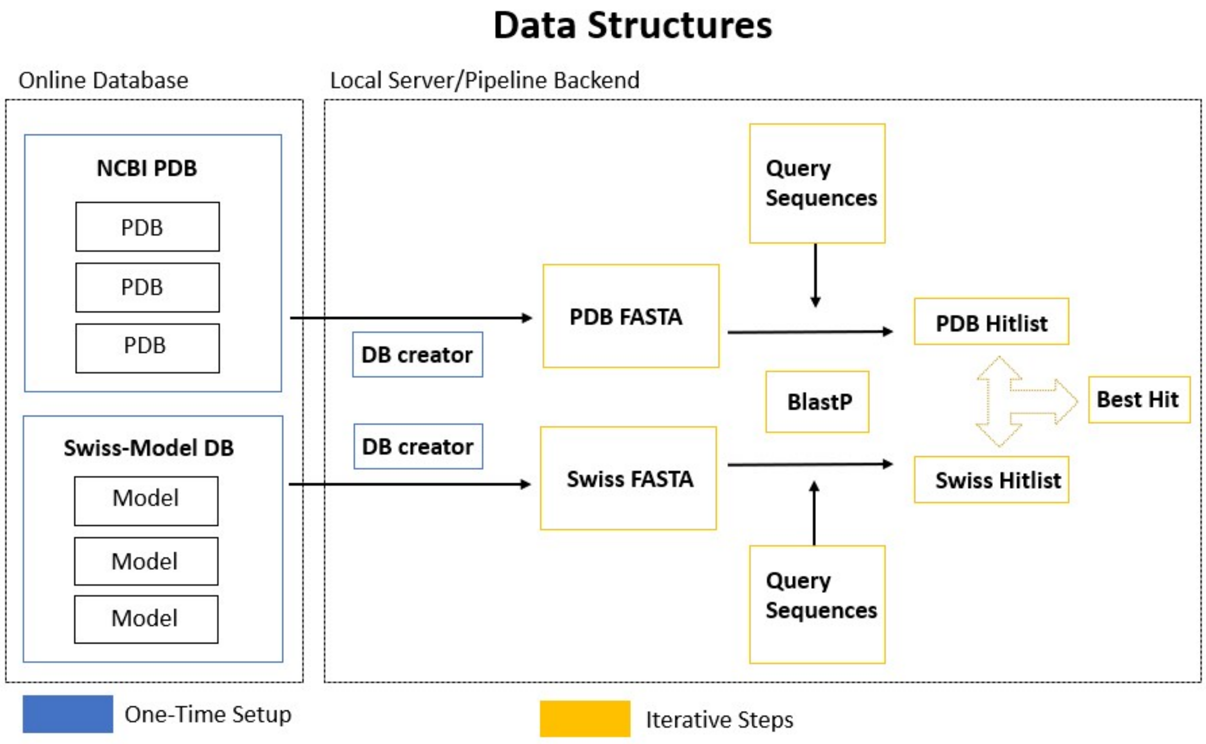
Mind the gap - linking crystal structures and sequences without misrepresentation in antibody research

Simon Schäfer[a,b], Thomas Winkler[a] and Heinrich Sticht[b]

[a]Chair of Genetics and [b]Division of Bioinformatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Protein crystal structure and model databases link structural with sequence information. Numerous crystal structures have unresolved regions (gaps) in their structural coordinates that are not represented in their corresponding sequences. This phenomenon is frequent in the C- and N-terminal regions as well as in solvent exposed loop structures [1]. In protein families with highly homologous regions and very variable loops structures, this mismatch between sequence and structure information is problematic for high throughput analyses.

In antibody structures, this problem affects the complementary determining regions. The NCBI Protein Database (PDB) entries for antibody heavy chains contain only 391 complete chains out of 586 chains in 122 PDB entries. To address the problem of possible misrepresentation for 195 chains containing gaps, we converted the PDB and the Swiss-Model Databases to a sequence data base containing only amino acids resolved in the crystal structures. The resulting database was queried with the BlastP algorithm and antibody reference sequences to find structures with high identity to the query sequences without the risk of obtaining unresolved structural elements [2]. The database utilizes the FASTA format and is compatible to sequence and structural analysis pipelines for any protein family. In addition, the database is offline deployable for closed scientific and clinical networks handling proprietary or patient data.



[1] K. Djinovic-Carugo; O. Carugo, (2015), *Intrinsically disordered proteins* 3 (1), e1095697. DOI: 10.1080/21690707.2015.1095697

[2] C. Camacho et al., (2009), *BMC bioinformatics* 10, p. 421. DOI: 10.1186/1471-2105-10-421