File format definitions for the source data in the integrin database

The following files contain the primary data that needs to be included in the database:

**integrin\_monomers.xlsx**

Holds information about both *alpha* and *beta* integrin subunits. Contains the following columns:

* **Integrin\_name** – the name of the subunit, in the form of alpha-x or beta-x, where x is a bunch of alphanumeric characters (usually one, but not always)
* **UniProt\_accession** – the primary accession of the protein
* **Ensembl\_accession** – Ensembl gene accession, as provided by UniProt ID mapping
* **Gene\_name** – taken from UniProt, always a single run of alphanumeric characters
* **Alternative\_names** – taken from UniProt, can contain multiple values separated by the pipe character ('|')
* **Function** – taken from UniProt, free text
* **Length** – taken from UniProt, single integer number
* **Possible\_dimerization\_partners** – contains one or more integrin names (in the format alpha-x or beta-x). For alpha subunits only beta partners are allowed, and vice versa. If there are more than one dimerization partners, they are separated by the pipe character ('|')
* **Domains\_and\_regions** – contains the location of domains in the sequence. First a domain name is specified (e.g. SPa), followed by the symbol '@', followed by the sequence boundaries of the domain. Multiple domains are separated by pipe ('|'). E.g. in the case of the first three domains of alpha-X: SPa@1-19|bP1@35-85|bP2@86-145. Possible domains for alpha and beta subunits are defined in the domain\_shorthands.xlsx file.
* **Sequence** – standard FASTA format, taken from UniProt. Formatting is preserved (header line+sequence lines of max 60 residues), but newline characters are replaced by '&'.
* **Expression** – details on the expression pattern of the protein, based on HPA annotations, followed by link to the appropriate HPA page in parantheses.
* **Structures** – IDs of PDB structures that contain any domain of the subunit. Multiple PDB IDs are separated by '|'.
* **Notes** – additional information to be included on the server page. Currently not used.

**integrin\_dimers.xlsx**

Holds information about all 24 human integrin dimers. Contains the following columns:

* **Dimer\_name** – the name of the dimer, being composed of two monomer names, separated by a slash, e.g.: alpha-X/beta-2
* **Expression** – details on the expression pattern of the protein, based on HPA annotations.
* **Notes** – additional information to be included on the server page. Currently not used.

**integrin\_structures.xlsx**

Holds information about all PDB entries that contain at least one domain/fragment of any alpha or beta integrins. Contains the following columns:

* **PDB\_ID**
* **ExpTech** – experimental technique used to determine the structure. Can be 'X-ray', 'NMR', 'Electron microscopy' or 'Model'.
* **Resolution** – resolution of X-ray and electron microscopy structures, measured in Angstroms. For all other techniques, it is 'N/A'.
* **alpha\_integrin** – the type of the included alpha integrin subunit (if any). Only one subunit name is allowed.
* **alpha\_domains** – a list of alpha subunit domains visible in the structure. Possible domains are defined in the domain\_shorthands.xlsx file.
* **beta\_integrin** – the type of the included beta integrin subunit (if any). Only one subunit name is allowed.
* **beta\_domains** – a list of beta subunit domains visible in the structure. Possible domains are defined in the domain\_shorthands.xlsx file.
* **Protein\_interactors** – other non-integrin interactors that are present in the structure. Two types are defined:
  + native – the interacting protein is a human protein. It is defined by the keyword 'native' followed by a colon, followed by a UniProt accession, region boundaries, and a protein name in parentheses. E.g.: native:Q96A83@308-329(Collagen alpha-1(XXVI) chain)
  + non-native – the interacting protein comes from an organism other than human. It is defined by the keyword 'non-native' followed by the organism name and taxonomic group separated by '\*' in parentheses. This is followed by a colon, and the UniProt accession, region boundaries, and protein name, similarly to native interactors. E.g.: non-native(Foot-and-mouth disease virus - type O\*Virus):Q6PMW3@287-933(Genome polyprotein)

In case of multiple interaction partners, interactors are separated by pipe '|'.

