino acids are allowed at these two positions, e.g. proline (P) or valine (V) at position 2. At position 8 there has to be a glutamate (E), and at position 9 must be an asparagine (N). Many sequences conform with this particular motif description, e.g. `APYEILDEN` or `SVRKTSQEN`. In general, motifs should be expressed in the same way using the elements `x`, `[]`, and capitals `ACDE...Y`. `SeqFeatR` shows a bump of the yellow line to 1.0 if several conditions are fulfilled simultaneously: (a) the HLA type in the sequence alignment matches the HLA type in the csv file, (b) the motif occurs in one of the aligned sequences, and (c) in the sequence window covered by the motif there is at least one significant association of an alignment position with the HLA.

3.2 Advanced `SeqFeatR` plotting, e.g. odds ratio plot

The R-package `SeqFeatR` offers a number of more advanced commands that are not yet available through the web interface of `SeqFeatR`. A nice example is the odds ratio plot that requires the use of the function `orPlot` of the R-package `SeqFeatR`. This is why you could be interested in the odds ratio plot:

[5] A similar example is part of the R-package `SeqFeatR` and there called `Example_HLA_binding_motifs_aa.csv`

[6] SYFPEITHI is just an example of a possible reference – the motif was not really taken from SYFPEITHI

While the Manhattan plot is a useful means to gain an overview over the distribution of significant sequence–feature associations along the alignment, there is still important information missing: Which amino acid is characteristic for the positions with low p-value? Is an amino acid overrepresented or underrepresented at such a position in sequences with a certain feature? All this information can be extracted from the csv file produced by `SeqFeatR`, but the program also provides a new plot, that we call *odds ratio plot*, to visualize this information (Figure 2).

We had introduced the odds ratio (OR) in section 2.1 as a way to quantify the strength of the association. The OR tells us whether a feature is over- or underrepresented at a position ($OR > 1$ or $OR < 1$, respectively), with $OR = 1$ indicating vanishing association. For easier visual recognition, we show the logarithms ($\log_{10} OR$) of the OR values along the sequence. To given an example: using this logarithmic representation, a tenfold overrepresentation of an amino acid at a sequence position in sequences with a certain feature (e.g. HLA type) shows up in the OR plot as *upwards* pointing bar of length 1, a tenfold underrepresentation as *downwards* pointing bar of length 1.

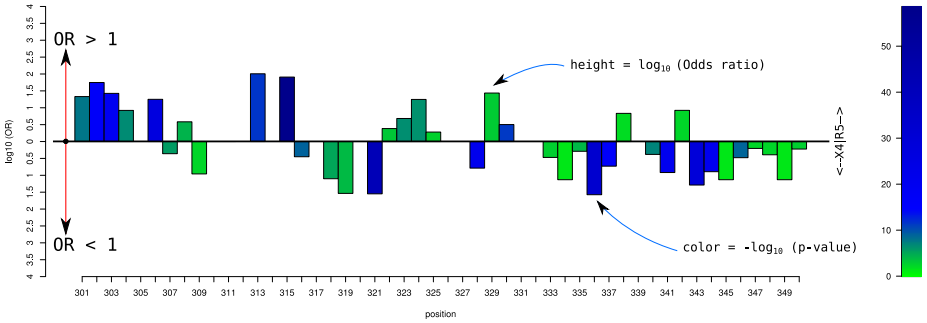

Figure 2: Example odds ratio plot. Here we analyze amino acid sequences of HIV-1 gp120 protein variants, and we have as feature the so-called co-receptor tropism of HIV-1, which can be "R5" or "not R5" (the latter is often called "X4"). The odds ratio (OR) plot shows for each sequence alignment position the association strength $\log_{10}(OR)$ as bar height and the p-value as bar color. The plot demonstrates that high values of $\log_{10} OR$ (long bars) and high statistical significance (blue color) are not the same.

There is one important caveat: even if a long OR bar indicates strong association, it still may not be *statistically significant*. This may be confusing at first, but think about a situation in which you have a small set of sequences, say two sequences, with a certain feature, and one without the feature. In the sequences with the feature we have amino acid A at the first alignment position, in the sequences without the feature we have G at that position. Thus A is overrepresented and G underrepresented in the sequences with the feature, right? In fact your OR would be infinity (see example in section 2.2). But would you believe this? Probably not, since your sequence set is so small that the p-value from the Fisher's exact test is 0.33.

Therefore the odds ratio plot allows you to combine OR information and p-value information. In Figure 2 OR bars with low p-values (highly significant) are filled

with a blue color, while OR bars with higher p-values (not so highly significant) are filled with a green hue. As you can see, by far not all long bars are blue, i.e. only a subset of positions may have strong and significant associations with the feature.

4 Input: sequences and features

The whole `SeqFeatR` analysis is critically dependent on the input. Therefore we summarize here how to prepare the input data properly.

First, before you submit sequences to `SeqFeatR`, you have to *align* them, i.e. do not submit sequences to `SeqFeatR`, but a multiple sequence alignment (MSA). There exist several popular tools for this task, for `mafft` (`http://www.ebi.ac.uk/Tools/msa/mafft/`). Make sure that the resulting MSA is in FASTA format (or Pearson/FASTA) (for a description of FASTA format, see `https://en.wikipedia.org/wiki/FASTA_format`, an example is shown on page 2). Usually, this output format can be chosen as option in the input form of `mafft` and other MSA programs, so you do not have to do this manually.

