

BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data

Supplementary Material

Adam Hospital^{1,2}, Pau Andrio³, Cesare Cugnasco^{3,4}, Laia Codo³, Yolanda Becerra^{3,4}, Pablo D. Dans^{1,2}, Federica Battistini^{1,2}, Jordi Torres^{3,4}, Ramón Goñi^{2,3}, Modesto Orozco^{1,2,3,5*}, Josep Ll. Gelpi^{2,3,5*}

¹Institute for Research in Biomedicine (IRB Barcelona). Baldiri Reixac 10-12, 08028 Barcelona, Spain.

²Joint BSC-IRB Research Program in Computational Biology. Baldiri Reixac 10-12, 08028 Barcelona, Spain.

³Barcelona Supercomputing Center. Jordi Girona 29, 08034 Barcelona, Spain.

⁴Dept. Computer Architecture, Technical University of Catalonia (UPC-BarcelonaTech), 08034 Barcelona, Spain.

⁵Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain.

* To whom correspondence should be addressed. Josep Ll. Gelpi Tel: 34934034009; Fax: 34934021559; Email: gelpi@ub.edu. Correspondence may also be addressed to Modesto Orozco Tel: 34934037155; Fax: 34034037175; Email: modesto.orozco@irbbarcelona.org

BIGNASim Supplementary Material can be also accessed at

<http://mmb.irbbarcelona.org/BIGNASim/SupMaterial/>

Table of contents:

1	Database structure	3
1.1	Cassandra. The trajectory subsystem.....	3
1.2	MongoDB. The analysis and metadata subsystem.....	4
1.3	Representation of molecular fragments	5
2	Data Ontology	6
3	Procedure and instruction to deposit new trajectories in BIGNASim	7
3.1	Submission requeriments	7
3.2	Submission procedure.....	8
4	Software Download	10
4.1	Hardware and software requirements	11
5	Examples of Use	11
5.1	Obtaining information about the Drew-Dickerson dodecamer	12
5.1.1.	Drew Dickerson Dodecamer (DDD)	12
5.1.2.	AT Base-Pair Step (central in DDD)	14
5.1.3.	Naked duplex B-DNA structure with a particular nucleotide fragment	17
5.2	Visualization of global analysis based in xCGy fragments.....	20
5.3	Retrieval, and downloading of xCGy Meta-trajectories	24
5.4	Analysis using fragments hierarchy. Correlation between CpG twist and ζ torsions. 27	
5.5	Combining Experimental and MD analysis.....	29
6	Supplementary Tables	40
6.1	Table S1. Analysis types available in BIGNASim	40
6.2	Table S2. Cassandra trajectory database structure.....	41
6.3	Table S3. Structure of main MongoDB collections.....	41
6.4	Table S4. Representative data structures stored in the MongoDB analysis subsystem.	42
6.5	Table S5. Complete structure of data objects showing hierarchical relationships derived from the central tetramer of the Drew-Dickerson dodecamer (A ⁵ ATT)	44
6.6	Table S6. BIGNASim ontology terms	46

6.7	Table S7. Deposition procedure	51
7	Supplementary Figures	53
7.1	Figure S1. BIGNASim portal screenshot. Expanded view of simulation summary.	53
7.2	Figure S2. Example of fragment definition on the analysis database.	54
8	References	55

1 Database structure

BIGNaSim is based on the combination of two database engines, Cassandra and MongoDB, and an adapted version of the analysis section of our Nucleic Acids MD portal, NAFlex (1). For trajectory data manipulation, the platform uses MDPlus, an in-house python library that integrates MDAAnalysis tools (2) with a developed Cassandra interface.

1.1 Cassandra. The trajectory subsystem.

Cassandra (3) is a distributed and highly scalable key-value database with a strong user community. Cassandra implements a non-centralized architecture, based on peer-to-peer communication, in which all nodes of the cluster are able to receive and serve queries. Data is stored in tables by rows, which are identified by a key chosen by the database user. In each row, users can add different attributes also identified by a chosen name. To each node of the cluster, a token is assigned and become responsible for hosting specific set of rows. The target node for each row is chosen through the partitioner algorithm and the decision is based on the row key and the node token. Cassandra also allows using compound keys and thus, more than one attributes to identify a row. In this case, the users have to specify the partition key (i.e. the attribute that will guide the node assignment) and the clustering keys. When a node receives a query, it finds out which node is responsible for each row involved in the query and partitions, and forwards appropriately the data request. This means that data modelling has a key influence on query performance (4). For this reason, users are encouraged to define their data models considering which queries they are going to perform. Moreover, a common practice is to replicate data in different data models to accommodate different queries. Recently, the authors proposed a mechanism to alleviate users from this requirement (5).

Table S2. Cassandra trajectory database structure.

Topology table (idSimulation)	Trajectory table (idSimulation)
atom_num (Partition Key) atom_name atom_type chain_code residue_code residue_num	frame (Partition Key) atom_id (Clustering Key) x y z (Box size data is included in the same frame as additional pseudo-atoms)

The Cassandra subsystem (Table S2) was organized in two tables: *Topology* holds the description of the molecular system using atom number as main indexing key, and storing the atom details, and the usual logical ways of grouping them (*residue*, *chain*). The *Trajectory* table stores the coordinates themselves indexed using frame and atom numbers. Cassandra is a distributed system, and the selection of the partition key has a strong influence on the retrieval efficiency. In our implementation, trajectory data is distributed using frame numbers, improving the retrieval of frame blocks. Indeed, by defining the frame number as Partition Key, we ensure that all the atomic coordinates at a given snapshot are stored contingently in the same node.

Additionally, each frame has atomic identifiers as a second level index, allowing efficient access to any subset of atoms. We have chosen to prioritize frame-based access, after analysing the pattern of access of the MDAnalysis software, used to handle trajectory data. MDAnalysis, constrained by its interface, always access to trajectory a frame at a time. Consequently, with our model the existing algorithms can access to a trajectory in Cassandra seamless, as if it was a common file. At the same time, algorithms that require data of only a subset of atoms may be optimized to take advantage of the second level indexing. To move trajectory data in and out of the Cassandra subsystem, the use of the Python package MDPlus assures a full compatibility with existing molecular dynamics software. Still, when dealing with massive bulk data loading into the database, the overhead introduced by the network communications and the data marshalling between different platforms can be a problem. For that reason, we developed a utility program that takes as input a trajectory file and converts it directly into SSTables, the Cassandra internal data format.

1.2 MongoDB. The analysis and metadata subsystem

The MongoDB database holds simulation metadata and pre-calculated analysis results. MongoDB is a fully flexibly engine and can store heterogeneous collections of documents. The internal structure of each document does not need to be defined beforehand and can match the data structure used in the interacting software, thus simplifying the use of database documents and external analysis software. MongoDB also allows to partition data among different servers (*data sharding*), using any of the fields as partition key. In our case, the data of the analysis requires both frame-based and atom-based access, hence we have chosen the complete document key as *sharding* key (See Table S3 below). Although MongoDB is configured with a single entry-point, it processes access queries in parallel among the available nodes, so maximum efficiency is achieved when data is spread evenly among them.

A condition to make the database usable is a very consistent indexing schema, which allows an easy recovery of such documents. Table S3 shows the database collection list together with the primary keys used to store the different objects. Table S4 shows representative data objects as stored in the DB.

Table S3. Structure of main MongoDB collections

MongoDB collection	Main Index components	Description
<i>simData</i>	idSim	Simulation metadata, following a specifically defined ontology
<i>analDefs</i>	idSim, idAnal	Analysis description, one document stored for every analysis result item available. Analysis available could differ from one simulation to another
<i>groupDef</i>	IdSim, (idGroup,nGroup)	Molecular groups (bases, base pairs, base-pair steps, molecular fragments) defined in the simulated system
<i>analData</i>	idSim, (idGroup,nGroup), nFrame (nFrame = 0: Averaged analysis data) (nGroup = 0: All system analysis)	Analysis results. The most appropriate data model for each analysis type is used.
<i>analBinFiles</i>	Id. Above	Binary files with pre-calculated analysis results (plots, images, etc.)

1.3 Representation of molecular fragments

MongoDB BIGNASim database has been populated using in-house scripts, and parsing the results obtained from the series of well-known software of analysis implemented in NAFlex (1). Definition of residues and standard groups (nucleotides, base-pair, and base-pair steps) are generated automatically from the simulated sequence and stored in the *groupDef* collection. Besides of predefined standard groups, the collection can store the definition of any relevant fragment of the simulated molecular system. As a representative example of *groupDef* structure, Figure S2 shows the complete hierarchy derived from the central tetramer of a Drew-Dickerson dodecamer.

The complete structure of such objects can be found in Table S5. Once the fragments are defined, their id (composed by *idSim*, *idGroup*, and *nGroup*, see Table S3) are used to index analysis results in *analData* and *analBinFiles* collections. As shown in Figure S2 above, and Table S5, the collection also holds a hierarchic relationships indicating which are the components of each fragment from the immediate lower level; this allows to navigate from any group down to its composing parts, and to the individual bases (see Use Case 4, for an example of such usage). This would allow linking together analysis corresponding to the related hierarchical levels. At the residue level, the MongoDB analysis subsystem is consistent with data hold in the trajectory subsystem (i.e. *idGroup* + *nGroup* corresponds to *residue_code* + *residue_num*). As shown in Table S2, results of the analyses are again stored in the three axes space: simulation, the analysed group (split in group id and sequence number for convenience), and frame number. In addition, averages along the trajectory and analysis spanning the whole system can be stored in the same structure. This layout will allow retrieving easily any set of results for any given set of groups and frames and performing the appropriate post-process. Although most data can be retrieved from the BIGNASim portal, more specialized data combination would require specific scripts. See Examples of Use below for sample codes of

such scripts using MongoDB JavaScript. Similar scripts can be prepared using other programming languages (Python, Perl, PHP, or Java). Analyses that may lead to non-numerical results (XY plots, 3D grids, etc.) are also stored under the same coordinate system, although they are kept in a separate collection (*analBinFiles*) for efficiency reasons. This database layout could be extended to any new type of analysis, without modification, after an appropriate mapping of each individual data item in the group/frame axes. Additionally, the GridFS system provided by MongoDB has been used to handle file based data transfers between application modules, and to hold the temporary user space used for downloading data.

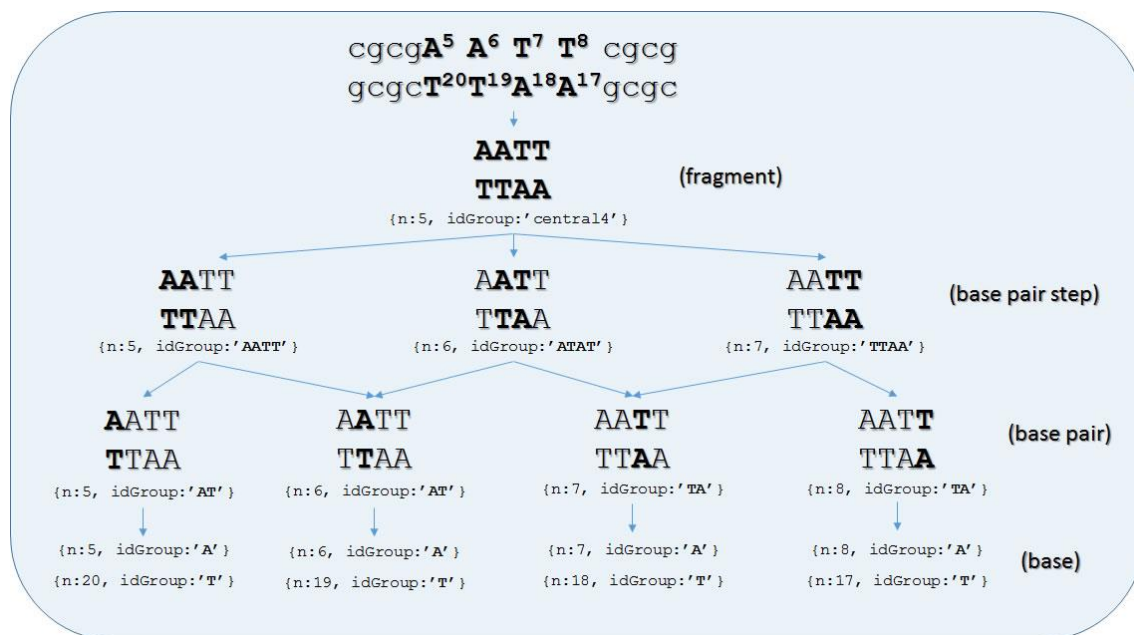


Figure S2. Example of fragment definition on the analysis database. Database entries in *groupDef* collection derived from the central tetramer of a Drew-Dickerson dodecamer. Primary keys of each data item are indicated. Arrows indicate a *container* relationship between data objects. Simulation id has been deleted from keys for simplicity.

2 Data Ontology

A strong requirement to organize any field of knowledge is the agreement in the terminology used. This has been a concern in Bioinformatics and has led to the development of a number of ontologies describing several aspects of the discipline (6-9). Here we have developed a partial data ontology to describe Nucleic Acids simulations. Its contents are used to qualify simulations, and power the search facility. Simulation browser tables, and simulation details (Figure S1B) include the set of keywords derived from the ontology. In its present state ontology represents merely a set of normalized terms, but it has been developed in a separate project and will be documented elsewhere. Table S6 show a categorized list of ontology terms.

3 Procedure and instruction to deposit new trajectories in BIGNASim

BIGNASim accepts submissions to incorporate new trajectory data to the repository. Submitted datasets should correspond ideally to finished studies, described by one or more journal publications. Datasets can consist in one or several trajectories. To maintain the consistency of the repository, uploaded trajectories will be limited to 5,000 frames. Authors however, should assure accessibility to the complete trajectories on-demand. Solvent should be removed but ions may remain when its presence is relevant to the study. Authors should assure that the trajectory is of high quality, fulfilling a specific quality checklist (see below), and provide enough metadata to allow to efficiently index the study within the database.

3.1 Submission requirements

The following requirements should be met for a simulated dataset be acceptable in BIGNASim

1. Dataset should be supported by a scientific publication. Publications in press or submitted could be acceptable, but data will be kept on-hold until the publication is publicly available, and eventually removed if this does not happens after a reasonable time. One or more datasets can be derived from a single publication, but all of them should be referred explicitly in the paper or Supplementary Material. In this sense, BIGNASim could be used as on-line Supp. Material for the publication, provide that enough processing time is allowed.
2. Dataset should correspond to Molecular Dynamics simulation trajectories of nucleic acids (DNA, RNA, PNA), alone or complexed with protein or small ligands.
3. Submitted trajectories should correspond to 5,000 frames, representing the complete simulation time. Solvent should be removed. Ions can be maintained if necessary to the aim of the study. Trajectories should be imaged and individual frames superimposed. Trajectory can be split in multiple files if necessary. Table S7 indicates the available formats.
4. Trajectories should be accompanied by a topology file in the appropriate format that should match the contents of the trajectory files. PDB files are acceptable as topologies. BIGNASim uses MDAnalysis (2) for handling trajectory and topology formats. Refer to MDAnalysis documentation for additional information (Table S7).
5. Submission should include a series of mandatory metadata items (Table S7), but we encourage providing metadata covering the entire ontology (Table S6).
6. Trajectories should fulfil the quality requirements shown in Table S7. Quality will be further checked as part of the analysis step performed in the database.

3.2 Submission procedure

Before starting the submission procedure, authors should prepare the following material:

- One or several trajectory files with a total length of 5000 frames. Chose the frequency of frames to cover the entire simulation. Solvent should be removed, and trajectories properly imaged and superimposed. Acceptable formats appear in Table S7.
- A topology file matching the trajectory file(s). Acceptable formats appear in Table S7
- A quality checklist (Table S7)
- A copy of relevant publications regarding the dataset, unless openly available

1. Register in BIGNASim. Indicate in the registry the intention of deposit new datasets, and follow the instructions.

2. Upload the necessary files (trajectories, and topology) to the workspace. Http upload is provided. For large datasets where transmission could be problematic, other procedures can be arranged.

3. Click on Initiate Deposition to open the Deposit form, and follow the instructions herein. Most Fields are mandatory. Alternatively, a file containing such information can be uploaded to the workspace. A template for this file can be obtained from the web site.

4. The form will allow indicating the uploaded files that correspond to the submission. As a submitted dataset could contain more than one simulation, topology and metadata should be provided for every simulation included. Facilities are provided to allow re-using metadata among trajectories.

5. Click on *Complete Submission* to start the process. After completing the submission, an accession number will be issued, and a status file will appear in the workspace. Data will be blocked and maintained on-hold until it has been processed. To modify eventually the contents of the submission, contact our team.

6. Trajectories will be checked and the set of BIGNASim analysis will be performed. A detailed log of the analysis performed will be available. The status file will be updated accordingly. In the eventual case that any of the analysis will fail due to issues in the incoming data, authors will be informed and asked to amend the problem. In the absence of errors, the dataset will be incorporated to de database and made public immediately or at the indicated date if any. Note that analysis procedure may take some time.

