Package 'EpitopeMatcher'

November 21, 2014

Title Epitope Matcher

Description A package that can be used to find out how well the epitopes in a patient's virus' will be recognized by the HLA's present in the patient.

Version 0.1

Author Phillip Labuschagne <philliplab@gmail.com>

Maintainer Phillip Labuschagne <philliplab@gmail.com>

Depends digest,
testthat,
Biostrings,
methods

Imports shiny

License GPL-3

URL https://github.com/philliplab/EpitopeMatcher

Collate 'data_documentation.R'
'lanl_hla_data.R'
'list_jobs.R'
'patient_HLA.R'
'msa.R'
'matcher.R'
'scoring_job.R'
'shiny_app.R'

R topics documented:

`BETWEEN10` *HIV Scoring Matrix 10 for between host evolution*

Description

HIV Scoring Matrix 10 for between host evolution

Usage

```
data(BETWEEN10)
```

Format

A 24x24 matrix

`BETWEEN38` *HIV Scoring Matrix 38 for between host evolution*

Description

HIV Scoring Matrix 38 for between host evolution

Usage

```
data(BETWEEN38)
```

Format

A 24x24 matrix

build_scoring_jobs *Builds scoring jobs from pre-processed inputs*

Description

Called by list_scores_to_compute

Usage

```
build_scoring_jobs(matched_patients, flat_lanl_hla)
```

Arguments

matched_patients	As produced by match_patient_hla_to_query_alignment.
flat_lanl_hla	As produced by flatten_lanl_hla

epitope_found *A function that indicates if an epitope was found in the reference sequence.*

Description

An epitope is seen as not found if the alignment between the epitope and the reference does not contain the same number of bases as the epitope. This should maybe be relaxed in the future.

Usage

```
epitope_found(epitope, alignment)
```

Arguments

epitope	The epitope that was searched for in the reference
alignment	The alignment to check - usually the output from pairwiseAlignment

find_epitope_in_ref *Finds the position of the epitope in the reference sequence*

Description

It uses pairwiseAlignment with the default settings. See Biostrings manual.

Usage

```
find_epitope_in_ref(epitope, query_alignment, alignment_type = "overlap")
```

Arguments

epitope	The epitope to find in the sequence. Either a character string or an AAString
query_alignment	The query alignment
alignment_type	The type of alignment to try. Defaults to 'overlap' use 'global' if 'overlap' alignment cannot be found. See Biostrings manual.

Details

Lots of things to test and investigate that can potentially improve this function.

flatten_lanl_hla *Flattens the LANL HLA file*

Description

Sometimes the same hla has a number of different names. Hence the hla_genotype column in the lanl file must be processed before the matches can be made.

Usage

```
flatten_lanl_hla(lanl_hla)
```

Arguments

lanl_hla	The data.frame (of class LANL_HLA_data) that contains the descriptions of the different HLA genotypes

Details

This function takes the LANL HLA file and transforms it so that each row correspond to one and only one hla. This means that row in which the 'hla_genotype' column is unpopulated gets discarded and that rows in which the 'hla_genotype' column contains the names of more than one hla_genotype (assumed to be seperated by commas) will be duplicated and each duplicate will be assigned to one hla_genotype.

get_epitope *Returns the epitope associated with a scoring_job*

Description

Returns the epitope associated with a scoring_job

Usage

```
get_epitope(the_scoring_job)

## S4 method for signature 'Scoring_Job'
get_epitope(the_scoring_job)
```

Arguments

the_scoring_job	The scoring job which epitope must be extracted

get_hla_details *Returns the hla details associated with a scoring_job*

Description

Returns the hla details associated with a scoring_job

Usage

```
get_hla_details(the_scoring_job)

## S4 method for signature 'Scoring_Job'
get_hla_details(the_scoring_job)
```

Arguments

the_scoring_job	The scoring job whose details must be extracted

---

`get_patient_ids,AAStringSet-method`
Returns the ids of the patients in the data structure

Description

It parses the FASTA header when called on a XStringSet object using the sep and id_position arguments.