In the FASTA formatted MSA, we have one block for each sequence. Such a block consists of a header line (starting with >) that can be used to describe the sequence, and the sequence itself on the following lines. `SeqFeatR` expects in each FASTA header line a label that tells it which feature this sequence carries. There are two different types of features that can be put into FASTA headers:

1. A feature can be given by a letter or word at the end of each FASTA header after a semicolon, as in:

   - >some information;feature
   - >patient 1;f
   - >HCVA;n
   - etc.

   Anything after a semicolon in the FASTA-header is interpreted as name of a feature.

2. HLA types are a special case of features accepted by `SeqFeatR`. Here an example snippet from a FASTA file with the encoded HLA type information in the FASTA headers (see also example on page 2):

   ```
   >P1 HLA_A0403_B0403 donor 1
   SNFT-TTKLRDQHNLYPAGLQEIEHKVDHQILGIYGQIWY
   ETSTALRTQDQTFMLALRANYMVMLKVLDCISVKLFICWR
   DSSTMDAECSTLQRFIWWHAHYAWIRVAKKPYCLDCPYAV
   >P2 HLA_A0403_B0404 donor 2
   ...
   ```

   Now let us focus on the first header and for orientation write numbers `123...` beneath the header that give us the positions of the characters in that line:

```
>P1 HLA_A04_03_B04_03 donor 1
123456789.123456789.123456789.1234
```

Here, `HLA_A04_03` corresponds to HLA-A*04:03[7] with locus A, group number 04, and variant 03. The group number "04" covers positions 10 and 11 of the line, the variant "03" is at positions 13 and 14. Analogously, for the B locus we have group "04" at positions 17 and 18, and variant "03" at positions 20 and 21. We give `SeqFeatR` these four position intervals of A and B groups and variants, as shown in Figure 3. Importantly, the HLA types in *all* FASTA blocks in one MSA file have to take the *same* positions in their respective lines to be properly recognized by `SeqFeatR`.

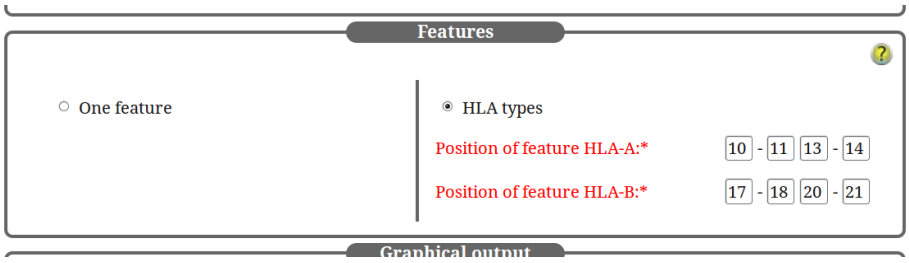

Figure 3: Screenshot of `SeqFeatR` web interface with filled HLA positions.

5 Multiple comparison correction

`SeqFeatR` applies the same statistical test to *many* sequence alignment positions. Under these conditions it is likely that some of the statistical test yield a low p-value just by chance, and not because of a real association.

An example may illustrate this so-called *multiple testing problem* (or *multiple comparison problem*): imagine you toss a fair coin (fair = fifty-fifty chance for head and tail) four times. The most probably result is that you have 2 times head and 2 times tail. The probability to have 4 times head is only 1 in 16 (1/2 for the first toss times 1/2 for the second ... $= (1/2)^4 = 1/16$). Now let us do 16 of these 4-toss experiments, i.e. we toss the coin 16×4 times. Then we expect that in one of these 4-toss experiments we will see 4 times head, just by chance. If we only consider the experiment with the outcome of 4 times head, we have the wrong impression that the coin is *not* fair. In the same way, a set of sequence alignment positions corresponds to a set of random experiments, and if we carry out many association tests, some of them seemingly indicate an association, but that association is not real.

`SeqFeatR` offers various types of corrections (graphical user interface/web interface: P-value correction) as available through the standard R-stats package function `p.adjust` (in R call help page of p.adjust for more information). This includes e.g. very conservative Bonferroni correction or the probably more useful False Discovery Rate correction.

[7] See e.g. `https://en.wikipedia.org/wiki/Human_leukocyte_antigen#Nomenclature`

6 Hints

- Do **not** use a word processor such as Word or LibreOffice/OpenOffice to prepare the sequence alignment input. These programs likely destroy the FASTA format by (invisible) extra characters and invalidate the input. Instead, use an editor for raw text such as `notepad` in Windows systems, or `gedit`, `vim` etc. on Linux systems.
- `SeqFeatR` can only understand letters from canonical DNA/RNA (A, C, G, T, U) and amino acid alphabets (A, C, D, E, ...) in the FASTA sequences. Do not use special characters (?!; etc.) or characters for wobbels (R, K, Y, ...) in sequences. You can however use X or B for unidentified nucleic acids or amino acids.