Example screenshots of the submission procedure

Step1 | File submission

Dataset - Group a collection of simulations into a dataset to better search and administrate your data

Dataset name (*)

Dataset description

Dataset publication (*) Release once submission is accepted
 Hold dataset until certain date

Publication/s (*) The dataset has not a reference publication yet. (Project documentation will be uploaded on the following section)
 Reference publication/s are the following

Files to submit - Select the files that will compose your submission. Only those files previously uploaded to the [workspace](#) can be here selected

Trajectory File/s (*)	<table><thead><tr><th>File Name</th><th>File Format</th></tr></thead><tbody><tr><td>Select from uploaded files ▼</td><td>Select file format ▼ (x)</td></tr></tbody></table> <p>[Add new]</p>	File Name	File Format	Select from uploaded files ▼	Select file format ▼ (x)
File Name	File Format				
Select from uploaded files ▼	Select file format ▼ (x)				
Topology File/s(*)	<table><thead><tr><th>File Name</th><th>File Format</th></tr></thead><tbody><tr><td>Select from uploaded files ▼</td><td>Select file format ▼ (x)</td></tr></tbody></table> <p>[Add new]</p>	File Name	File Format	Select from uploaded files ▼	Select file format ▼ (x)
File Name	File Format				
Select from uploaded files ▼	Select file format ▼ (x)				
Simulation metadata	<table><thead><tr><th>File Name</th><th>File Format</th></tr></thead><tbody><tr><td>Select from uploaded files ▼</td><td>Select file format ▼ (x)</td></tr></tbody></table>	File Name	File Format	Select from uploaded files ▼	Select file format ▼ (x)
File Name	File Format				
Select from uploaded files ▼	Select file format ▼ (x)				
Dataset documentation	<table><thead><tr><th>File Name</th><th>File Format</th></tr></thead><tbody><tr><td>Select from uploaded files ▼</td><td>Select file format ▼ (x)</td></tr></tbody></table> <p>[Add new]</p>	File Name	File Format	Select from uploaded files ▼	Select file format ▼ (x)
File Name	File Format				
Select from uploaded files ▼	Select file format ▼ (x)				

(*) Indicates the obligatory fields

Step 2 | Simulation Metadata [previous step]

General description:

PDB	<input type="text"/>
Ligands (or modified nucleotides)	<input type="text"/>
Additional Solvent	<input type="text"/>
Counter Ions	<input type="text"/>
Number of Frames	<input type="text"/>
Frame Step (ns)	<input type="text"/>
Comments	<input type="text"/>

Trajectory annotation: Check the ontological terms that better describe the uploaded trajectory

1.- System

NA_Type (+)

- DNA
- RNA (+)
- DNA-RNA_Hybrid
- PNA
- OtherNAType
- Not specified

Architecture (+)

System_Type (+)

OriginalHelicalConformation (+)

SequenceModifications (+)

SequenceFeatures (+)

Local Structures (+)

2.- Simulation

SimConditions (+)

TrajectoryType (+)

3.- Analysis

TimeScope (+)

FragmentScope (+)

AnalysisType (+)

>> Save metadata

Step3 | Confirmation [step 1] [step 2]

Submission process

Submission Identifier	BNS_1TRR_0000
Status	COMMITTED
Submission date	08 Oct 2015 - 09:28

Files attached to the submission

topology	1TRR_300K.top
trajectory	1TRR_300K-1_200ns.xtc
metadata	1TRR_300K-1_200ns.csv
documentation	manuscript.pdf

Submission successfully sent!

Your simulation data has been sent. During the following days the petition will be validated. Return to your personal [workspace](#) to keep track of the status of all your submissions.

4 Software Download

BIGNASim is composed of a rather large and complex set of software units, besides of requiring the installation of two NoSQL database platforms. The installation of local copy of the

server is not generally recommended. However, in cases where data is sensitive, and must be kept private, or for testing purposes an installable reduced version of the software is available. Local stand-alone installations will not be connected with the main BIGNASim database, unless specific agreement in such direction is achieved. Downloadable software will be tuned to avoid the need of an external database system, although they will have the full functionality of the analysis portal. This would obviously limit the amount of data to be handled, but offers a system capable enough to fulfil the needs of a simulations group to a certain extent. Due to the complexity of the software structure, the installable package is being prepared in prepacked Docker images.

4.1 Hardware and software requirements

Local BIGNASim could be installed in any modern Linux system (it has been tested on Ubuntu, and openSUSE). Note that the installation procedure requires reasonable skills in system administration, and root privileges. The downloadable system is designed to run in a single computer or inside a virtual machine.

1. To run BIGNASim Docker images, a Docker server is required. Instructions to install Docker can be found at Docker web site (<http://www.docker.com>). Note that Docker requires root privileges to be installed. Docker images can be run as part of a virtualized system, refer to Docker instructions to choose the appropriate configuration. Docker should be installed in all nodes performing analysis, but it is not required in database nodes.
3. Download installation package from BIGNASim site. Use the information provided to download Docker images, and perform the basic installation. A series of launching scripts will be created.
4. Install script will ask for data directories for raw trajectory data, and analysis results. These directories should be accessible locally or NFS mounted. Raw trajectory data could be removed after the analysis is performed, but analysis results should be a permanent storage.
5. Follow README files for detailed instructions to install trajectory data.

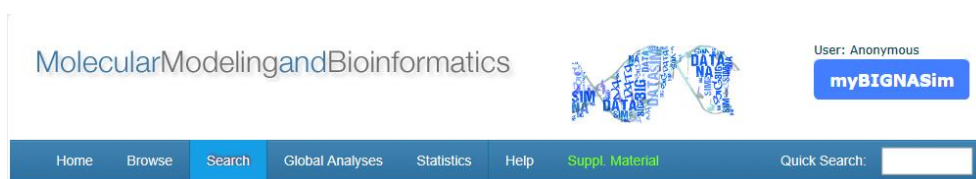
5 Examples of Use

BIGNASim has been designed in a way that advanced data analysis can be performed straightforwardly, even for types of analysis that were not known at the time of the design. The following cases show some representative examples of the possibilities of data manipulation. Most of the examples can be done within the portal, and are already available as tutorials in BIGNASim web site. The power of the BIGNASim is to store raw data in structures that allow an

easy recovery and combination, either from the analysis or from the trajectory databases. Simple scripts can be used to query the databases and obtain the required analyses. Scripts examples shown here are written in Javascript, the natural language for MongoDB, but most popular scripting languages can be also used. Additional information can be found at [BIGNASim Help site](#)

5.1 Obtaining information about the Drew-Dickerson dodecamer

This use case shows several examples on using the BIGNASim search engine to locate a particular nucleic acid sequence, nucleic acid fragment or base pair step. The search section of the portal is accessed through the main menu:



The **search section** contains three different possibilities:

- Search by **sequence or specific sequence fragments** (using regular expressions)
- Search by **specific base-pair-steps** (with or without flanking regions)
- Search by an extensive **nucleic acids ontology** (see below)

Examples will show the way of finding three different types of information from the database:

- Information related to the well-known Drew-Dickerson Dodecamer (DDD)
- Information related to the AT Base-Pair Step (central in DDD)
- Information related to a naked duplex B-DNA structure with a particular nucleotidic fragment

5.1.1 Drew Dickerson Dodecamer (DDD)

In this case, the nucleotide sequence of the DDD (CGCGAATTTCGCG) can be just searched using the "Search by Sequence" section of the portal:

Search by Sequence

Sequence (Regular Expressions Allowed)
Ex: GG[GTAC]GG

1naj

CGCGAATTCGCG

Javascript code examples

```
// Finding DDD simulations
SimulationList = db.simData.find({'sequence' : 'CGCGAATTCGCG'}).toArray()
// Finding simulations containing DDD sequence using regular expressions
SimulationList = db.simData.find({'sequence' : /CGCGAATTCGCG/}).toArray()
// Finding simulations containing DDD sequence using possible variations
SimulationList = db.simData.find(
  {'sequence' : /^CGCGA[AC]TTCGCG$/}
).toArray()
```

Due to the importance of DDD in the field, it is specifically included in the *Sequence Features* section of the nucleic acids ontology, and can be located directly:

Search by Ontology

Ontology Search: Results found: (9)

- > Nucleic Acid Type
- > Structure
- > System Type
- > Trajectory Type
- > Original Helical Conformation
- > Sequence Modifications
- > Sequence Features (Click to Collapse)
 - Any
 - Poly Adenine Track (0)
 - Poly Guanine Track (0)
 - Drew Dickerson Dodecamer (9)
 - Sequence Mismatch (0)
 - Other
- > Local Structures
- > Simulation Conditions

Javascript example code

```
// Finding DDD simulation from ontology search
SimulationList = db.simData.find({'ontology' : '10603'}).toArray()
```

Both access ways open a **browse page** showing the simulations stored in the database for this particular sequence. In this case, 5 different trajectories are found, the longest one having

10 μ s. Each of the simulations can be open individually to look at the MD simulation metadata and trajectory analyses. Combined information from more than one simulation can be obtained, by selecting the desired entries and clicking at **Open analyses for selected simulations**.

Click on Id to open Simulation Metadata and Analyses

Show 10 entries Search:

<input type="checkbox"/>	Id.	PDB	Type	SubType	ForceField	Solvent	Description	Time (ns)	Sequence
<input type="checkbox"/>	<input type="text" value="id"/>	<input type="text" value="Sel"/>	<input type="text" value="Sel"/>	<input type="text" value="Select"/>	<input type="text" value="Select"/>	<input type="text" value="Select"/>	<input type="text" value="Description"/>	<input type="text" value="Time (ns)"/>	<input type="text" value="Sequence"/>
<input type="checkbox"/>	NAFlex_DDD_bsc1	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC0 TIP3P Electroneutral	10,000	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_DDD_II	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	1,200	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_DDD_II_1	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	1,000	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_DDD_II_2	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	1,000	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_DDD_800ns	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC0 TIP3P Electroneutral	800	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_DDD_GMX_GPU	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	100	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_DDD_GMX_CPU	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	100	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_DDD_Amber_GPU	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD Parm99 TIP3P Electroneutral	100	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_DDD_Amber_CPU	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD Parm99 TIP3P Electroneutral	100	CGCGAATTCGC G

Showing 1 to 9 of 9 entries Previous Next

5.1.2 AT Base-Pair Step (central in DDD)

The central base-pair step of DDD (CGCGA**AT**TCGCG) can be obtained from the **Search by**

Search by Base Pair Step

Base Pair Step / Flanking Region (Number of bases)

AT


Base Pair Step section:

Javascript equivalent code

```
// Finding simulations containing AT BpStep with 2 flanking bases
SimulationList = db.simData.find(
  {'sequence': /..AT../},
  {'_id': 1}
).toArray();
// Finding simulations containing AT BpStep on any strand
// (not needed for ApT due to symmetry)
SimulationList = db.simData.find(
  { $or: [ {'sequence': /..AT../}, {'rev-sequence': /..AT../} ] },
```

```
{_id:1}
).toArray();
```

In the selector, the desired base-pair step, in this case AT, must be chosen. There is also the possibility to add a number of required **flanking nucleotides**, to ensure that information obtained will not be from base-pair steps placed at terminal regions, which can show distorted flexibility parameters. In this example, two flanking nucleotides are forced. The same procedure can be applied for any base-pair step.

Click on Id to open Simulation Metadata and Analyses  Retrieve MetaTrajectory

Show 10 entries Search:

<input type="checkbox"/>	Id.	PDB	Type	SubType	ForceField	Solvent	Description	Time (ns)	Sequence
<input type="checkbox"/>	<input type="text" value="id"/>	<input type="text" value="Sel"/>	<input type="text" value="Sel"/>	<input type="text" value="Select"/>	<input type="text" value="Select"/>	<input type="text" value="Select"/>	<input type="text" value="Description"/>	<input type="text" value="Time (ns)"/>	<input type="text" value="Sequence"/>
<input type="checkbox"/>	NAFlex_DDD_bsc1	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	10,000	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_D05M	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P AddedSalt	3,000	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_DTipDang015M	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P AddedSalt	2,000	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_BSC1-BSC0	1NAJ	Dna	B	parmBSC0	TIP3P	DNA-B Duplex Naked TIP3P Electroneutral	1,500	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_BSC1-OL4	1NAJ	Dna	B	ParmBSC0-OL4	TIP3P	DNA-B Duplex Naked OL4 TIP3P Electroneutral	1,500	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_BSC1-OL1	1NAJ	Dna	B	ParmBSC0-OL1	TIP3P	DNA-B Duplex Naked OL1 TIP3P Electroneutral	1,500	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_BSC1-OL1-OL4	1NAJ	Dna	B	ParmBSC0-OL1-OL4	TIP3P	DNA-B Duplex Naked OL1+OL4 TIP3P Electroneutral	1,500	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_BSC1-Garcia	1NAJ	Dna	B	ParmBSC0-CG	TIP3P	DNA-B Duplex Naked Cheng-Garcia TIP3P Electroneutral	1,500	CGCGAATTCGC G
<input type="checkbox"/>	NAFlex_BSC1-C36	1NAJ	Dna	B	Charmm36	TIP3P	DNA-B Duplex Naked Charmm36 TIP3P Electroneutral	1,500	CGCGAATTCGC GCGCGAATTCGC CG
<input type="checkbox"/>	NAFlex_DDD_II	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	1,200	CGCGAATTCGC G

Showing 1 to 10 of 65 entries Previous 1 2 3 4 5 6 7 Next

Open Analyses for selected simulations Mark all Unmark all Reset

The results obtained for the **AT** base-pair step with the current content of the database are 51 different simulations. Looking at the sequence column, the interesting **AT pair**, together with the flanking region, can be easily identified thanks to the marking in yellow and orange colours, respectively. The first thing that can be seen in the browse page is that the database contains simulated systems different from **DDD** containing also the **AT base pair**. Specifically 46 sequences, some of them having more than one occurrence of it, are recovered. That offers enough information to compare between the flexibility parameters obtained for just the 5 sequences of **DDD** obtained in the previous section of this example with the remaining sequences having the **AT base pair**. To exclude DDD simulations from the recovered analyses, **selector of records shown** should be set to the maximum (100 records): 1) select all

simulations by using the checkbox placed at the left part of the table header, next to the **Id. title**,
 2) Filter results using DDD in the Search box, and uncheck all, 3) Remove the DDD filter.

Click on Id to open Simulation Metadata and Analyses

Retrieve MetaTrajectory

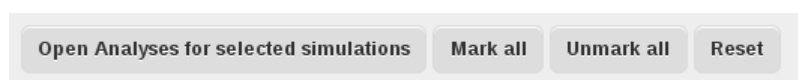
Show 10 entries Search:

<input type="checkbox"/>	Id.	PDB	Type	SubType	ForceField	Solvent	Description	Time (ns)	Sequence
<input type="checkbox"/>	id.	Sel	Sel	Select	Select	Select	Description	Time (ns)	Sequence
<input type="checkbox"/>	NAFlex_DDD_bsc1	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC0 TIP3P Electroneutral	10,000	CGCGAATTCGC
<input checked="" type="checkbox"/>	NAFlex_D05M	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P AddedSalt	3,000	CGCGAATTCGC
<input checked="" type="checkbox"/>	NAFlex_DTipDang015M	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P AddedSalt	2,000	CGCGAATTCGC
<input checked="" type="checkbox"/>	NAFlex_BSC1-BSC0	1NAJ	Dna	B	parmBSC0	TIP3P	DNA-B Duplex Naked TIP3P Electroneutral	1,500	CGCGAATTCGC
<input checked="" type="checkbox"/>	NAFlex_BSC1-OL4	1NAJ	Dna	B	ParmBSC0-OL4	TIP3P	DNA-B Duplex Naked OL4 TIP3P Electroneutral	1,500	CGCGAATTCGC
<input checked="" type="checkbox"/>	NAFlex_BSC1-OL1	1NAJ	Dna	B	ParmBSC0-OL1	TIP3P	DNA-B Duplex Naked OL1 TIP3P Electroneutral	1,500	CGCGAATTCGC
<input checked="" type="checkbox"/>	NAFlex_BSC1-OL1-OL4	1NAJ	Dna	B	ParmBSC0-OL1-OL4	TIP3P	DNA-B Duplex Naked OL1+OL4 TIP3P Electroneutral	1,500	CGCGAATTCGC
<input checked="" type="checkbox"/>	NAFlex_BSC1-Garcia	1NAJ	Dna	B	ParmBSC0-CG	TIP3P	DNA-B Duplex Naked Cheng-Garcia TIP3P Electroneutral	1,500	CGCGAATTCGC
<input checked="" type="checkbox"/>	NAFlex_BSC1-C36	1NAJ	Dna	B	Charmm36	TIP3P	DNA-B Duplex Naked Charmm36 TIP3P Electroneutral	1,500	CGCGAATTCGC
<input type="checkbox"/>	NAFlex_DDD_II	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	1,200	CGCGAATTCGC

Showing 1 to 10 of 65 entries Previous 1 2 3 4 5 6 7 Next

Open Analyses for selected simulations Mark all Unmark all Reset

The final step consists on clicking at the **Open Analyses for selected simulations** button at the bottom of the **browse page**, which will lead to the **analysis section** of **BIGNASim**.



In this section, the **AT base-pair step** button will open the available analyses for the AT bp-step. In order to compare the results with the ones corresponding to the **AT base-pairs** from just **DDD sequences**, the procedure can be repeated for these particular simulations, using an additional browser window.



