



HAL
open science

NAfragDB: A Multi-Purpose Structural Database of Nucleic-Acid/Protein Complexes for Advances Users

Antoine Moniot, Sjoerd de Vries, Dave Ritchie, Isaure Chauvot de Beauchêne

► To cite this version:

Antoine Moniot, Sjoerd de Vries, Dave Ritchie, Isaure Chauvot de Beauchêne. NAfragDB: A Multi-Purpose Structural Database of Nucleic-Acid/Protein Complexes for Advances Users. 21e congrès du Groupe de graphisme et modélisation moléculaire (GGMM), Apr 2019, Nice, France. hal-02393039

HAL Id: hal-02393039

<https://hal.science/hal-02393039v1>

Submitted on 1 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NAfragDB: a multi-purpose structural database of nucleic-acid – protein complexes for advanced users

Antoine Moniot ^{1,*} Sjoerd De Vries ^{2,*} Dave Ritchie ^{1,*}, Isaure Chauvot De Beauchene ^{1,*}

1 : Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) - Centre National de la Recherche Scientifique : UMR7503, Université de Lorraine, Institut National de Recherche en Informatique et en Automatique, Campus Scientifique BP 239 54506 Vandoeuvre-lès-Nancy Cedex - France

2 : Molecules Thérapeutiques in-Silico (MTi) - INSERM, Université Paris Diderot, Université Denis Diderot 35 rue Hélène Brion, case 7113, 75205 Paris Cedex 13 - France

Auteur correspondant: antoine.moniot@loria.fr, sjoerd.de-vries@inserm.fr, dave.ritchie@inria.fr, isaure@beauchene.fr

Many structural bioinformatics databases support automated searches via a limited number of pre-defined criteria and their combinations. Here, we present NAfragDB, a structural database of nucleic-acid (NA) - protein complexes that supports arbitrarily advanced queries and any combination thereof via python-written requests directly on the raw data.

We have extended our earlier pipeline for automated creation of NA fragment libraries to include many "low level" combinable information units. To build a database, we retrieve all protein-NA structures from the PDB, extract relevant data (resolution, NA type, etc), clean each structure (add missing atoms, list incomplete nucleotides, etc), and characterize the interface (sugar/phosphate/base - protein distances, water contacts, etc). We then use the 3DNA program [1] for NA structure description that gives exhaustive data in easily parsable Json format. Finally, we rearrange the data per nucleotide (eg. "nucleotides 5 to 15 make a stem-loop" → "nucleotide 5 is at position 1 in an 11-nucleotides stem-loop").

We provide a description at both the structural and nucleotide levels of the full set of protein-bound NAs from the PDB in a single Json file, and a tool to build customised queries. This allows the user to:

* create specific benchmarks for targeted purpose (ex. Retrieve all complexes with a stretch of 7 consecutive single-stranded nucleotides, of which at least 5 are within 4 Angstrom of the protein).

* compute various statistics (ex. Do single-stranded RNA establish H-bonds more preferentially via the base/sugar than via phosphates compared to double-stranded RNA?).

* build targeted fragment libraries

We will provide a live demo of building a database, applying search queries, and computing statistics. The source code to create, update, and search a database is available at <https://github.com/isaureCdB/NAfragDB.git>. We hope to encourage the sharing of such database capabilities to increase interoperability and ease complex use-cases by advanced users.

[1] X-J Lu & WK Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Research (2003) 31(17), 5108-21.

Mots-Clés : protein ; nucleic acid complexes ; nucleic acids structure ; structural database