Usage

```
## S4 method for signature 'AAStringSet'
get_patient_ids(x, sep = "\\|", id_position = 1)

get_patient_ids(x, sep = "\\|", id_position = 1)

## S4 method for signature 'Patient_HLA'
get_patient_ids(x)
```

Arguments

x	The data structure to interrogate
sep	The symbol used to separate elements in the sequence names
id_position	After the sequence name has been split on the 'sep' character, which element of the resulting vector contains the patient id?

`get_test_lanl_hla_data`
A function that returns a test lanl hla genotype dataset

Description

A function that returns a test lanl hla genotype dataset

Usage

```
get_test_lanl_hla_data(dataset_name = "base")
```

Arguments

dataset_name	The name of the test dataset to return

get_test_patient_hla_data

A function that returns a test patient hla dataset

Description

A function that returns a test patient hla dataset

Usage

```
get_test_patient_hla_data(dataset_name = "base")
```

Arguments

dataset_name	The name of the test dataset to return

get_test_query_alignment

A function that returns a test query alignment

Description

A function that returns a test query alignment

Usage

```
get_test_query_alignment(dataset_name = "base")
```

Arguments

dataset_name	The name of the test dataset to return

`LANL_HLA_data-class` *The class for the data from LANL that describe the HLA genotype's epitopes*

Description

This class is an extension of the data.frame class placing some extra restrictions on the data format. This data is typically obtained from http://www.hiv.lanl.gov/content/immunology/tables/tables.html.

Details

The data.frame has the following columns:

- Epitope
- Protein
- HXB2.start
- HXB2.end
- Subprotein
- HXB2.DNA.Contig
- Subtype
- Species
- HLA

`list_scores_to_compute` *Processes the three input files (query_alignment, patient_hla and lanl_hla) in to a list of scoring jobs.*

Description

First the patient_hla data is matched to the query_alignment `match_patient_hla_to_query_alignment`, then the lanl_hla file is flattened `flatten_lanl_hla`, and finally, the jobs are built `build_scoring_jobs`.

Usage

```
list_scores_to_compute(query_alignment, patient_hla, lanl_hla)
```

Arguments

`query_alignment`	An AAStringSet that contains the multiple sequence alignment of the patient's viral sequences
`patient_hla`	The data.frame that specifies which query sequence to check against which hla genotypes
`lanl_hla`	The data.frame (of class LANL_HLA_data) that contains the descriptions of the different HLA genotypes

Details

This list of jobs can then be used to perform the comparisons.

Value

A list of Scoring_Jobs

`match_epitopes` *Computes similarities between certain epitopes and sequences*

Description

A query alignment and a file specifying which hla_genotypes should be checked for different patients are first compared to construct a list of scores that must be computed. This list is then passed to the score_all_epitopes function which computes the scores. The results and error logs are then returned as output.

Usage

```
match_epitopes(query_alignment, patient_hla, lanl_hla_data,
  range_expansion = 5, update_progress_bar = NULL,
  substitutionMatrix = "BLOSUM50")
```

Arguments

`query_alignment`	The query alignment
`patient_hla`	The data.frame (of class Patient_HLA) that contain lists all the HLA genotypes each patient has.
`lanl_hla_data`	The data.frame (of class LANL_HLA_data) that contains the descriptions of the different HLA genotypes
`range_expansion`	After the epitope is found in the reference seqeuence, search in each of the query sequences for the same epitope, but expand the range with this number of amino acids
`update_progress_bar`	A closure passed in from a reactive shiny expression that allows a progress bar to be updated when using the shiny web ui.
`substitutionMatrix`	substitution matrix representing the fixed substitution scores for an alignment. It cannot be used in conjunction with <e2><80><98>patternQuality<e2><80><99> and <e2><80><98>subjectQuality<e2><80><99> arguments.