Continuation Javascript code to retrieve analysis data for ApT steps

```
(...)
idSim = SimulationList[i]. id.idSim;
// retrieve the position of the AT bpSteps (stored as class:'ATAT')
ATPos = db.groupDef.find(
  {'_id.idSim': idSim, 'class': 'ATAT'},
  { id: 1}
).toArray()
(...)
// obtain available data for a given group
dataCur = db.analData.find(
  {'_id.idSim': idSim, '_id.nGroup': ATPos[i]._id.n,
  '_id.idGroup': ATPos[j]._id.idGroup}
);
while (dataCur.hasNext()) {
  printjson(dataCur.next());
}
```

5.1.3 Naked duplex B-DNA structure with a particular nucleotide fragment

The third example shows a more specific search: trajectories having the **DDD** central tetramer (**AATT**), computed on **naked B-DNA duplex** structures, simulated in **equilibrium** conditions, and **electroneutral** charge schema. **AATT** sequence should be included in the **Search by Sequence** section; and then the search refined using the **Search by Ontology**

section.

In the **Search by Ontology** section, search can be refined using keywords organized in a series of groups. In this case, the keywords chosen should be **DNA** in **Nucleic Acid Type** area, **Duplex** in **Structure** area, **Naked** in **System Type** area, **Equilibrium** in **Trajectory Type**, **B** in **Helical Conformation**, and, finally, **Electroneutral** in **Simulation Conditions**, **Ionic Concentration**. Every time a search parameter is chosen, the **search engine** computes the number of results stored for the current selected refinement specification and shows it on-the-fly in the top right part of the **Search by Ontology** section.

Search by Ontology

Ontology Search: Results found: (20)

> Nucleic Acid Type (Click to Collapse)

Any DNA (139) RNA (14) DNA-RNA Hybrid (1) Other

> Structure (Click to Collapse)

Any Lineal (Single Strand) (0) Duplex (138) Triplex (4) Quadruplex (5) Holliday Junction (0)

3-Way Junction (0) RNA Pseudoknot (0) Ribozymes (0) Large Ribosomal RNA (0) Riboswitch (0)

tRNA (0) G1introns (0) G2introns (0) RNA nanostructures (0) Other

> System Type (Click to Collapse)

Any Naked (144) Complex: Protein-Nuc (6) Complex: LigandNuc (0) Other

> Trajectory Type (Click to Collapse)

Any Equilibrium (154) Un/Folding (1) Transition (0) Other

> Original Helical Conformation (Click to Collapse)

Any A (16) B (118) Z (3) Hoogsteen (2) Mixed (5) Other

> Sequence Modifications

> Sequence Features

> Local Structures

> Simulation Conditions (Click to Collapse)

Select Force Field

Select Simulation Length

Select Temperature

Select Solvent Type

Electroneutral (80)

Select Ions Parameters

Javascript equivalent code

```
// Finding simulations containing TTAA fragment with 2 flanking bases
// TTAA is palindromic, only one strand need to be considered
// ontology tags: 'DNA' (10101), 'Duplex' (10202),
//               'Naked' (10301), 'B' (10402)
//               'Equilibrium' (20201), 'B', 'Electroneutral' (2010501)
// further check on subclasses has been eliminated for clarity
SimulationList = db.simData.find(
  {
    'sequence': /.TTAA./,
    'ontology': {$all:['10101','10202','10301','2010501','20201']}
  },{_id:1}).toArray();
```

Again, the results are shown in a **browse page**. Descriptions show the keywords assigned to each simulation, and confirm that results are indeed **duplex naked B-DNA structure** simulations, as defined in the **search**. Still, results obtained contain a sequence different than the **DDD** having the AATT tetramer: 1rvh (GCAAATTTTGC). For the rest of **DDD** trajectories, the differences rely on the particular simulation parameters used in the MD, e.g. solvent type, ionic parameters or total length.

Click on Id to open Simulation Metadata and Analyses Retrieve MetaTrajectory

Show entries Search:

Id.	PDB	Type	SubType	ForceField	Solvent	Description	Time (ns)	Sequence	
<input type="checkbox"/>	<input type="text" value="id"/>	<input type="text" value="Sel"/>	<input type="text" value="Sel"/>	<input type="text" value="Select"/>	<input type="text" value="Select"/>	<input type="text" value="Description"/>	<input type="text" value="Time (ns)"/>	<input type="text" value="Sequence"/>	
<input type="checkbox"/>	NAFlex_DDD_bsc1	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC0 TIP3P Electroneutral	10,000	CGCGAATTGCGG
<input type="checkbox"/>	NAFlex_BSC1-BSC0	1NAJ	Dna	B	parmBSC0	TIP3P	DNA-B Duplex Naked TIP3P Electroneutral	1,500	CGCGAATTGCGG
<input type="checkbox"/>	NAFlex_BSC1-OL4	1NAJ	Dna	B	ParmBSC0-OL4	TIP3P	DNA-B Duplex Naked OL4 TIP3P Electroneutral	1,500	CGCGAATTGCGG
<input type="checkbox"/>	NAFlex_BSC1-OL1	1NAJ	Dna	B	ParmBSC0-OL1	TIP3P	DNA-B Duplex Naked OL1 TIP3P Electroneutral	1,500	CGCGAATTGCGG
<input type="checkbox"/>	NAFlex_BSC1-OL1-OL4	1NAJ	Dna	B	ParmBSC0-OL1-OL4	TIP3P	DNA-B Duplex Naked OL1+OL4 TIP3P Electroneutral	1,500	CGCGAATTGCGG
<input type="checkbox"/>	NAFlex_BSC1-Garcia	1NAJ	Dna	B	ParmBSC0-CG	TIP3P	DNA-B Duplex Naked Cheng-Garcia TIP3P Electroneutral	1,500	CGCGAATTGCGG
<input type="checkbox"/>	NAFlex_BSC1-C36	1NAJ	Dna	B	Charmm36	TIP3P	DNA-B Duplex Naked Charmm36 TIP3P Electroneutral	1,500	CGCGAATTGCGG CGCGAATTGCGG
<input type="checkbox"/>	NAFlex_DDD_II	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	1,200	CGCGAATTGCGG
<input type="checkbox"/>	NAFlex_DDD_II_1	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	1,000	CGCGAATTGCGG
<input type="checkbox"/>	NAFlex_DDD_II_2	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC1 TIP3P Electroneutral	1,000	CGCGAATTGCGG

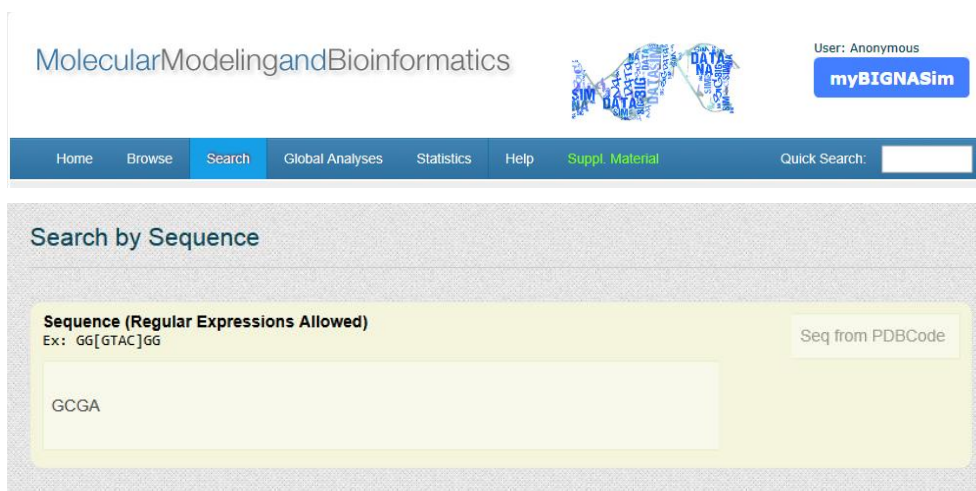
Showing 1 to 10 of 20 entries Previous Next

From the list of simulations, flexibility analyses can be obtained independently, or combined, clicking at **Open Analysis for selected simulations** button, or a **meta-trajectory** with this particular nucleotide fragment can be generated, joining atomic coordinates of the selected set

of simulations. More details on how to build a **meta-trajectory** with **BIGNASim** can be found at the section 5.3.

5.2 Visualization of global analysis based in xCGy fragments

The example shows the procedure to extract information from the **Global Analyses** section of **BIGNASim** portal. The simple study shown here can be extended to a real use case: the importance of flanking nucleotides in the flexibility of base-pair steps. The effect of the **tetranucleotide environment** in the sequence-dependent polymorphism of particular base-pair steps has been the target of recent studies. The **CG base pair step** used for example, shows an interesting bimodal behaviour in one of the six helical base-pair step parameters: **Twist** (10). In the study, authors claim that the effect of the flanking bases in the **CG base pair step** is crucial for the existence of two different conformers: **High Twist** (HT: $\sim 40^\circ$) and **Low Twist** (LT: $\sim 20^\circ$). Behaviours for each of the 16 possible tetramers including **CG** are reported. To illustrate the power of **BIGNASim** database and its interface, two analyses have been chosen: **ACGC** showing almost no bimodality, and **GCGA** showing a clear bimodality. The first step uses the **Search section** of the portal → Search by sequence (GCGA).



MolecularModelingandBioinformatics

User: Anonymous

myBIGNASim

Home Browse Search Global Analyses Statistics Help Suppl. Material Quick Search:

Search by Sequence

Sequence (Regular Expressions Allowed)
Ex: GG[GTAC]GG

Seq from PDBCode

GCGA

Javascript equivalent code

```
// Direct search of GCGA fragments on both strands
SimulationList = db.simData.find(
  {$or: [{'sequence':/GCGA/}, {'rev-sequence': /GCGA/}]}
);
// Alternatively search both on only one strand using complementarity
SimulationList = db.simData.find({'sequence':/(GCGA|TCGC)/});
```

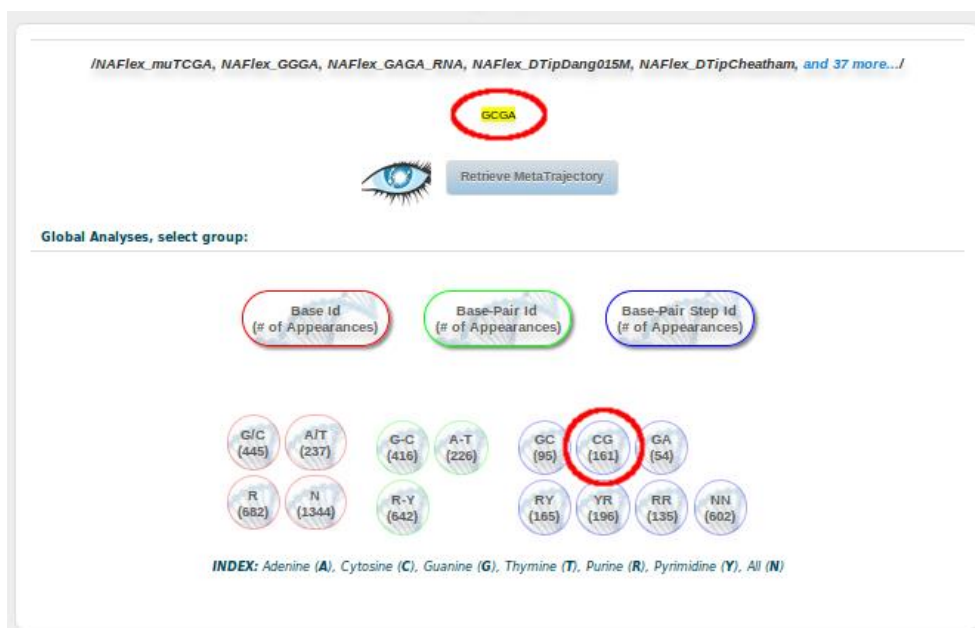
In this case, more than 40 simulations containing this particular fragment are available for selection.

Show 10 entries

Search:

Id.	PDB	Type	SubType	ForceField	Solvent	Description	Time (ns)	Sequence	
<input type="checkbox"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	
<input checked="" type="checkbox"/>	NAFlex_DDD_bsc1	1NAJ	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked DDD ParmBSC0 TIP3P Electroneutral	10,000	CGCGAATTCGCG
<input checked="" type="checkbox"/>	NAFlex_D05M	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P AddedSalt	3,000	CGCGAATTCGCG
<input checked="" type="checkbox"/>	NAFlex_DTipDang015M	1BNA	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P AddedSalt	2,000	CGCGAATTCGCG
<input checked="" type="checkbox"/>	NAFlex_BSC1-BSC0	1NAJ	Dna	B	parmBSC0	TIP3P	DNA-B Duplex Naked TIP3P Electroneutral	1,500	CGCGAATTCGCG
<input checked="" type="checkbox"/>	NAFlex_BSC1-OL4	1NAJ	Dna	B	ParmBSC0-OL4	TIP3P	DNA-B Duplex Naked OL4 TIP3P Electroneutral	1,500	CGCGAATTCGCG
<input checked="" type="checkbox"/>	NAFlex_BSC1-OL1	1NAJ	Dna	B	ParmBSC0-OL1	TIP3P	DNA-B Duplex Naked OL1 TIP3P Electroneutral	1,500	CGCGAATTCGCG
<input checked="" type="checkbox"/>	NAFlex_BSC1-OL1-OL4	1NAJ	Dna	B	ParmBSC0-OL1-OL4	TIP3P	DNA-B Duplex Naked OL1+OL4 TIP3P Electroneutral	1,500	CGCGAATTCGCG

Retrieve Analysis for the selected simulations at the bottom of the page, leads to a **Global Analyses** page, showing the results for the particular **CGGA** fragment. Since the interest is studying the possible bimodality showed by the **CG** base pair step in its **Twist** parameter when it is surrounded by G and A (**CGGA**), the **CG** button should be selected:

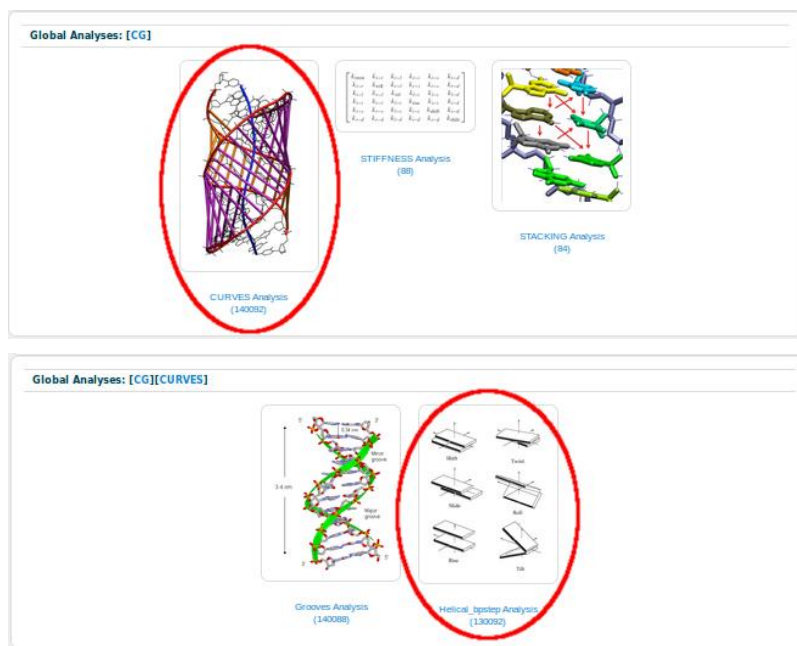


Javascript code hint

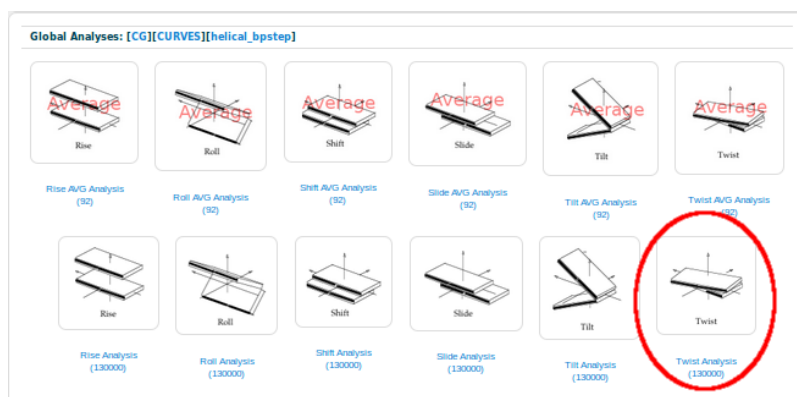
```
// Available data for all CpG bpstep (in any simulation, and any sequence
// position) can be retrieved at using just its idGroup: CGCG
DataforAllCpG = db.analData.find( {'_id.idGroup' : 'CGCG' } );

// For a simulation SIM and position POS
Datafor1CpG = db.analData.find( {'_id.idSim': SIM, '_id.nGroup': POS,
'_id.idGroup' : 'CGCG' } );
```

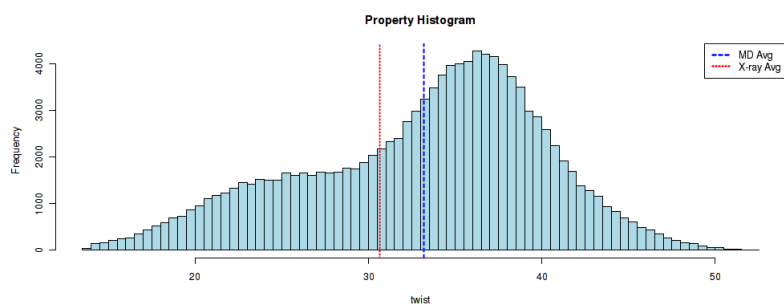
Twist data can be obtained from “Curves → Helical_bpstep”



BIGNASim in its current version contains two kind of analysis for each of the **six helical base-pair step parameters**: one with the values for every snapshot of all the selected simulations, and one with the **time-averaged** values for each simulation. To show the bimodality, histogram with all the values for the **Twist** parameter should be chosen:



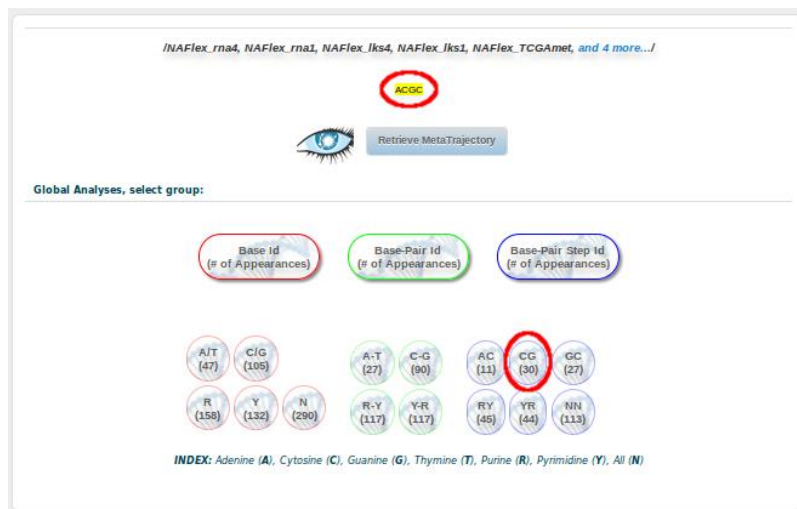
In the histogram plot, the average is represented as a vertical blue line, and the experimental value, used as reference, is represented as a vertical red line (see Example of use 5 for a detailed description of the use of experimental data).



