EpiGRAPH Attribute Reference Sheet

All attributes calculated by EpiGRAPH are given names according to the following hierarchical naming schema:

<full-attribute-identifier> ::= [<window>.]<attribute-group-name>.<attribute-name>.<column-name>

- At the top level (**<attribute-group-name>**), an attribute group pools a set of biologically related attributes, e.g. DNA sequence patterns in the attribute group "DNA_Sequence" or gene-related attributes in the attribute group "Genes".
- The intermediate level (**<attribute-name>**) refers to a set of attributes and columns that are derived from the same dataset, e.g. "RefSeq_Genes" and "CCDS", both belonging to the attribute group "Genes".
- The bottom level (**<column-name>**) refers to a specific column in the table of attributes that EpiGRAPH calculates. All column names are given based on rules identifying a certain mode of calculation (e.g. frequency of overlap or average score), which are described in more detail below.
- An additional top level (**<window>**) is used when the attribute calculation includes not only the genomic regions provided by the input dataset, but also adjacent windows upstream and downstream.

Because the full attribute names are often quite long, we also use a shorthand, which is for example used in the column header of EpiGRAPH's dataset of calculated attributes (downloadable from the EpiGRAPH website via the "Download Data Table" button on EpiGRAPH's results overview page). A complete mapping from long to short attribute names is provided in the corresponding X-GRAF file, which can be downloaded via the "Download XML Documentation" button on the results overview page.

Below, we describe four types of attributes that are calculated in different ways as indicated by the **<column-name>** segment of their full attribute name. Attribute types 1. and 3. are most common within EpiGRAPH.

1. DNA sequence attributes (calculated from pattern frequencies)

The columns of DNA sequence attributes represent the frequency of appearance of a specific DNA sequence pattern (e.g. "CGCG") within the genome sequence of the genomic region specified in the input file. The names of the corresponding columns are composed according to the following rule ("+" stands for string concatenation):

"Pat_"+<pattern-string>+"_freq"

For every sequence pattern, EpiGRAPH provides the option to compute additional statistics on the pattern occurrence throughout the region, such as standard deviation, skewness and kurtosis of the frequency values. These statistics are computed by dividing the genomic region into subregions and calculating the frequency of the pattern in each subregion. This results in a set of pattern frequencies from which standard deviation, skewness and kurtosis are computed. The names of the corresponding columns are composed according to the following rule:

"Pat_"+<pattern-string >+"_std"	–	for standard deviation
"Pat_"+<pattern-string >+"_skew"	–	for skewness
"Pat_"+<pattern-string >+"_kurt"	–	for kurtosis

All pattern frequencies can be computed either in a strand-specific or non-strand-specific way (strand specificity refers to the genomic plus-strand, not to the direction of transcription of the nearest gene). Strand specificity is indicated in the column names according to the following rule:

<strand-specific-pattern-string > ::= "plus"|"minus"+<pattern-string>

2. DNA structure attributes (calculated from oligomers with known structure)

DNA structure predictions are calculated for a given genomic region by sliding a window of fixed size over the region and comparing the DNA sequence pattern in this window with a set of oligomers with known structure (which is described by numerical score values). For example, the predicted helix structure of all possible octamers has been quantified by a set of six numeric scores: twist, roll, tilt, rise, slide and shift (Gardiner, et al. 2003, J Mol Biol).

For each score, a new column named **<score-name>** is added, its value being the mean of the scores corresponding to all oligomer hits observed while shifting the sliding window over the genomic region. Similar to the pattern frequency attributes described above, we also report standard deviation (**<score-name>+”_std”**), skewness (**<score-name>+”_skew”**) and kurtosis (**<score-name>+”kurt”**).

3. Patch attributes (quantifying overlap with all kinds of genomic regions)

Patch attributes describe the frequency of overlap between the genomic regions in the input dataset and various types of other genome annotations that take the form of genomic regions (e.g. CpG islands, repetitive regions and SNPs). For every patch attribute, three basal columns are introduced, which report general statistics about the overlap between the regions in the input dataset and the patch attribute:

“overlapRegionsCount” – total number of patch attribute regions overlapping the input region, standardized to 1kb

“overlapTotalLength” – total length of patch attribute regions overlapping the input region, standardized to 1kb

“overlapAverageSize” – average size of the overlapping regions

In addition to its three basic features (chromosome, start position, end position), a patch attribute may also contain additional columns that can be numeric (referred to as score attributes), binary (class attributes) or categorical (category attributes), giving rise to additional columns during the attribute calculation.

Score attributes give rise to columns with the same name as the score column in the patch attribute and are calculated as weighted averages of the patch regions overlapping with the region specified in the input dataset. Weighting is performed according to the length of overlap.

Class attributes give rise to columns with the same name as the class column in the patch attribute and are calculated as the dominant class among the patch regions overlapping with the region specified in the input dataset. Furthermore, distribution statistics for each class are reported in additional columns:

<class-name>+”_”+<class-value>+”_overlapRegionsCount” – total number of patch attribute regions with value **<class-value>** for class **<class-name>** overlapping with the input region, standardized to 1kb

<class-name>+”_”+<class-value>+"_overlapTotalLength” – total length of patch attribute regions with value **<class-value>** for class **<class-name>** overlapping the input region, standardized to 1kb

<class-name>+”_”+<class-value>+"_overlapAverageSize” – average size of the overlapping patch attribute regions with value **<class-value>** for class **<class-name>**

Category attributes split the patch attribute into several sub-attributes, and the standard measures of overlap are calculated separately for each category, giving rise to the following columns:

<category-name>+”_”+<category-value>+"_overlapRegionsCount” – total number of patch attribute regions with value **<category-value>** for class **<category-name>** overlapping with the input region, standardized to 1kb

<category-name>+”_”+<category-value>+"_overlapTotalLength” – total length of patch attribute regions with value **<category-value>** for class **<category-name>** overlapping the input region, standardized to 1kb

<category-name>+”_”+<category-value>+"_overlapAverageSize” – average size of the overlapping patch attribute regions with value **<category-value>** for class **<category-name>**

Furthermore, for each category EpiGRAPH reports separate averages for all score attributes. The names of the corresponding columns are composed according to the following rule:

“c”+<category-name>+<category-value>+”_o”+<score-name> – mean score of all patch attribute regions belonging to the category **<category-value>** and overlapping with the input region

Genomic strand columns are treated as special type of categories, and the columns derived from a strand-specific patch attribute are named by the same rules, except for the prefix **“c”** being changed to **“s”**:

“s”+<strand-name>+<strand-value>+”_o”+<score-name>

4. Gene attributes (quantifying overlap with genes and exons)

Gene attributes are a special case of patch attributes that take the specific structure of eukaryotic genes (exons and introns) into account. They contain a number of additional columns, with names composed according to the following rules

<attribute-name>+”_elen” – total length of exonic DNA within the region, standardized to 1kb

<attribute-name>+”_eno” – total number of exons within the region, standardized to 1kb

<attribute-name>+”_eavg” – average length of the exons overlapping the target region

<attribute-name>+”_estd” – standard deviation of the lengths of the exons overlapping the region

<attribute-name>+”_glen” – total length of genic DNA within the region, standardized to 1kb

<attribute-name>+”_gno” – total number of genes within the region, standardized to 1kb

<attribute-name>+”_gavg” – average full of the genes overlapping the region

<attribute-name>+”_gstd” – standard deviation of the lengths of the genes overlapping the region

<attribute-name>+”_gcav” – average number of exons per gene

<attribute-name>+”_gcsd” – standard deviation of the exon number per gene