`match_patient_hla_to_query_alignment` *Matches the patient_hla data to the query sequence to check which hla's must be checked for in which sequences.*

Description

It treats the patient_id column in the patient_hla data as a regular expression and performs the lookup to the query sequence names.

Usage

```
match_patient_hla_to_query_alignment(query_alignment, patient_hla)
```

Arguments

`query_alignment`	An AAStringSet that contains the multiple sequence alignment of the patient's viral sequences
`patient_hla`	The data.frame that specifies which query sequence to check against which hla genotypes

Value

A list of lists. The inner lists contains the elements 'hla_genotype' and 'query_sequence_names'

`Patient_HLA-class` *The class for the data that describes which patients have which HLAs*

Description

It is an extension of data.frame and has these columns:

- patient_id
- hla_genotype

Details

The value of the patient_id column gets treated as a regular expression when it is matched to the FASTA headers in the query_alignment. If the value is set to .* then the hla_genotype corresponding to this entry will be matched to all the sequences in the query alignment.

The hla_genotype column should be a name from the LANL table.

read_lanl_hla *A function that reads a HLA genotype specification file*

Description

This function converts the file into an object of class LANL_HLA. See the documentation of that function for more details: `.LANL_HLA_data`

Usage

```
read_lanl_hla(file_name)
```

Arguments

`file_name`	Name of the file

read_patient_hla *A function that reads a patient HLA genotype specification file.*

Description

The function will convert the file into an object of class Patient_HLA. See the class documentation for more details about the format: `.Patient_HLA`

Usage

```
read_patient_hla(file_name)
```

Arguments

`file_name`	Name of the file

`read_query_alignment` *Reads in the query alignment*

Description

This function currently just calls readAAStringSet in the Biostrings package. See that function for more details.

Usage

```
read_query_alignment(file_name)
```

Arguments

`file_name`	Name of the fasta file

Details

Must be a valid FASTA file.

The FASTA headers must be in some delimited form with a special character used for delimiting different fields. Further more, the patient id must always be in the same column in the FASTA header. For example: PATID_TIME_PID_CONSENSUSDETAILS

`run_EpitopeMatcher_app` *Runs the shiny app for this package*

Description

Runs the shiny app for this package

Usage

```
run_EpitopeMatcher_app(port = 5436)
```

score_all_epitopes *Given a list of scoring jobs, compute the similarities*

Description

This function is mostly just a wrapper for score_epitope.

Usage

```
score_all_epitopes(the_scoring_jobs, query_alignment, range_expansion,
  update_progress_bar = NULL, substitutionMatrix = "BLOSUM50")
```

Arguments

the_scoring_jobs	A list of scoring_jobs
query_alignment	The query alignment
range_expansion	After the epitope is found in the reference seqeuence, search in each of the query sequences for the same epitope, but expand the range with this number of amino acids
update_progress_bar	A closure passed in from a reactive shiny expression that allows a progress bar to be updated when using the shiny web ui.
substitutionMatrix	substitution matrix representing the fixed substitution scores for an alignment. It cannot be used in conjunction with *patternQuality* and *subjectQuality* arguments.

score_epitope *Computes the similarities between the epitope and the sequences in the alignment*

Description

Computes the similarities between the epitope and the sequences in the alignment

Usage

```
score_epitope(the_scoring_job, query_alignment, range_expansion = 0,
  substitutionMatrix = "BLOSUM50")
```

Arguments

`the_scoring_job`
: A scoring job as a object of type 'Scoring_Job'

`query_alignment`
: The query alignment

`range_expansion`
: After the epitope is found in the reference seqeuence, search in each of the query sequences for the same epitope, but expand the range with this number of amino acids

`substitutionMatrix`
: substitution matrix representing the fixed substitution scores for an alignment. It cannot be used in conjunction with <e2><80><98>patternQuality<e2><80><99> and <e2><80><98>subjectQuality<e2><80><99> arguments.