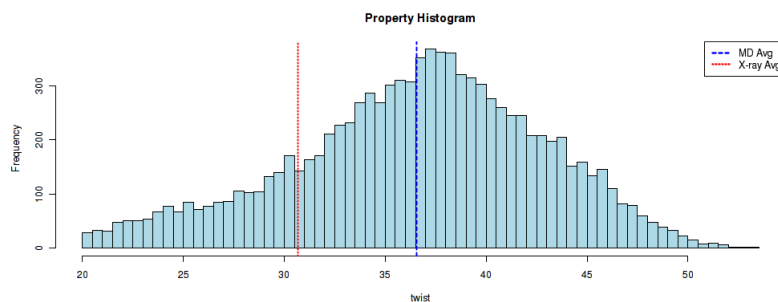
Javascript code

```
// Code to retrieve twist values for a complete trajectory for a given Simulation "SIM",
// and CpG at position "POS"
twistData c = db.analData.find(
  {'id.idSim': SIM, 'id.nGroup': POS, 'id.idGroup': 'CGCG'})
).sort('_id.frame':1);
while (twistData_c.hasNext() {
  Data = twistData_c.next();
  printjson (Data.id.frame + ' ' + Data.CURVES.helical_bpstep.twist);
}
```

The histogram shows two well defined populations, centred at $\sim 25^\circ$ and $\sim 35^\circ$, in good agreement with the previously presented study (10). To analyse the influence of the surrounding bases in the **CG** base pair step (tetramer influence), the procedure will be repeated seeking for the fragment **ACGC**. (*Search by Sequence* \rightarrow *Select All* \rightarrow *Open Analysis for the selected simulations*).



And CG \rightarrow Curves \rightarrow Helical_bpstep Analysis \rightarrow Twist Analysis.



The new histogram do seems to follow a **normal distribution**, although a small shoulder to the **low twist** conformation can still be identified. The clear difference between the two plots show that the GCGA tetramer shows a clear bimodality, whereas the **ACGC** tetramer is more inclined to be in a **High Twist** conformation. Additionally, raw histogram data can be downloaded for further analysis.

Complete Javascript code to retrieve twist data from ACGC tetramers

```
SEQ = 'ACGC';
RSEQ = 'GCGT';
// Search for simulations bearing ACGC
SimulationList = db.simData.find(
  {$or: [
    {'sequence': {$regex: SEQ}},
    {'rev-sequence': {$regex: RSEQ}},
    {'id':1}
  ]}
).toArray();
// search for CpG fragments
FragmentsList = db.groupDef.find(
  {'_id.idSim': {$in: SimulationList}, 'class': 'CGCG'}
).toArray()
// Iterate over fragments
for (i=0; i < FragmentsList.length; i++) {
  twistData_c = db.analData.find(
    {'_id.idSim': FragmentsList[i]._id.idSim,
     'id.nGroup': FragmentsList[i].id.n,
     'id.idGroup' : 'CGCG'}
  ).sort('id.frame':1);
  while (twistData_c.hasNext() {
    Data = twistData_c.next();
    printjson (Data._id.frame + ' ' + Data.CURVES.helical_bpstep.twist);
  }
}
```

5.3 Retrieval, and downloading of xCGy Meta-trajectories

The example show the procedure to build a meta-trajectory containing a particular nucleotide fragment using the **BIGNASim** portal. Example 3.2 has shown how to extract directly from the database the distribution of a particular helical parameter, recovering the **Twist bimodality**. However, one can be interested in study different properties, not pre-calculated in **BIGNASim** database. For that, there is the possibility to generate, download, or send to the **NAFlex server** a meta-trajectory containing just the nucleotide fragment of interest. It is build joining together the cartesian coordinates from a set of selected simulations stored in our database enclosing the fragment. To illustrate the power of **BIGNASim** database, the DB is searched for the **16 possible tetramers** including **CG** base-pair, going from 2 (**TCGG**) to 49 (**TCGC**) occurrences (see table below), in the present database release.

Tetramer	#Occurrences	Tetramer	#Occurrences
ACGA	5	ACGG	17
CCGA	6	CCGG	7
GCGA	43	GCGG	15
TCGA	9	TCGG	2
ACGC	10	ACGT	4
CCGC	6	CCGT	11
GCGC	19	GCGT	5
TCGC	49	TCGT	3

As a quick example, we show here the generation of a **meta-trajectory** of one of the tetramers having less occurrences (**ACGT**, 4 occurrences). The same procedure could be used to generate meta-trajectories for all of the possible tetranucleotides including **CG**, thus obtaining

a valuable set of ensembles to analyse and compare. The first step of the procedure consists on accessing the **search section** of the portal from the main menu (see Example 3.1), to locate simulations of normal duplex B-DNA systems containing ACGT.



MolecularModelingandBioinformatics

User: Anonymous

myBIGNASim

Home Browse Search Global Analyses Statistics Help Suppl. Material Quick Search:

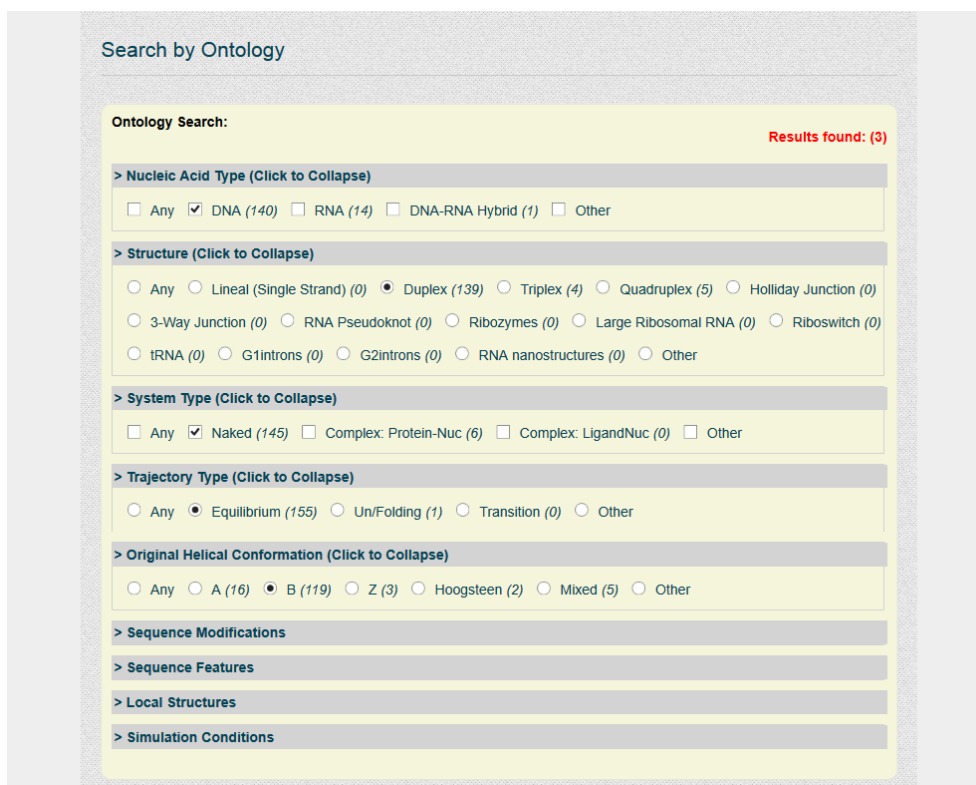


Search by Sequence

Sequence (Regular Expressions Allowed)
Ex: GG[GTAC]GG

Seq from PDBCode

ACGT



Search by Ontology

Ontology Search: Results found: (3)

> Nucleic Acid Type (Click to Collapse)

Any DNA (140) RNA (14) DNA-RNA Hybrid (1) Other

> Structure (Click to Collapse)

Any Linear (Single Strand) (0) Duplex (139) Triplex (4) Quadruplex (5) Holliday Junction (0)

3-Way Junction (0) RNA Pseudoknot (0) Ribozymes (0) Large Ribosomal RNA (0) Riboswitch (0)

tRNA (0) G1introns (0) G2introns (0) RNA nanostructures (0) Other

> System Type (Click to Collapse)

Any Naked (145) Complex: Protein-Nuc (6) Complex: LigandNuc (0) Other

> Trajectory Type (Click to Collapse)

Any Equilibrium (155) Un/Folding (1) Transition (0) Other

> Original Helical Conformation (Click to Collapse)


Any A (16) B (119) Z (3) Hoogsteen (2) Mixed (5) Other

> Sequence Modifications

> Sequence Features

> Local Structures

> Simulation Conditions

Click on Id to open Simulation Metadata and Analyses  Retrieve MetaTrajectory


Show 10 entries Search:

Id.	PDB	Type	SubType	ForceField	Solvent	Description	Time (ns)	Sequence
<input type="checkbox"/> <input type="text" value="Id."/>	<input type="text" value="Sel"/>	<input type="text" value="Sel"/>	<input type="text" value="Select"/>	<input type="text" value="Select"/>	<input type="text" value="Select"/>	<input type="text" value="Description"/>	<input type="text" value="Time (ns)"/>	<input type="text" value="Sequence"/>
<input type="checkbox"/> NAFlex_56merTIP3P	NONE	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P Electroneutral	500	CGCGATTGCCTAACGGA CAGGCATAGACGCTCTAT GCCTGTCCGTTAGGCAA TCGCG
<input type="checkbox"/> NAFlex_ACGT	NONE	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P Electroneutral	500	CGCGACGTCGCG
<input type="checkbox"/> NAFlex_56merSPCE	NONE	Dna	B	parmBSC1	SPCE	DNA-B Duplex Naked ParmBSC1 SPC/E Electroneutral	290	CGCGATTGCCTAACGGA CAGGCATAGACGCTCTAT GCCTGTCCGTTAGGCAA TCGCG

Showing 1 to 3 of 3 entries Previous 1 Next

Open Analyses for selected simulations Mark all Unmark all Reset

Results for this specific search are **3 simulations**. The information that will constitute the final meta-trajectory is the combination of all the selected trajectories, **1,290 μ s** (500 + 500 + 290 ns). Note that if with the most represented tetramer including **CG (TCGC)** was chosen, the generated meta-trajectory would correspond to up to **39 μ s** of simulated time, so this option should be used with care. Meta-trajectory can be obtained from the “**Retrieve Meta-trajectory**” button, after selecting the appropriate simulations.

Click on Id to open Simulation Metadata and Analyses  Retrieve MetaTrajectory

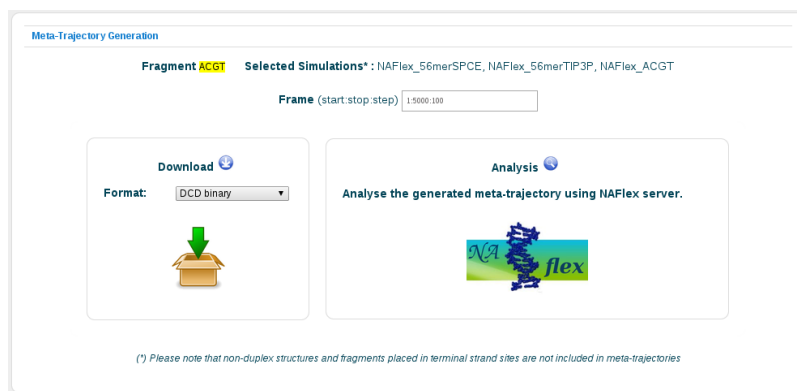
Show 10 entries Search:

Id.	PDB	Type	SubType	ForceField	Solvent	Description	Time (ns)	Sequence
<input checked="" type="checkbox"/> <input type="text" value="Id."/>	<input type="text" value="Sel"/>	<input type="text" value="Sel"/>	<input type="text" value="Select"/>	<input type="text" value="Select"/>	<input type="text" value="Select"/>	<input type="text" value="Description"/>	<input type="text" value="Time (ns)"/>	<input type="text" value="Sequence"/>
<input checked="" type="checkbox"/> NAFlex_56merTIP3P	NONE	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P Electroneutral	500	CGCGATTGCCTAACGGA CAGGCATAGACGCTCTAT GCCTGTCCGTTAGGCAA TCGCG
<input checked="" type="checkbox"/> NAFlex_ACGT	NONE	Dna	B	parmBSC1	TIP3P	DNA-B Duplex Naked ParmBSC1 TIP3P Electroneutral	500	CGCGACGTCGCG
<input checked="" type="checkbox"/> NAFlex_56merSPCE	NONE	Dna	B	parmBSC1	SPCE	DNA-B Duplex Naked ParmBSC1 SPC/E Electroneutral	290	CGCGATTGCCTAACGGA CAGGCATAGACGCTCTAT GCCTGTCCGTTAGGCAA TCGCG

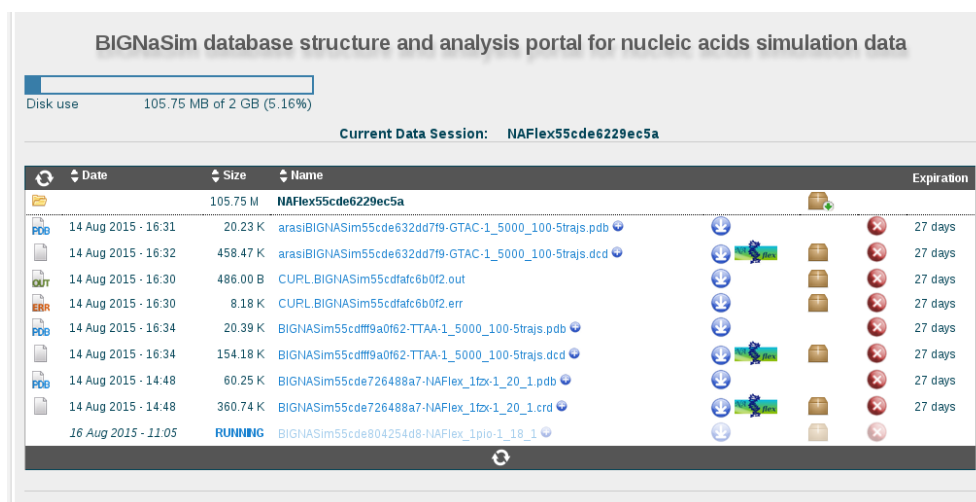
Showing 1 to 3 of 3 entries Previous 1 Next

Open Analyses for selected simulations Mark all Unmark all Reset

The next step for generating the meta-trajectory is to define a desired frame step. In this case, **100 snapshots** as representatives of each trajectory will be extracted (**Frame** input section: **1:5000:50** (for snapshot 1 to snapshot 5000, every 50 snapshots)). At this step, trajectory can be downloaded or sent to the **NAFlex** nucleic acid flexibility server.



As the generation of the trajectory is potentially a lengthy process, downloaded trajectories or meta-trajectories, a [private user space](#) is created where data is maintained. For registered users, the workspace is permanent, and data can be added in multiple sessions.