* **Other\_interactors** – other interactors that are present in the structure, but do not have a UniProt accession for whatever reasons. Four types are defined:
  + peptide – the interacting protein is a (usually synthetic) peptide. It is defined by the keyword 'peptide' followed by a colon, followed by a free text defining the peptide. E.g.: peptide:Cyclic Arg-Gly-Asp Peptide
  + antibody – the interacting protein is an antibody. It is defined by the keyword 'antibody' followed by a colon, followed by some free text defining the antibody. E.g.: antibody:Monoclonal Antibody 10e5 Heavy Chain
  + drug – the interacting protein is an known pharmaceutical agent. It is defined by the keyword 'drug' followed by the drug name in parentheses. This is followed by a colon, followed by some free text defining the drug. The drug name in parentheses must match the name of exactly one drug name from the first column of the integrin\_drugs.xlsx file. E.g.: drug(Efalizumab):Efalizumab Fab Fragment, Light Chain
  + other – the interacting protein is something that fits none of the above definitions. It is defined by the keyword 'other' followed by a colon, followed by a free text defining the interactor. E.g.: other:Collagen

In case of multiple interaction partners, interactors are separated by pipe '|'.

**domain\_shorthands.xlsx**

Holds information about possible domains in both alpha and beta integrins. This information is intended to translate the domain shorthands used in the integrin\_monomers.xlsx and integrin\_structures.xlsx for names that should be displayed. Contains the following columns:

* **Subunit\_type –** the type of subunit the domain belongs to. Can be either 'alpha' or 'beta'
* **Domain\_shorthand** – the ID of the domain
* **Domain\_name** – the regular name of the domain

**integrin\_drugs.xlsx**

Holds information about known drugs that were developed against integrins. Contains the following columns:

* **Name** – the standard name of the drug
* **Marketing\_name** – the name under which the drug was/is marketed to the general public. If none exists, this field should be 'N/A'.
* **Status** – the status of the drug, can be marketed, followed by the information if it is prescription only. If the drug has been withdrawn, the year of the withdrawal should be indicated.
* **Type** – can be 'cyclic peptide', 'monoclonal antibody', or 'small molecule'.
* **Administration** – the standard route of administration for patients.
* **ATC\_code** – the code of the drug in the Anatomical Therapeutic Chemical (ATC) Classification System.
* **ATC\_definition** – the class the drug belongs to, according to ATC nomenclature.
* **Target\_integrin** – integrin(s) the drug is known to bind to. In case of multiple integrins, values are separated by pipe '|'.
* **Lauch\_year** – the year the drug has entered clinical use (not trials!).
* **Notes** – notes describing the drug.

**integrin\_interactors.xlsx**

Holds information about proteins known to interact with integrins. Contains the following columns:

* **Protein\_name** – the name of the interacting protein taken from UniProt.
* **Organism** – the name of the source organism. If this is human, the interactor is *native*, otherwise it's *non-native*.
* Taxonomic\_group – a higher level taxonomic group the source organism belongs to. E.g.: 'Viperidae' (for snake venoms), 'Bacteria', 'Virus', 'Human', etc.
* **UniProt\_accession** – the primary accession of the protein
* **Gene\_name** – taken from UniProt, always a single run of alphanumeric characters
* **Alternative\_names** – taken from UniProt, can contain multiple values separated by the pipe character ('|')
* **Function** – taken from UniProt, free text
* This is followed by 3\*24 columns, defining the interaction between the protein and each of the 24 integrins:
  + one column defines if there is an interaction (header: '**alpha-x/beta-x**', values: '1' for interaction, '0' for non-interaction, <blank> for no info)
  + a second column lists references to papers supporting the quoted interaction data (header: '**alpha-x/beta-x\_ref**', values are links to PubMed)
  + a third column contains free text describing the binding strength, if applicable.
* **Site\_boundaries** – location of the core motif in the sequence.
* **Site\_definition** – free text from UniProt describing the site
* **Structural\_state** – the structural state of the protein around the core motif. Can either be 'Ordered' or 'Disordered', followed by the basis of the assertion in parentheses. E.g.: Disordered (based on prediction)
* **Motif\_type** – a 3-letter definition of the core motif
* **Sub-sequence** – the sequence of the core motif, surrounded by 20 residues of flanking regions to both sides. If the motif is N- or C-terminal, missing residues are marked by dashes '-'.
* **Structures** – Structures that contain the interacting region defined in the entry. PDB ID followed by dash '-', followed by the chain IDs in the structure that represent the interactor. Can contain multiple chain IDs concatenated. E.g.: 3C05-AC Can contain multiple structures separated by pipe '|'.