Value

The output from this function is a list with two data.frames. The first is the results data.frame that contains these columns:

- sequence_id - The sequence description from the FASTA file
- score - The similarity score produced by the alignment
- score_type - The type of similarity score as returned by pairwiseAlignment
- eregion_in_refseq - The region of the reference sequence that was attempted to be aligned to the query sequence as returned by pairwiseAlignment
- candidate_substr - The candidate substring that was obtained by expanding the coordinates found in the reference sequence by 'range_expansion' AAs on each side (unless at the end or beginning of the sequence)
- matched_substr - The part of the candidate substring that was matched to the epitope as returned by pairwiseAlignment
- comparison - A comparison between the epitope and the query sequence indicating where there were mismatches
- pid - The percentage of amino acids that were identical (Percentage IDentity) between the epitope and query sequences
- simple_distance - 100 - PID
- nmatch - The number of matches in the alignment
- nmismatch - The number of mismatches in the alignment
- leven.dist - The Levenshtein distance (or edit distance) between the two sequences
- start_pos_in_ref - The starting position in the reference sequence of the matching substring that was found for the epitope
- end_pos_in_ref - The end position in the reference sequence of the matching substring that was found for the epitope
- start_pos_in_candidate - The starting position in the candidate subsequence of the query sequence that was obtained by expanding the range of the reference that matches the epitope by starting a number of amino acids earlier in the query sequence. The number of amino acids is controlled by the range_extention parameter.

- end_pos_in_candidate - The end position in the candidate subsequence of the query sequence that was obtained by expanding the range of the reference that matches the epitope by stopping a number of amino acids later in the query sequence. The number of amino acids is controlled by the range_extention parameter.
- range_expansion - The number of amino acids by which the range of the query sequence that is compared to the epitope is larger than then match found for the epitope in the reference sequence.
- These three column are usually added to the table by the score_sequence_epitopes function:
  - epitope - The epitope from the lanl file that was searched for in the reference sequence
  - hla_genotype - The name of the hla genotype the epitope is associated with
  - lanl_start_pos - The start position of the epitope according to the lanl file
  - lanl_end_pos - The end position of the epitope according to the lanl file

The second element of the list is the error log data.frame that contains these columns:

- pattern - The epitope as aligned to the reference sequence when a less restrictive alignment algorithm is used than the one that failed when aligning to the reference sequence the first time
- subject - The portion of the reference sequence to which the epitope was aligned to when a less restrictive alignment algorithm is used than the one that failed when aligning to the reference sequence the first time
- global_alignment_start - The starting position in the reference sequence of the subsequence of the reference sequence that the epitope was aligned to when a less restrictive alignment algorithm is used than the one that failed when aligning to the reference sequence the first time
- global_alignment_end - The end position in the reference sequence of the subsequence of the reference sequence that the epitope was aligned to when a less restrictive alignment algorithm is used than the one that failed when aligning to the reference sequence the first time
- These three column are usually added to the table by the score_sequence_epitopes function:
  - epitope - The epitope from the lanl file that was searched for in the reference sequence
  - hla_genotype - The name of the hla genotype the epitope is associated with
  - lanl_start_pos - The start position of the epitope according to the lanl file
  - lanl_end_pos - The end position of the epitope according to the lanl file

---

`Scoring_Job-class` *A class that contains all the required information to run an epitope scoring job.*

Description

Three pieces of information is required for a scoring job to be valid:

- hla_genotype - The name of the hla_genotype to investigate
- query_sequence_names - A character vector of the names of the query sequences in which this hla_genotype must be looked for
- hla_details - A list of further details about this hla_genotype

Details

The hla_details is forced to have the following values: "end_pos", "epitope", "gene_name", "hla_genotype", "hxb2_dna_position", "organism", "start_pos", "subprotein", and "sub_type". This is reasonable since the HLA data will always come from the LANL file and these details must be added to the results and error logs.

WITHIN5 *HIV Scoring Matrix 5 for within host evolution*

Description

HIV Scoring Matrix 5 for within host evolution

Usage

```
data(WITHIN5)
```

Format

A 24x24 matrix

Index