5.4 Analysis using fragments hierarchy. Correlation between CpG twist and ζ torsions.

The analysis of nucleic acids flexibility involves a large number of properties, from helical parameters, to hydrogen bonding or stacking energies, dihedral torsions, and distances, among others. Expert users usually combine such analyses in consistent ways, as they are highly correlated, and the combined study gives additional insight in the basis of conformational shifts. This practice requires to combine analyses done at different levels, the base-pair step (two contiguous base pairs), like twist or roll and analysis at the base pair level, like the hydrogen bonding pattern, or backbone torsions of the involved nucleotides. An expert DNA modeller has no problem on doing this manually, but it is a tedious and error-prone activity. BIGNASim includes the hierarchic relationships between sequence fragments (from individual bases to bp-steps), taken automatically (see Figure S1 and Table S5) from the simulation topology, and can be used to perform this kind of analysis in a straightforward way. We present here the necessary pipeline to analyse the correlation between CpG step twist and the ζ torsions of the neighbouring nucleotides. The high twist/low twist conformational change in the d(CpG) base

pair step is one of the most surprising sources of polymorphism in B-DNA. A detailed study of the phenomena (11) is available. The study relates the twist polymorphism with the BI/BII transitions related to ζ/ϵ backbone torsions. The key correlation analysis (See Figure 3 on reference (10)) implies to obtain the Twist helical parameter of the CpG bp-step and the ζ torsion of the two G bases at 3' of the CpG step. This specific analysis could be done through the portal, obtaining the twist and ζ as described in the previous examples, and downloading the corresponding raw data. However to illustrate the power of the MongoDB database structure the following Javascript code shows the pipeline required to obtain such combined set of data from BIGNASim database.

Javascript code pipeline to generate a combined analysis of a CpG base pair step Twist and ζ torsions of 3'G nucleotides in the "NAFlex_DDD_800ns" simulation.

```
//Step 1. Locate available CpG bp-steps
// CpG bp-steps are indicated with an idGroup of CGCG.
// CpG is symmetrical, but in other
// cases both strands should be considered.
// The "Class" field combines both possible orientations.

var IDSIM = "NAFlex_DDD_800ns";
var BPStep = "CGCG";
var BP1 = "CG";
Bpstps = db.groupDef.find({'_id.idSim': IDSIM, 'class': BPStep}).sort({'_id.n':
  1}).toArray();

//Alternatively search can be extended to All simulations with CG bpSteps
//Bpstps = db.groupDef.find({'class': BPStep}).sort({'_id.n': 1}).toArray();

var TwCG = [];
var ZetaW = [];
var ZetaC = [];
printjson(Bpstps.length + ' ' + BPStep + ' found')

// Iterate over all CpG steps in the sequence.
for (i = 0; i < Bpstps.length; i++) {
  IDSIM = Bpstps[i]._id.idSim;

//Step 2. Obtain relevant sequence positions.
// CGpos: Sequence position (n) of first nucleotide in Watson strand.
// It is part of the _id fragment key.
// G1pos: Sequence position of the Watson 3'G: n + 1 nucleotide
// G2pos: Sequence position of the Crick 3'G. It corresponds to the
// complementary nucleotide on the first (n) base pair in the CpG step.
// "comps" field relates CG base pair with its component nucleotides,
// is this case comps[1] is the G nucleotide.
CGpos = Bpstps[i]._id.n;
G1pos = CGpos + 1;
G2pos = db.groupDef.findOne(
  {'_id.n': CGpos,
   '_id.idSim': IDSIM,
   '_id.idGroup': BP1
  }).comps[1].n;

//Step 3. Collect twist values ordered by frame number.
TwCG[i] = db.analData.find(
  {'_id.idSim': IDSIM,
   '_id.nGroup': CGpos,
   '_id.idGroup': Bpstps[i]._id.idGroup
  }, {'CURVES.helical bpstep.twist': 1}
  ).sort({'_id.frame': 1}).toArray();

// Step 4. Obtain  $\zeta$  torsions values ordered by frame number
ZetaW[i] = db.analData.find(
  {'_id.idSim': IDSIM,
```

```

        '_id.nGroup': G1pos,
        '_id.idGroup': 'G'
    }, {'CURVES.backbone_torsions.zeta': 1}
    ).sort({' id.frame': 1}).toArray();
ZetaC[i] = db.analData.find(
    {'_id.idSim': IDSIM,
     '_id.nGroup': G2pos,
     '_id.idGroup': 'G'
    }, {'CURVES.backbone_torsions.zeta': 1}
    ).sort({' id.frame': 1}).toArray();

// Step 4. Output data as "Frame Twist G1Zeta G2Zeta
for (i = 0; i < TwCG.length; i++) {
    printjson(
        Bpstps[i]. id.idSim + ' ' +
        Bpstps[i]. id.idGroup + ' ' +
        Bpstps[i]._id.n
    );
    for (k in TwCG[i]) {
        if (TwCG[i][k]. id.frame > 0) {
            printjson(
                TwCG[i][k]._id.frame + ' ' +
                TwCG[i][k].CURVES.helical_bpstep.twist + ' ' +
                ZetaW[i][k].CURVES.backbone_torsions.zeta + ' ' +
                ZetaC[i][k].CURVES.backbone_torsions.zeta
            );
        }
    }
}
}

```

5.5 Combining Experimental and MD analysis

The example shows the information integrated in **BIGNASim** to compare **theoretical results** obtained with MD simulations with values obtained from **experimental structures/trajectories**. Comparison with experimental structures in **BIGNASim** is done at three different levels:

1. Individual helical parameters obtained from the corresponding experimental structure from the PDB
2. Averaged analysis of helical parameters from a complete dataset made from available nucleic acids experimental structures from the PDB
3. Experimental ensembles with PDB structures of equivalent sequences, analysed as pseudotrajectories, superimposed to the appropriate analysis

Individual helical parameters obtained from the corresponding experimental structure from the PDB

In those simulations having a **reference PDB structure**, the most direct analysis are the comparison between geometrical values extracted from the crystallographic structure and those averaged from the simulation trajectory. From the broad range of possible analyses, the **helical parameters**, and more specifically, the base-pair step helical parameters are the most used in the last scientific studies (see [references section](#), Ascona B-DNA Consortium articles). In these cases, **BIGNASim** database store not only the analysis for the simulated trajectory, but also the ones computed on the original PDB structure. Then, just joining both information, a direct comparison can be obtained. The most interesting cases where this comparison can give valuable information are those **nucleic acids complexed with protein or ligands**. The typical helical conformation of the nucleic acid structure can be **distorted** due to the influence of the

attached molecule. Studying the time-averaged values for a determined simulated trajectory, with its standard deviation, can give a clue about whether the nucleic acid structure explores the conformation needed to accept the interacting molecule, or it needs a complete change of shape. Additionally, docking regions can be easily identified due to the helical parameters distortions shown (e.g. **intercalator** ligands). To illustrate the procedure a protein-nucleic acid complex, PDB Code 1hlv, will be chosen. The **Quick Search section** can be used to select all simulations available related to such PDB entry. Quick search does a global search in any of fields in the database, including ontology terms, or even sequences. It is particularly useful to identify simulations based on experimental structures.

The screenshot shows the myBIGNASim website interface. At the top, there is a navigation bar with links: Home, Browse, Search, Global Analyses, Statistics, Help, and Suppl. Material. A search bar on the right contains the text "1hlv". Below the navigation bar, there is a section titled "Click on Id to open Simulation Metadata and Analyses". A table displays search results for the query "1hlv".

Id.	PDB	Type	SubType	ForceField	Solvent	Description	Time (ns)	Sequence
NAFlex_1hlv	1HLV	Dna	B	parmBSC1	SPCE	DNA-B Duplex Protein-Nuc ParmBSC1 SPC/E AddedSalt	1,000	AATCCCGTTTCCAACGAAGGC

Showing 1 to 1 of 1 entries

In this case, “1hlv” is found both as PDB reference and in the internal identifier of the simulation. The description column shows the details of the specific simulation, showing that corresponds to “naked” DNA, meaning that the available simulation was done with just the nucleic acid portion of the complex; otherwise, the description would indicate “complex”. This can also be seen in the simulation details, where a straight DNA molecule is depicted.

The screenshot shows the simulation details page for entry "NAFlex_1hlv". The page is titled "Entry: NAFlex_1hlv" and contains a section for "Nucleic Acid Data".

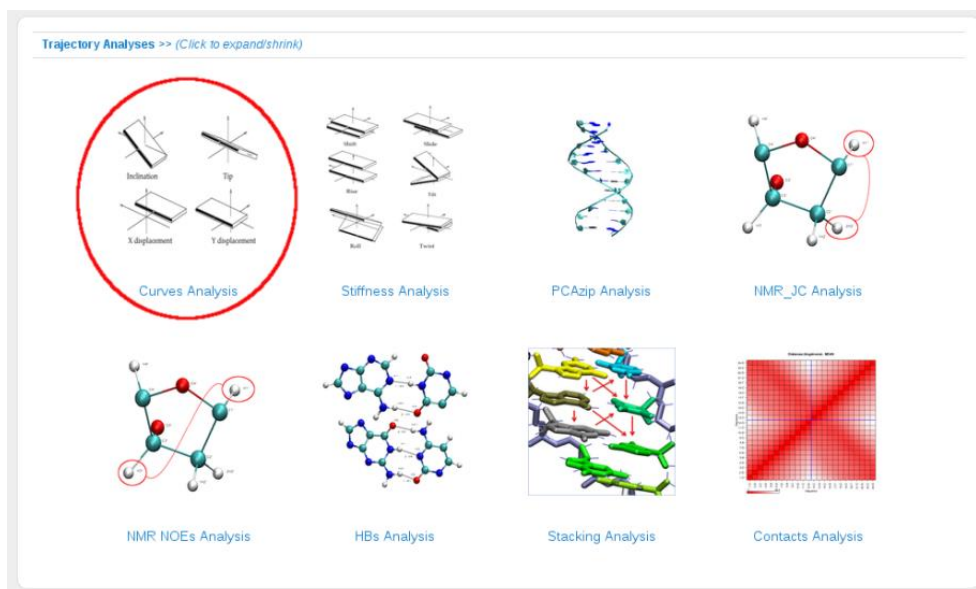
Sequence	AATCCCGTTTCCAACGAAGGC
Rev. Sequence	GCCTTCGTTGGAACGGGATT
Type	DNA
SubType	B
Chains	duplex
Pdb	1HLV [PDB] [NUCDB]
Ligands	No
Keywords	DNA-B Duplex Protein-Nuc ParmBSC1 SP C/E AddedSalt

Below the table, there is a 3D visualization of a DNA double helix structure. At the bottom of the page, there are three expandable sections:

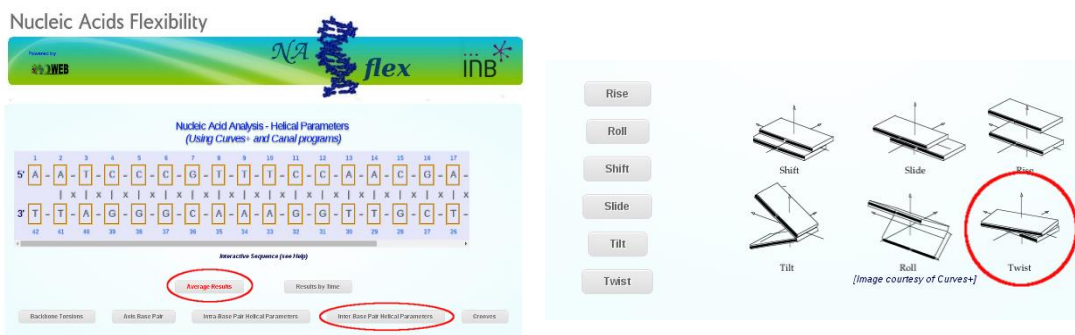
- MD Simulation >> (Click to expand/shrink)
- Trajectory Analyses >> (Click to expand/shrink)
- Trajectory Selection >> (Click to expand/shrink)

This will be more evident looking at the helical parameters comparison with the experimental values, which correspond to the complexed conformation.

Trajectory Analyses → **Curves** gives the set of pre-computed helical analyses available

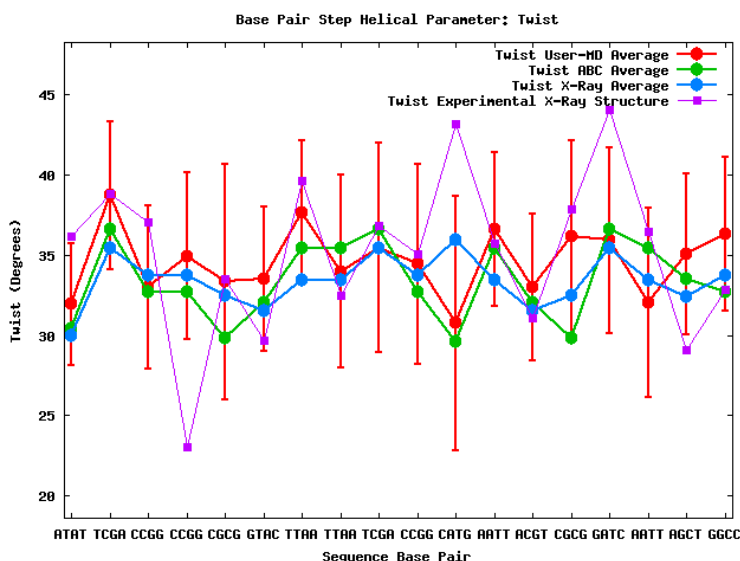


This will open the extended NAFlex interface included in BIGNASim. For a **complete description of NAFlex analysis interface** and its interactive sequence possibilities, please refer to NAFlex help pages (1). **Average Results** and **Inter-Base Pair Helical Parameters** should be selected.



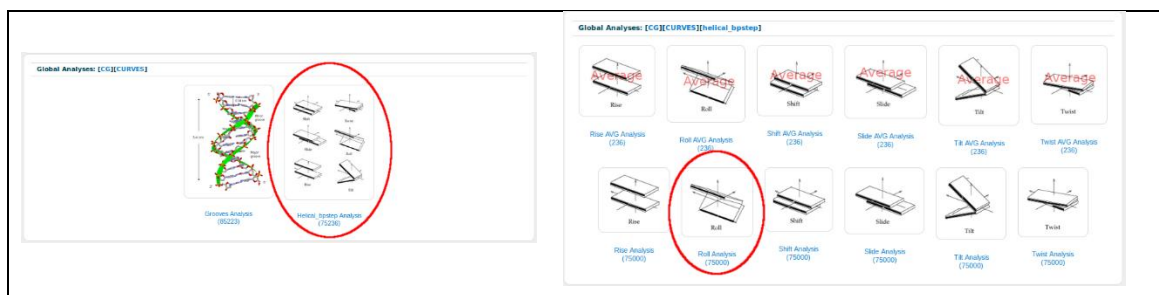
This will show a set of **six different inter-base pair step helical parameters**: *Rise*, *Roll*, *Shift*, *Slide*, *Tilt* and *Twist*. Each of them have an associated plot with the comparison between the time-averaged theoretical results with the experimental ones. As example, the figure below shows results of the **Twist** parameter along the sequence. Four lines with different colours are shown: **NAFlex_1hlv theoretical MD simulation** in red, with standard deviation values; average values coming from a set of **MD simulations** (green) and **X-ray structures** (blue) (11) and finally the values computed to the corresponding 1hlv PDB experimental structure in violet. One can easily identify the regions of the nucleic acid where the protein is attached, disturbing

the helical conformation. The same analysis can be obtained for the rest of base-pair parameters.

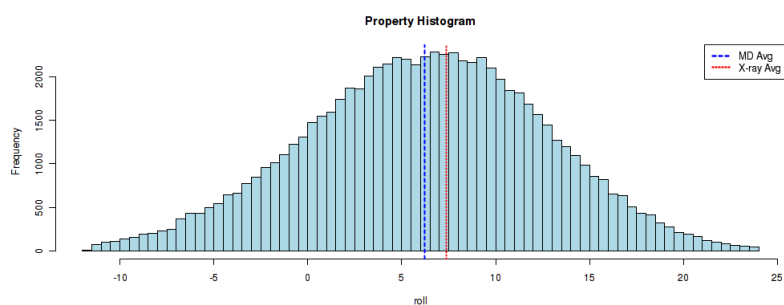


Averaged analysis of helical parameters from a complete dataset made from available nucleic acids experimental structures from the PDB

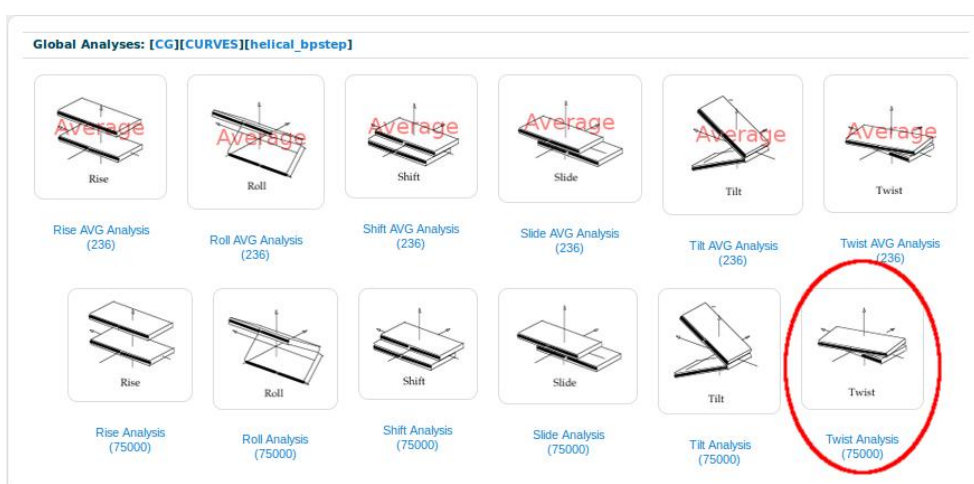
The second way **BIGNASim** offers to compare theoretical and experimental values covers a more **global vision** of nucleic acids flexibility. For that, an **averaged analysis** of standard fragments (**base, base pairs, and base-pairs steps**) of a complete dataset made from available nucleic acids **experimental structures** (taken directly from the PDB), is included in global analyses. See **BIGNASim** corresponding help section for details of the pipeline to obtain such structures. These averaged values can be used to compare **global values** extracted from simulations having information from many different simulations on a single **base, base pair, and base-pairs step** and thus obtain a direct representation of how much our simulations reproduce the experimental observations. To illustrate how this information is added in the **BIGNASim** interface, the set of global analyses for the **CG** base-pair step will be used, but any other fragment can be analysed likewise. The procedure required is “Global Analyses → CG → CURVES → Helical_bpstep → Roll”

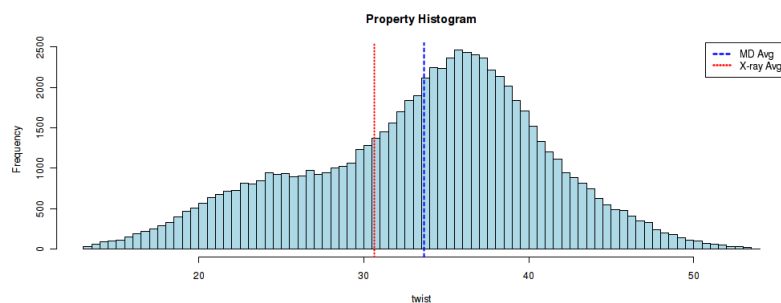


BIGNASim automatically generates a histogram with all the **Roll** values extracted from the database. Two vertical lines are plotted on the histogram: a **blue one**, indicating the **mean** for the represented set of values (**MDs**) and a **red one**, which corresponds to the **experimental** average value obtained from experimental structures.



In this case it can be easily seen that the observed values from the simulated systems agree pretty well with the **experimental** average. The distribution of values follows an almost perfect **normal distribution**, with the mean value (6.225) distant just 0.5 degrees from the **experimental** average value (7.35). A second base-pair step parameter, **Twist** show a different behaviour. It is known from reference (10) that **CG-Twist** show bimodality with a polymorphism between a **high twist** (~20°) and a **low twist** (~40°).





The difference with the previous **Roll** distribution is clear. In this case, two different distributions can be easily identified, centred at $\sim 25^\circ$ and $\sim 35^\circ$. Interestingly, although the **theoretical mean value** (33.64) is close to the **experimental one** (30.66), both averages hide information about the bimodality.

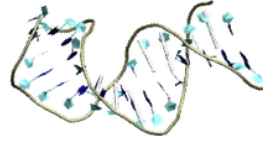
Experimental ensembles with PDB structures of equivalent sequences, analysed as pseudotrajectories, superimposed to the appropriate analysis

The last **experimental** comparison in **BIGNASim** is done using analysis taken from **experimental ensembles**. The procedure to build such ensembles can be found in [BIGNASim experimental trajectories help section](#). Having these **experimental ensembles** allows to compare theoretical values not just against experimental static information but also the ensemble of conformation found in experimental data. If a particular helical parameter exists in two different conformations (see previous section, **CG-Twist** example), that information can be obtained from the experimental ensemble and, thus, be compared with the MD simulation values. Unfortunately, due to the scarce number of deposited PDB structures for a determined nucleotide sequence, the number of generated **experimental trajectories** goes just up to 15 different ensembles (see [experimental trajectories table](#)). In this example, a particular structure will be used: a protein-DNA complex with PDB code 2LEF, the structure of a Lymphoid Enhancer-Binding Factor (Lef1 hmg domain, from mouse), complexed with DNA (15bp). As the deposited structure in the PDB was obtained by Nuclear Magnetic Resonance (NMR), it contains 12 different structures, what allows us to construct the experimental ensemble and compute the corresponding helical parameters. The simulation can be easily found from the **search section** (*Search by Ontology* → *System Type* → *Complex: Protein-Nuc*), and using its nucleotide sequence (*CACCCTTTGAAGCTC*) in the *Search by Sequence* section, or just browsing the whole database and ordering the results by the id. Once in the **simulation** page, **Trajectory Analyses** show a specific section about comparison to experiment when this is available:

Entry: NAFlex_2lef

Nucleic Sequence:

Sequence	CACCCCTTGAAGCTC
Rev. Sequence	GAGCTTCAAAGGGTG
Type	DNA
SubType	B
Chains	duplex
Pdb	2LEF
Ligands	No
Keywords	DNA-B Duplex Naked ParmBSC1 SPC/E AddedSalt

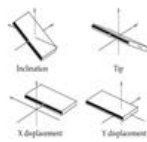


[MD Simulation >>](#) (Click to expand/shrink)

[Trajectory Analyses >>](#) (Click to expand/shrink)

[Trajectory Selection >>](#) (Click to expand/shrink)

Trajectory Analyses >>



Curves Analysis



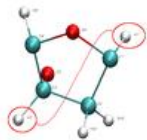
Stiffness Analysis



PCAzip Analysis



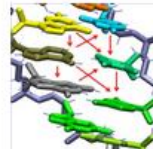
NMR_JC Analysis



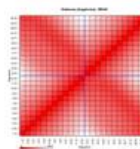
NMR NOEs Analysis



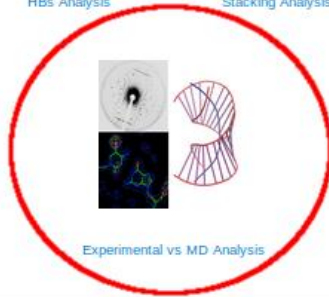
HBs Analysis



Stacking Analysis



Contacts Analysis



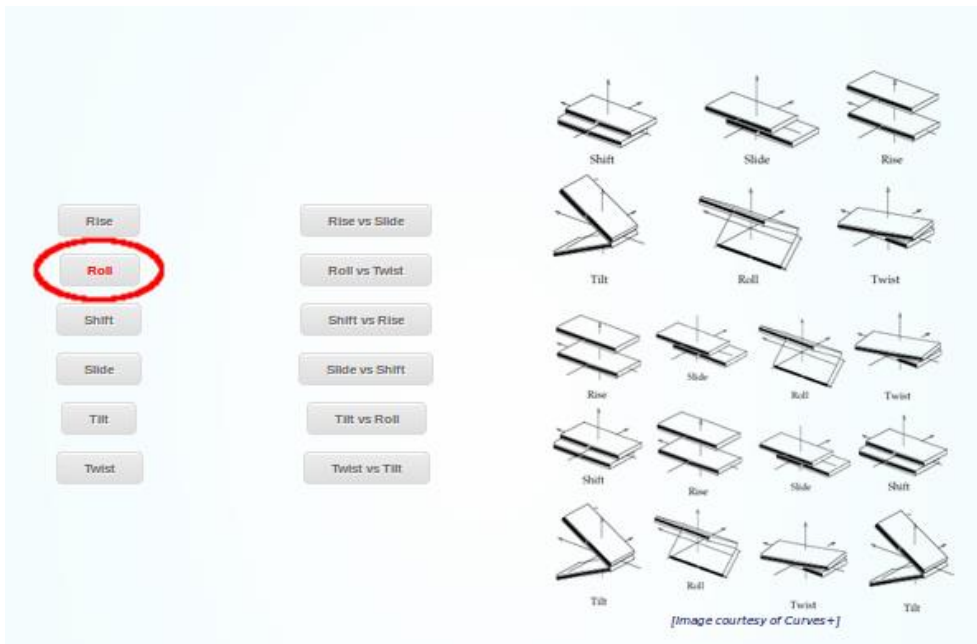
Experimental vs MD Analysis

The path to obtain the compared analysis is “Average Results → Inter-BasePair Helical parameters”

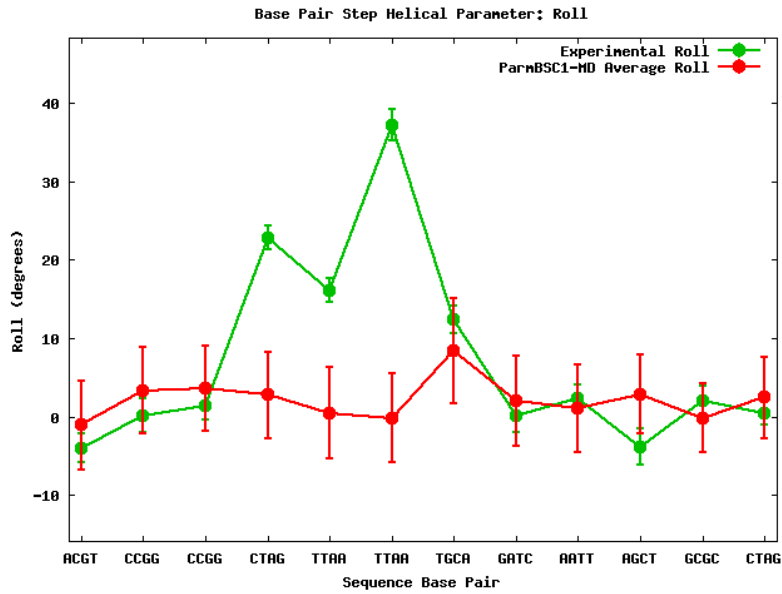
Nucleic Acids Flexibility



Please note that direct access to this **NAFlex interface** for all the sequences having a calculated experimental trajectory is also available in the [Experimental Trajectories](#) help section. As this text is intended to be a tutorial, all the previous steps to get to this point are detailed. Two examples will be described: one single parameter (Roll), or two parameters (Roll and Twist).

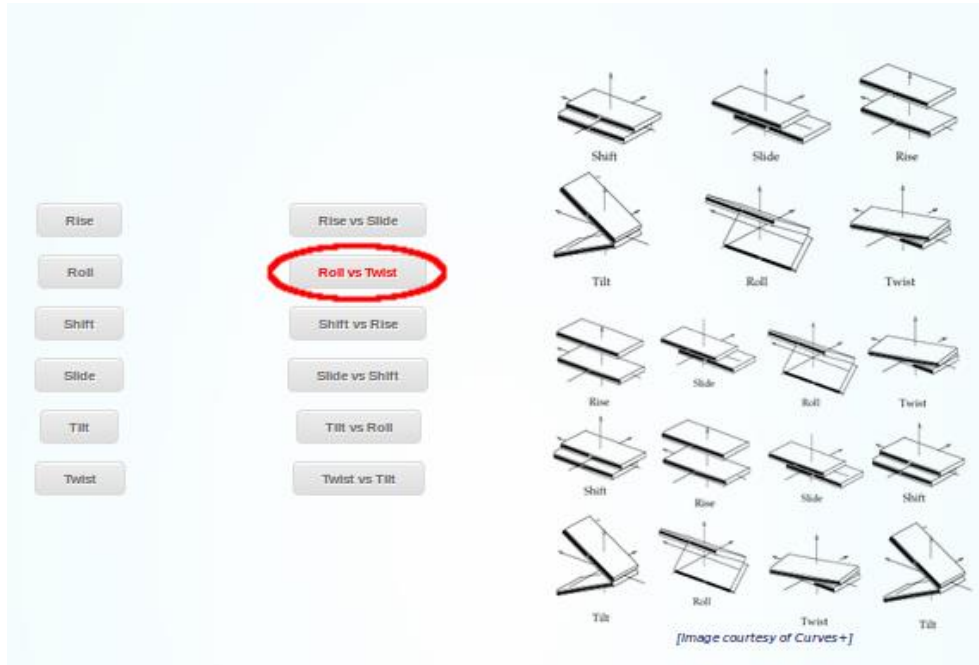


This is the resulting **Roll** plot:

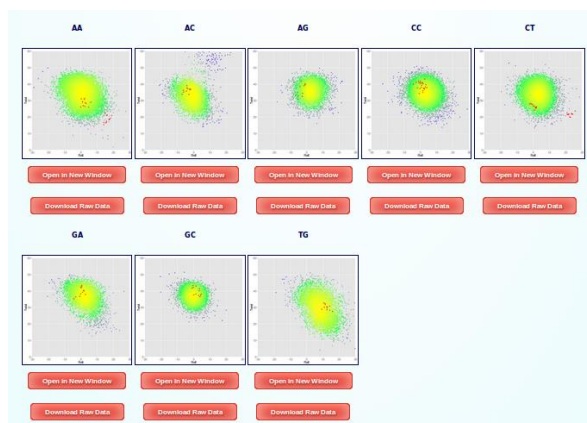


The plot shows two lines representing **time-averaged** values for the **Roll** parameter: **red line** corresponds to the **MD simulation** and **green line** corresponds to the **experimental ensemble**. It can be clearly identified a **distorted region** of the DNA, from the 4th to the 6th base-pair steps of the sequence, probably influenced by the attached protein. The available MD simulation (naked B-DNA) is not exploring these specific **Roll** values, so the experimental behaviour of the nucleic acid is in this case clearly influenced by the protein.

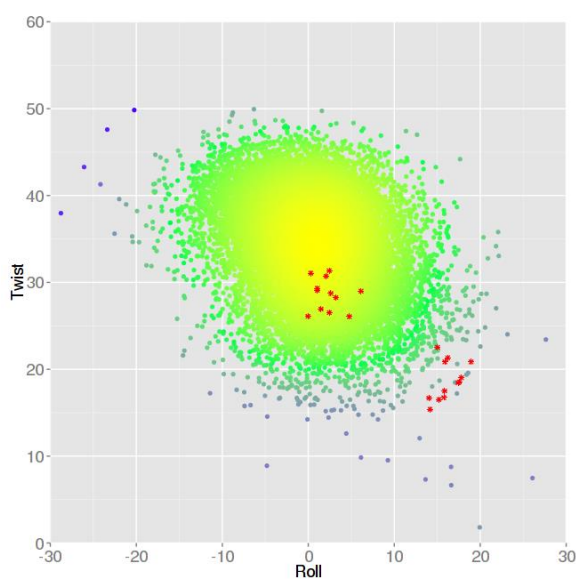
The **Roll and Twist** example can be followed in a similar way.



Analyses of **helical parameters pairs** are done for each of the base-pairs present in the sequence. In this case, we have **6 different plots**, corresponding to (AA/TT), AC, AG, CC/GG, CT, GA, GC and TG).

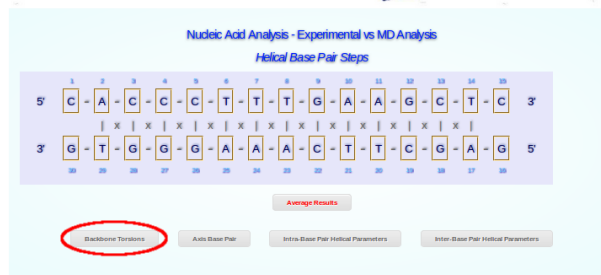


The following figure shows the base-pair step **AA (TT) Twist vs Roll** in more detail:

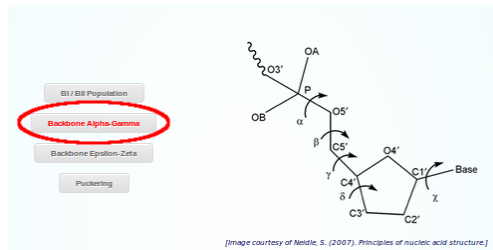


Information shown in these **3D plots** are points corresponding to the correlation between both parameters, where colours represent different population densities, from **higher density in yellow**, to **lower density in blue**. Experimental values are shown as **asterisk symbols in red colour**. As expected from the results found in the previous **Roll** plot, there is a strong deviation in the **Roll** parameter for some of the base-pair steps, placed in a region of the plot almost unexplored by the complete MD simulation. Unsurprisingly, they also have distorted values of the **Twist** parameter.

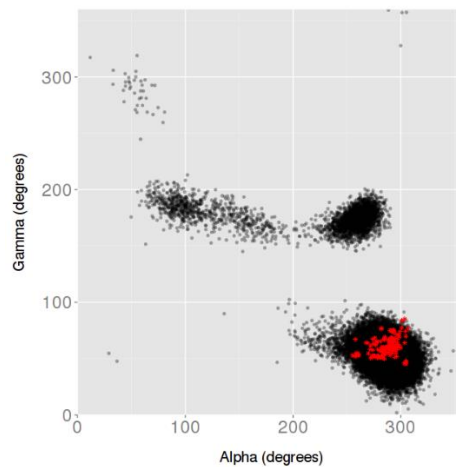
But comparisons are made also for **backbone torsions, axis base pairs and intra-base pair helical parameters**. As a last example, the following screenshots show a similar example with **Backbone Torsions**:



There are 4 different analyses available: **BI/BII population** and **Puckering** percentages, and two **angle analysis**: α/γ , ϵ/ζ . We choose **Backbone α/γ** analysis:



The resulting plot shows the **MD (black)** vs the **experimental (red)** values:



No distortion is observed in this case, with the **experimental ensemble** exploring just a **single conformation**, whereas the **MD simulation** is exploring different well-known regions of the conformational space.

6 Supplementary Tables

6.1 Table S1. Analysis types available in BIGNASim

Analysis Type	Fragment Scope	Available Data Types	Software
Cartesian analysis <ul style="list-style-type: none"> RMSd, RMSf, Radius of Giration, SASA 	Base	Average values, Plot (Sequence), Plot (Time)	pTraj (12), Gromacs (13)
Backbone geometry <ul style="list-style-type: none"> BI/BII population Sugar puckering (N,E,S,W population) Proportion of alpha/gamma torsions Torsions ($\alpha, \beta, \gamma, \varepsilon, \zeta, \chi$, phase) 	Base	Average values, Plot (Sequence), Plot (Time)	Curves+ (14)
Translation Base Pair Axis <ul style="list-style-type: none"> X-Displacement, Y-Displacement 	Base Pair	Plot (Sequence), Plot (Time)	Curves+ (14)
Rotational Base Pair Axis <ul style="list-style-type: none"> Inclination, Tip 	Base Pair	Plot (Sequence), Plot (Time)	Curves+ (14)
Translation Base Pair Helical parameters <ul style="list-style-type: none"> Shear, Stretch, Stagger 	Base Pair	Plot (Sequence), Plot (Time)	Curves+ (14)
Rotation Base Pair Helical parameters <ul style="list-style-type: none"> Buckle, Opening, Propeller 	Base Pair	Plot (Sequence), Plot (Time)	Curves+ (14)
Translation Base Pair Step parameters <ul style="list-style-type: none"> Rise, Slide, Shift 	BP Step	Plot (Sequence), Plot (Time)	Curves+ (14)
Rotational Base Pair Step parameters <ul style="list-style-type: none"> Roll, Tilt, Twist 	BP Step	Plot (Sequence), Plot (Time)	Curves+ (14)
Groove dimensions <ul style="list-style-type: none"> Major and Minor grooves (depth, width) 	BP Step	Plot (Sequence), Plot (Time)	Curves+ (14)
Interactions <ul style="list-style-type: none"> Base Pair Canonical Hydrogen Bonds (distances) Hydrogen Bond energies Stacking Energies 	Base, BP Step	Plot (sequence), Plot (Time), Contact Maps	pTraj (12), in-house
Stiffness analysis <ul style="list-style-type: none"> Stiffness force constants (Twist, roll, Tilt, Rise, Shift, Slide) Complete stiffness Matrix 	Base Pair Step		Curves+ (14)
NMR Observables <ul style="list-style-type: none"> Vicinal J-Couplings Nuclear Overhauser Effect (NOE) 	Base	Values, Plot (Time) 2D NOE Plots	in-house
Principal component analysis <ul style="list-style-type: none"> EigenValues & EigenVectors Collectivity index EigenVector Stiffness constant Animated Trajectories Trajectory projections 	Collective	Values. Animated Traj. 3D view (JsMol) Projections Time Plots	PCASuite (15,16)
Trajectory video	Collective	Standard video formats, 3D view	VMD (17)

6.2 Table S2. Cassandra trajectory database structure.

Topology table (idSimulation)	Trajectory table (idSimulation)
atom_num (Partition Key) atom_name atom_type chain_code residue_code residue_num	frame (Partition Key) atom_id (Clustering Key) x y z (Box size data is included in the same frame as additional pseudo-atoms)

6.3 Table S3. Structure of main MongoDB collections

MongoDB collection	Main Index components	Description
<i>simData</i>	idSim	Simulation metadata, following a specifically defined ontology
<i>analDefs</i>	idSim, idAnal	Analysis description, one document stored for every analysis result item available. Analysis available could differ from one simulation to another
<i>groupDef</i>	IdSim, (idGroup,nGroup)	Molecular groups (bases, base pairs, base-pair steps, molecular fragments) defined in the simulated system
<i>analData</i>	idSim, (idGroup,nGroup), nFrame (nFrame = 0: Averaged analysis data) (nGroup = 0: All system analysis)	Analysis results. The most appropriate data model for each analysis type is used.
<i>analBinFiles</i>	Id. Above	Binary files with pre-calculated analysis results (plots, images, etc.)

6.4 Table S4. Representative data structures stored in the MongoDB analysis subsystem.

Examples correspond to NAFlex_DDD_bsc1 simulation (Drew Dickerson dodecamer).
Analysis correspond to fragments labelled at position 5.

```
{
  _id: 'NAFlex_DDD_bsc',
  dataset: 'ParmBSC',
  NucType: 'DNA',
  moleculeType: 'Dna',
  SubType: 'B',
  PDB: 'lnaj',
  sequence: 'CGCGAATTCGCG',
  rev sequence: 'CGCGAATTCGCG',
  Parts: 'DNA+iones',
  Chains: 'duplex',
  soluteResidues: 24,
  soluteAtoms: 758,
  soluteCharge: -22,
  CounterIons: 'Na+',
  IonicConcentration: 'Electroneutrality',
  totalIons: 0,
  totalAtoms: 758,
  AdditionalSolvent: 'No',
  solventResidues: 0,
  solventAtoms: 0,
  totalResidues: 24,
  AdditionalMolecules: 'No',
  Ligands: 'No',
  AdditionalIons: 'No',
  IonsParameters: 'Dang',
  ionsModel: '-',
  Water: 'TIP3P',
  Author: 'P.D.',
  Temperature: 298,
  forceField: 'parmbsc0',
  date: '06/23/15',
  saltConcentration: 0,
  totalCharge: -22,
  time: 10000,
  Format: 'netcdf',
  FrameStep: '20ps',
  Frames: 500000,
  Trajectory: 'DDD10micros_nowat.nc',
  Topology: 'DDD10micros_nowat.prmtop',
  RMSd_avg: 1.734,
  RMSd_stdev: 0.368,
  Comments: ,
  ontology: ['10101','10402','10202','10301','10603','201010103','2010401','2010501'],
  description: 'DNA|B|Duplex|Naked|DDD|ParmBSC1|TIP3P|Electroneutral'
}

// Analyses at base level (5-A)
{
  id: {frame: 0,idSim: 'NAFlex_DDD_bsc1',idGroup: 'A',nGroup: 5},
  CURVES: {
    backbone_torsions: {
      BI_population: [82.26],
      canonical_alpha_gamma: [97.22],
      puckering: [4.34,6.64,88.96,0.06]
    }
  },
  NMR_JC: {
    J1p2p-DNA: [10.011,1.764], J1p2pp-DNA: [5.213,1.430],
    J2p3p-DNA: [6.309,0.999], J2pp3p-DNA: [2.276,1.734],
    J3p4p-DNA: [1.883,1.693]
  },
  NMR_NOE: {
    H1p-H2p: [3.0252,0.0806], H1p-H2pp: [2.3772,0.1157],

```

```

H2p-H3p: [2.3771,0.1072], H2p-H4p: [3.7905,0.0933],
H2pp-H3p: [2.7312,0.1112], H2pp-H4p: [3.9616,0.3060],
H3p-H4p: [2.7804,0.1090], H1p-H4p: [3.1969,0.2762],
H1p-H3p: [3.7813,0.1239], H1p-H8: [3.5212,0.6081],
H2p-H8: [3.6553,0.5865], H2pp-H8: [2.4727,0.4316],
H3p-H8: [4.6430,0.4418], H4p-H8: [5.7722,0.6122]
}
}
// Analyses at base pair level (5-AT)
{
  _id: {frame: 0,idSim: 'NAFlex_DDD_bscl',idGroup: 'AT',nGroup: 5},
  CURVES: {
    axis bp: {
      inclin avg: [3.44,4.6],
      xdisp avg: [-0.96,0.7],
      ydisp_avg: [0.14,0.5]
    },
    helical bp: {
      buckle avg: [5.62,10.2], opening avg: [1.87,5.4],
      propel avg: [-14.99,7.7], shear avg: [0.14,0.4],
      stagger_avg: [-0.01,0.4], stretch_avg: [0.03,0.1]
    }
  }
}
// Analyses at base-pair step level (5-AATT)
{
  _id: {frame: 0,'idSim: 'NAFlex_DDD_bscl',idGroup: 'AATT',nGroup: 5},
  CURVES: {
    grooves: {
      majd avg: [5.71,1.4], majw avg: [11.93,1.6],
      mind_avg: [4.95,0.7], minw_avg: [5.22,1.4]
    },
    helical_bpstep: {
      rise avg: [3.37,0.3], roll avg: [1.21,5],
      shift avg: [-0.27,0.6], slide avg: [-0.48,0.5],
      tilt avg: [-2.38,4], twist avg: [36.35,4.6]
    }
  },
  STACKING: {
    stW: [-7.1088648,0.982759350869257],
    stC: [-1.43531339999999,0.731727353308896],
    HB: [-11.045809,1.11398008443553]
  },
  STIFFNESS: {
    FORCE_CTES: {
      PROD: [0.01606],
      MAT: [
        2.31756, 0., 0., 0.00173, 0., 0.,
        0., 5.57878, 2.97746, 0., -0.01652, -0.20697,
        0., 2.97747, 11.11783, 0., 0.04616, -0.16052,
        0.00173, 0., 0., 0.04624, 0., 0.,
        0., -0.01652, 0.04616, 0., 0.03597, 0.0064,
        0., -0.20696, -0.16051, 0., 0.0064, 0.06717
      ],
      rise avg: [11.11783], roll avg: [0.03597],
      shift avg: [2.31756], slide avg: [5.57878],
      tilt avg: [0.04624], twist avg: [0.06717]
    }
  }
}
}

```

6.5 Table S5. Complete structure of data objects showing hierarchical relationships derived from the central tetramer of the Drew-Dickerson dodecamer (A⁵ATT)

```
//Fragment definition
// 5-AATT
//   TTAA
{
  _id: {n: 5, idSim: '0001', idGroup: 'frag1'},
  type: 'fragment',
  fragment end: 8,
  comps: [
    {n: 5, idSim: '0001', idGroup: 'AATT'},
    {n: 6, idSim: '0001', idGroup: 'ATAT'},
    {n: 7, idSim: '0001', idGroup: 'TTAA'},
  ]
}

//Base pair step definitions
// 5-AA
//   TT
{
  id: {n: 5, idSim: '0001', idGroup: 'AATT'},
  class: 'AATT',
  type: 'stepid',
  id: '5-AATT',
  comps: [
    {n: 5, idSim: '0001', idGroup: 'AT'},
    {n: 6, idSim: '0001', idGroup: 'AT'}
  ]
}
// 6-AT
//   TA
{
  id: {n: 6, idSim: '0001', idGroup: 'ATAT'},
  class: 'ATAT',
  type: 'stepid',
  id: '6-ATAT',
  comps: [
    {n: 6, idSim: '0001', idGroup: 'AT'},
    {n: 7, idSim: '0001', idGroup: 'TA'}
  ]
}
// 7-TT
//   AA
{
  _id: {n: 7, idSim: '0001', idGroup: 'TTAA'},
  class: 'AATT',
  type: 'stepid',
  id: '7-TTAA',
  comps: [
    {n: 7, idSim: '0001', idGroup: 'TA'},
    {n: 8, idSim: '0001', idGroup: 'TA'}
  ]
}

//Base pair definitions
// 5-A
//   T
{
  id: {n: 5, idSim: '0001', idGroup: 'AT'},
  class: 'AT',
  type: 'bpid',
  id: '5-AT',
  comps: [
    {n: 5, idSim: '0001', idGroup: 'A'},
    {n: 20, idSim: '0001', idGroup: 'T'}
  ]
}
// 6-T
//   A
{
  _id: {n: 6, idSim: '0001', idGroup: 'AT'},
```

```

class: 'AT',
type: 'bpid',
id: '6-AT',
comps: [
  {n: 6, idSim: '0001', idGroup: 'A'},
  {n: 19, idSim: '0001', idGroup: 'T'}
]
}
// 7-T
// A
{
  _id: {n: 7, idSim: '0001', idGroup: 'TA'},
class: 'AT',
type: 'bpid',
id: '7-TA',
comps: [
  {n: 8, idSim: '0001', idGroup: 'T'},
  {n: 18, idSim: '0001', idGroup: 'A'}
]
}
// 8-T
// A
{
  _id: {n: 8, idSim: '0001', idGroup: 'TA'},
class: 'AT',
type: 'bpid',
id: '8-AT',
comps: [
  {n: 8, idSim: '0001', idGroup: 'T'},
  {n: 17, idSim: '0001', idGroup: 'A'}
]
}

```

```

//Base definitions
{
  id: {n: 6, idSim: '0001', idGroup: 'A'},
class: 'A', type: 'bid', id: '6-A'
}
{
  _id: {n: 7, idSim: '0001', idGroup: 'A'},
class: 'A', type: 'bid', id: '7-A'
}
{
  _id: {n: 8, idSim: '0001', idGroup: 'T'},
class: 'T', type: 'bid', id: '8-A'
}
{
  id: {n: 9, idSim: '0001', idGroup: 'T'},
class: 'A', type: 'bid', id: '6-A'
}
{
  id: {n: 17, idSim: '0001', idGroup: 'A'},
class: 'A', type: 'bid', id: '17-A'
}
{
  id: {n: 18, idSim: '0001', idGroup: 'A'},
class: 'A', type: 'bid', id: '18-A'
}
{
  _id: {n: 19, idSim: '0001', idGroup: 'T'},
class: 'T', type: 'bid', id: '19-T'
}
{
  id: {n: 20, idSim: '0001', idGroup: 'T'},
class: 'T', type: 'bid', id: '20-T'
}
}

```

6.6 Table S6. BIGNASim ontology terms

Hierarchic Id	Label	Description
1	System	Composition of simulated system
101	NA_Type	Type of Nucleic Acid
10101	DNA	DNA
10102	RNA	RNA
1010201	Viral	Viral RNA
1010202	Synthetic	Synthetic RNA
1010203	tRNA	tRNA
1010204	Messenger	Messenger RNA
1010205	Ribosomal	Ribosomal RNA
10103	DNA-RNA_Hybrid	DNA-RNA Hybrid
10104	PNA	PNA
10199	OtherNAType	Other NA Types
102	Architecture	Architecture (strand organization)
10201	SingleStrand	Single Strand
10202	Duplex	Duplex
1020201	Canonical	Canonical WC pairing
102020101	Linear	Linear duplex
102020102	Circular	Circular duplex
1020202	Hogsteen	Hogsteen pairing
10203	Triplex	Triplex
1020301	ParallelTrip	Parallel Triplex
1020302	AntiParallelTrip	Antiparallel Triplex
10204	Quadruplex	Quadruplex
1020401	Gloop	Gloop
1020402	ParallelQuad	Parallel Quadruplex
1020403	AntiparallelQuad	Antiparallel Quadruplex
1020404	IDNA	I-DNA
1020405	IRNA	I-RNA
10205	HollidayJnt	Holliday junction
10206	3WayJnt	3-Way junction
10207	RNA PseudoKnot	RNA PseudoKnot
10208	Ribozymes	Ribozymes
10209	Large Ribosomal RNA	Large Ribosomal RNA
10210	Riboswitch	Riboswitch
10211	tRNA	tRNA fold
10212	G1introns	G1introns
10213	G2introns	G2introns
10214	RNA nanostructures	RNA nanostructures
10299	OtherStructureType	Other structure types
103	System_Type	Type of complex involving NA
10301	Naked	Naked, uncomplexed
10302	Complex	Complexed Nuc. Acid
1030201	Protein-nuc	Complex Protein Nucl. Acid
103020101	Enzymes	Complexed with Enzyme

103020102	Binding	Binding Proteins
10302010201	Regulatory	Regulatory Proteins (Trans Factors, etc)
10302010202	SstrandBind	Single Strand Binders
10302010203	Nucleosome	Nucleosome proteins
10302010299	OtherBindProt	Other binding proteins
1030202	Ligand-nuc	Ligand - Nucleic Acid complexes
103020201	Intercalator	Intercalator
103020202	MinGBinder	Minor groove binder
103020203	MajGBinder	Major groove binder
103020204	HybridBinder	Hybrid binders
104	OriginalHelicalConformation	Original helical conformation of the Nucleic Acids
10401	A	A
10402	B	B
10403	Z	Z
10404	Hogsteen	Hogsteen
10405	MixedHConf	Mixed conformations
10499	OtherHConf	Other Conformations
105	SequenceModifications	Modifications of Nucleic Acids Sequence
10501	ModifiedNucleotides	Modified Nucleotides
10502	CrossLinked	CrossLinked
10503	EpigeneticVariants	EpigeneticVariants
10504	SequenceMismatches	Sequence Mismatches
10599	OtherSeqMod	Other modifications
106	SequenceFeatures	Relevant features related to sequence
10601	PolyA	Poly A Track
1060101	BrokenPolyA	Broken Poly A Track
10602	PolyG	Poly G Track
10603	DrewDickersonD	Drew Dickerson Dodecamer
10604	SeqMismatch	Sequence Mismatches
107	Local structures	Local structure features
10701	Kink Turn	Kink Turn
10702	Bulges	Bulges
10703	Internal loops	Internal loops
10704	Interacting loops	Interacting loops
1070401	Kissing loops	Kissing loops
1070402	Ring RNA	Ring RNA
10705	Hairpin loops	Hairpin or Stem Loops
1070501	Triloops	Triloops
1070502	Tetraloops	Tetraloops
1071503	Hexaloops	Hexaloops
1071504	tRNA Fragments	tRNA Fragments
1071505	Sarcin-Ricin	Sarcin-Ricin
1071506	TAR RNA	TAR RNA
1071507	IRES Domains	IRES Domains
2	Simulation	Simulation Data
201	SimConditions	Simulation settings
20101	ForceField	ForceField

2010101	Amber	Cornell ForceField family
201010101	Parm99	Parm99
201010102	ParmBSC0	ParmBSC0
201010103	ParmBSC1	ParmBSC1
201010104	ParmBSC0-OL1	ParmBSC0-OL1
201010105	ParmBSC0-OL4	ParmBSC0-OL4
201010106	ParmBSC0-OL1-OL4	ParmBSC0-OL1-OL4
201010107	ParmBSC0-CG	ParmBSC0-Cheng/Garcia
2010102	Charmm	Charmm ForceField family
201010201	Charmm36	Charm66
2010199	OtherFF	Other forcefields
20102	Length	Length of simulations
2010201	NanoSecondRange	Between 1 ns and 1 μ s
2010202	MicroSecondRange	Over 1 μ s
20103	Temperature	Simulation temperature
2010301	Physiological	Physiological (around 298, 300K)
2010302	NonPhysiological	NonPhysiological
20104	Solvent	Solvent used in the simulation
2010401	Water	Water only
2010402	Mixed	Mixture water and other solvent
201040201	Wat-Ethanol	Water Ethanol mixture
20105	Charge	Charge model
2010501	Electroneutral	Counter ions to compensate NA charge
2010502	AddedSalt	Added counterions over charge compensation
201050201	Physiological	Physiological (0.15M)
201050202	NonPhysiological	NonPhysiological
20106	IonParam	Parameter used for ion description
2010601	Dang	Dang
2010602	Cheatham	Cheatham
202	TrajectoryType	Type of trajectory related to conformation changes
20201	Equilibrium	Equilibrium (thermal fluctuations without major conf. Changes)
20202	Folding	Folding or Unfolding
20203	Transition	Transition between known conformations
20299	OtherTrajType	Other type of trajectory
3	Analysis	Analysis: any data derived from trajectories (simulated or experimental ensembles)
301	TimeScope	Time scope of the analysis
30101	Snapshot	Analysis made on a single snapshot
30102	TimeAvg	Time averaged analysis
302	FragmentScope	Fragment scope of the analysis
30201	SingleBase	Analysis done on a single residue (base)
30202	GroupAvg	Analysis done on a group of residues
3020201	BP	Analysis done on a base pair (considering the main NA pairing)
3020202	BPStep	Analysis done on a base pair step (2 consequent base pairs, considering the main pairing)
3020203	SeqFragment	Analysis done on other sequence fragments

30203	FullSystem	Analysis done on the complete system
30204	Metatrajectory	Analysis done on a group of trajectories
30205	ExpStructure	Analysis done on a experimental structure
303	AnalysisType	Type of analysis
30301	BackboneTorsions	Backbone Torsions
3030101	BI/BIIPopulation	Proportion of BI/BII population
3030102	SugarPuckering	Sugar puckering populations (N,E,S,W)
3030103	AGCanonical	Proportion of canonical alpha/gamma torsions
30302	HelicalParam	Helical parameters
3030201	AxisBP	Base Pair Axis parameters
303020101	AxisBPTras	Traslational Base Pair Axis parameters
30302010101	Xdisp	X-Displacement
30302010102	Ydisp	Y-Displacement
303020102	AxisBPRot	Rotational Base Pair Axis parameters
30302010201	Inclination	Inclination
30302010202	Tip	Tip
3030202	HelicalBP	Base Pair Helical parameters
303020201	HelicalBPTrans	Translational Base Pair Helical parameters
30302020101	Shear	Shear
30302020102	Stretch	Stretch
30302020103	Stagger	Stagger
303020202	HelicaBRRot	Rotational Base Pair Helical parameters
30302020201	Buckle	Buckle
30302020202	Opening	Opening
30302020203	Propeller	Propeller
3030203	HelicalBPStep	Base Pair Step Helical parameters
303020301	HelicaBPStepTrans	Translational Base Pair Step Helical parameters
30302030101	Rise	Rise
30302030102	Slide	Slide
30302030103	Shift	Shift
303020302	HelicaBPStepRot	Rotational Base Pair Step Helical parameters
30302030201	Roll	Roll
30302030202	Tilt	Tilt
30302030203	Twist	Twist
30303	GrooveAnalysis	Groove analysis
3030301	MajorGroove	Major Groove
303030101	MajGDepth	Depth of the Major Groove
303030102	MinGWidth	Width of the Major Groove
3030302	MinorGroove	Minor Groove
303030201	MinGDepth	Depth of the Minor Groove
303030202	MinGWidth	Width of the Mino Groove
30304	Interactions	Analysis of interactions
3030401	Hbonds	Hydrogen bonds (distances)
303040101	WC	Watson-Crick Hydrogen Bonds
303040199	Other	Other Hydrogend Bonds
3030402	Stacking	Stacking interactions
303040201	Wstrand	Stacking on the Watson Strand

303040202	Cstrand	Stacking on the Crick Strand
303040203	Crossed	Stacking between strands
30305	NMR	NMR observables
3030501	NOE	NOE
3030502	JC	J-Couplings
30306	Stiffness	Stiffness analysis
3030601	ForceConstant	Force Constants
303060101	Fctwist	Fctwist
303060102	Fcroll	Fcroll
303060103	Fctilt	Fctilt
303060104	Fcrise	Fcrise
303060105	Fcshift	Fcshift
303060106	Fcslide	Fcslide
3030602	ForceMatrix	Matrix of stiffness constants (twist, roll, tilt, rise, shift, slide)
3030603	ForceProduct	DiagonalProduct
30307	TrajectoryVideo	Video of Trajectory in standard formats
30308	TrajectoryData	Trajectory data
3030801	PCAZip	Trajectory in PCZ format
30309	Cartesian	Cartesian analysis
3030901	RMSd	Root Mean Square Deviation (RMSd)
3030902	RMSf	Root Mean Square Fluctuation (RMSf)
3030903	RadGyration	Radius of Gyration
3030904	Bfactor	B - Factor
3030905	AvgStruct	Average structure
30310	PCAnalysis	Principal Component analysis
3031001	EigenValues	PCA EigenValues
3031002	NumberEV	Number of PCA EigenValues for a given variance
3031003	EigenVal	Vector of eigenValues
3031002	EigenVector	PCA EigenVectors
3031003	TrajectoryProj	Projections of trajectory
3031004	Animated trajectory	Trajectory animated following given eigenvectors
3031005	Entropy	Entropy prediction
3031005	Schlitter	Entropy prediction using Schlitter protocol
303100502	Androcioaei	Entropy prediction using Androcioaei protocol
3031006	Variance	Variance measured in the trajectory
30311	ContactMaps	Contact Maps

6.7 Table S7. Deposition requirements

Acceptable topology formats	Preferred: PDB, Amber TOP, Gromacs GRO, ITP, RTP, NAMD PSF http://www.mdanalysis.org/mdanalysis/documentation_pages/topology/init.html
Acceptable trajectory formats	Preferred: DCD, CRD, XTC, PDB (Models), NetCDF, BINPOS http://www.mdanalysis.org/mdanalysis/documentation_pages/coordinates/init.html
Minimum set of metadata	
Dataset Description	Description/aim of the study References for supporting publication(s)
System Description	Reference experimental structure (PDB, NDB id) Type of Nucleic Acid, Main architecture (S. strand, Duplex, etc.), RNA type Composition (Naked NA, Complexes) Relevant sequence modifications or features Relevant local structures
Simulation conditions	ForceField (type and precise version) Simulation length Simulation temperature Solvent and ions Charge settings, added salt Type of trajectory (equilibrium, folding/unfolding, transition) Number of frames, Time per frame
Preliminary Analyses	RMSd, RMSd/bp, R. Gyration Variation, %Lost WC HBonds, %Lost 3D contacts Presence of fraying, Global avg. Roll (degrees), Global avg. Twist (degrees), Groove dimensions (specify measurement method)
Orientative quality Checklist (applicable to equilibrium trajectories)	
Simulation length	> 200 ns
RMSd*	< 5 Å
RMSd/bp*	< 0.3 Å /bp
R. Gyration	< 0.4 Å /bp

Lost H-Bonds SS*	< 20%
Lost 3D contacts*	< 30%
Maintenance of global fold	> 90% simulation time

*RMSd: All-heavy atoms mass weighted. References should be the available experimental structure when available. When not available refer to canonical fiber data.

& Applicable to average values in duplex segments

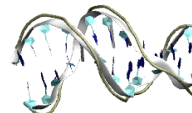
7 Supplementary Figures

7.1 Figure S1. BIGNASim portal screenshot. Expanded view of simulation summary.

A) Simulated system details; B) Simulation details and visualization; C) Access to analysis pages; D) Download trajectories or meta-trajectories

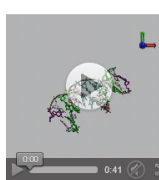
A Nucleic Acid Data:

Sequence	CGCGAATTCGCG
Rev. Sequence	CGCGAATTCGCG
Type	DNA
SubType	B
Chains	duplex
Pdb	1NAJ [PDB] [NUCDB]
Ligands	No
Keywords	DNA-B Duplex Naked DDD ParmBSC0 TI P3P Electroneutral



B MD Simulation >> (Click to expand/shrink)

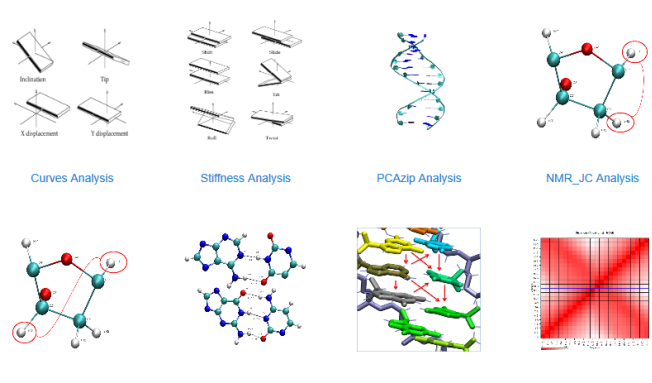
Simulation Metadata	
Force Field	parmBSC1
Simulation Date	05/07/15
Simulated Time	800
Time Step	1 ps
Parts	DNA+ions
Temperature	298K
Water	TIP3P
Additional Solvent	No
Counter Ions	Na+
Ionic Concentration	Electroneutrality
Additional Ions	No
Additional Molecules	No
Ions Parameters	Dang



Quality Control	
RMSd	1.787 (0.353) Å
Rgyr	13.380 (0.261) Å
SASA	4383.518 (69.231) Å ²
RMSf	

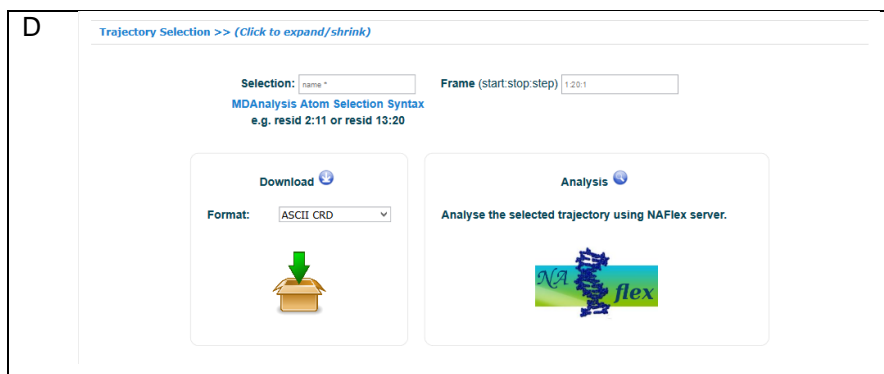
Unified Molecular Modeling (UMM) Metadata
(Click to see full UMM)

C Trajectory Analyses >> (Click to expand/shrink)



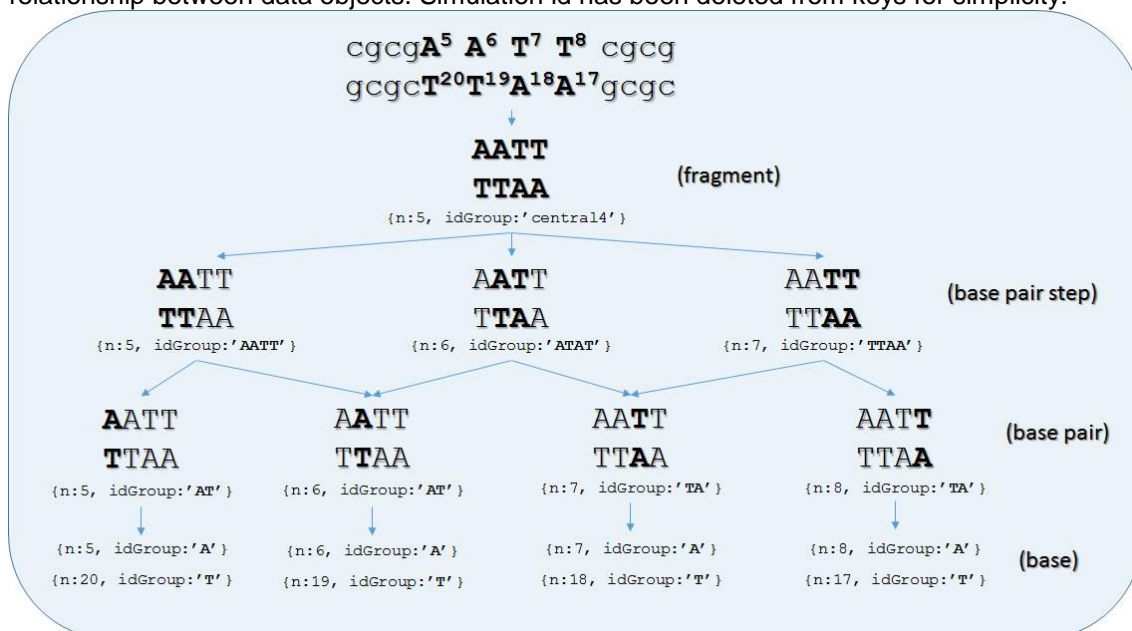
Curves Analysis Stiffness Analysis PCAzip Analysis NMR_JC Analysis

NMR NOEs Analysis HBs Analysis Stacking Analysis Contacts Analysis



7.2 Figure S2. Example of fragment definition on the analysis database.

Database entries in groupDef collection derived from the central tetramer of a Drew-Dickerson dodecamer. Primary keys of each data item are indicated. Arrows indicate a “container” relationship between data objects. Simulation id has been deleted from keys for simplicity.



8 References

1. Hospital, A., Faustino, I., Collepardo-Guevara, R., Gonzalez, C., Gelpi, J.L. and Orozco, M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Research*, **41**, W47-W55.
2. Michaud-Agrawal, N., Denning, E.J., Woolf, T.B. and Beckstein, O. (2011) Software News and Updates MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *Journal of Computational Chemistry*, **32**, 2319-2327.
3. Lakshman, A. and Malik, P. (2010), SIGOPS Oper. Syst. Rev, Vol. 44, pp. 35-40.
4. Hernandez, R., Cugnasco, C., Becerra, Y., Torres, J. and Ayguade, E. (2015), Proceedings of the 23rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, pp. 288-295.
5. Hernandez, R., Becerra, Y., Torres, J. and Ayguade, E. (2015), Proceedings of the International Conference on Computational Science, ICCS 2015, pp. 2822-2826.
6. Ison, J., Kalas, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S. and Rice, P. (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, **29**, 1325-1332.
7. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25-29.
8. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, **25**, 1251-1255.
9. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, **41**, D456-D463.
10. Dans, P.D., Faustino, I., Battistini, F., Zakrzewska, K., Lavery, R. and Orozco, M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Research*, **42**, 11304-11320.
11. Dans, P.D., Perez, A., Faustino, I., Lavery, R. and Orozco, M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic acids research*, **40**, 10668-10678.
12. Roe, D.R. and Cheatham, T.E., III. (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, **9**, 3084-3095.
13. Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D. *et al.* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**, 845-854.

14. Blanchet, C., Pasi, M., Zakrzewska, K. and Lavery, R. (2011) CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res*, **39**, W68-73.
15. Meyer, T., Ferrer-Costa, C., Perez, A., Rueda, M., Bidon-Chanal, A., Luque, F.J., Laughton, C.A. and Orozco, M. (2006) Essential dynamics: A tool for efficient trajectory compression and management. *Journal of Chemical Theory and Computation*, **2**, 251-258.
16. Camps, J., Carrillo, O., Emperador, A., Orellana, L., Hospital, A., Rueda, M., Cicin-Sain, D., D'Abramo, M., Lluís Gelpi, J. and Orozco, M. (2009) FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics*, **25**, 1709-1710.
17. Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling*, **14**, 33-38.