ccpB

Collaborative Computational Project for Biomolecular Simulation

PCA-based trajectory file compression

Introduction

This document outlines the theory behind MD trajectory file compression using principal component analysis (PCA), and the tools developed to implement this that are distributed by CCPB.

Background.

If a trajectory file consists of F snapshots of a system of N atoms, then the total number of (floating point) numbers in the trajectory file will be $3*N*F$.
If the trajectory is subjected to PCA, then it may be represented, exactly, by $3*N$ eigenvectors (each of size $3*N$), plus $3*N*F$ projections, which can regenerate the snapshots by operating on the time-averaged structure of the system ($3*N$ numbers).
But as:

$$3*N*F << (3*N)^2+3*N*F+3*N$$

this is a very inefficient way to represent the data. However, for a typical MD system, PCA will capture the vast majority of the motion of the system in a small number of eigenvectors, M ($M<<3*N$). So representation of the system, to an acceptable level of accuracy, becomes attractive in a PCA-based format if:

$$3*N*F > 3*N*M+M*F+3*N$$

For a typical biomolecular simulation - say 5000 snapshots of 700 atoms, we might expect the PCA analysis would be able to capture 90% of the variance in, say, 35 eigenvectors. In which case:

$$3*N*F = 10{,}500{,}000 \text{ numbers}$$
$$3*N*M+M*F+3*N = 250{,}600 \text{ numbers}$$

i.e., we have a compression to 2.3% of the original file size.

For a full discussion of this approach in practice, see: Mayer et al., *J. Chem. Theor. Comp.* 2006, **2**, 251-258.

PCAZIP

NAME

pcazip – compress MD trajectory files

SYNOPSIS

pcazip –i trajfile –o pczfile –n natoms [-v] [-q quality] [-e eigenvectors] [-mask maskfile] [-nofast] [-formatted][-nofit] [-lowmem] [-help]

OR:

pcazip –a albfile –o pczfile –n natoms [-v] [-q quality] [-e eigenvectors] [-mask maskfile] [-nofast][-formatted][-nofit] [-lowmem] [-help]

DESCRIPTION

Pcazip compresses trajectory files using the PCA method, as described in: Mayer et al., *J. Chem. Theor. Comp.* 2006, **2**, 251-258.

REQUIRED

Either:

-i trajfile

The MD trajectory file to be compressed. The program uses the VMD molfile plugins, so any format supported by these should read in OK.

Or:

-a albfile

An ‘album’ is simply a text file listing, one per line, the names of trajectory files to include in the analysis. Mixtures of different format files are permitted, as long as each contains the same number of atoms in the same order. The filenames can optionally be followed by a frame selection. So if one of the files listed in *albfile* is called “traj1.x”, then if the line containing it reads “traj1.x(23:50)” then just the 23rd through 50th snapshots in traj1.x will be used. Other acceptable syntaxes are:

traj1.x(n)	selects just snapshot *n*
traj1.x(:e)	selects snapshots 1-*e*
traj1.x(b:)	selects snapshots from *b* to the end of the file
traj1.x(b:e:s)	selects every *s*-th snapshot from *b* to *e*
traj1.x(::s)	selects every *s*-th snapshot in the file

Note that this should work for any trajectory file format

Plus:

-o pczfile

The output compressed file. No suffix for this is enforced, though ‘.pcz’ is encouraged, for consistency/transparency.

-n natoms

The number of atoms in one frame of the trajectory file *trajfile*. Unfortunately there is no foolproof way to determine this from examination of *trajfile* itself for all supported formats (e.g. AMBER *mdcrd*). *Pcazip* can however automatically detect the presence or otherwise of periodic box information.

OPTIONS

-v

Verbose option. Various information messages are printed (to standard error) as the compression procedure progresses.

-q quality

Quality option. By default, *pcazip* compresses files with enough eigenvectors to capture 90% of the variance in the data. This can be overridden by specifying *quality* (as an integer 1-99, i.e. 1%-99% of variance captured).

-e eigenvectors

Fixed number of eigenvectors option. This overrides the selection of the number of eigenvectors to use in the compression based on a quality measure. Exactly *eigenvectors* (specified as an integer <= 3**natoms*) eigenvectors will be used.

-mask maskfile

Only include the subset of atoms present in *maskfile*, not every atom in *trajfile*, in the compressed file. *Maskfile* should be a PDB-format file. *Pcazip* will use the atom numbers in this – the second field in each ATOM record – as indices into the coordinate arrays in *trajfile*. So it important that these are correct! The philosophy behind this approach is that most modelling packages will include utilities to create correctly-numbered PDB format files, and simple editing of these, retaining only the desired ATOM records, provides a simple way of selecting the required subset of atoms, and makes a visual check of this easy too. The mask files are also of use with the *pczdump* utility.

-nofast

If the number of frames in the trajectory, *F*, is less than *3*natoms*, then only (*F-1)* eigenvectors can be obtained. In such cases the PCA can sometimes be speeded up by diagonalising the *(F-1)x(F-1)* covariance matrix, rather than the conventional *3*natoms*x*3*natoms* one. *Pcazip* will automatically switch to ‘fastpca’ mode in suitable circumstances, but ‘slow’ diagonalisation can be enforced, if wanted, using this flag.

-formatted

By default, (as of pcazip version 3.0) the output files are in a compact binary format. Including this flag produces pcazip files in a simple ascii format, much less compact but maybe useful for input into other programs.

-nofit

Bypasses the least-squares fitting step that by default is performed before the covariance matrix is calculated. The assumption is that the user has prepared the input trajectory using his or her own favourite fitting process.

-lowmem

Pcazip can be quite memory-hungry – using this option saves some memory by writing intermediate information to a scratch file.

-help

Prints a quick summary of options, and exits.

SEE ALSO

pcaunzip, pczdump, pczcomp, pczformat, quickmask

DIAGNOSTICS/BUGS

Many and varied. *Pcazip* will complain about missing input files, overwriting existing output files, or damaged trajectory files. But the error catching is far from complete, so expect more-or-less obvious system error messages…

EXAMPLES

pcazip –i trajfile.traj –o pczfile.pcz –n 760

Compress *trajfile.traj*, which contains 760 atoms per snapshot, producing the compressed file *pczfile.pcz*. Enough eigenvectors will be included to capture 90% of the variance.

pcazip –i trajfile.traj –o pczfile.pcz –n 760 –v

As above, only produce progress messages as well.

pcazip –i trajfile.traj –o pczfile.pcz –n 760 –q 95

As above, only include enough eigenvectors to capture 95% of the variance.

pcazip -i trajfile.traj –o pczfile.pcz –n 760 –e 20

As above, only include exactly 20 eigenvalues in *pczfile.pcz*, whatever percentage of the variance this will capture.

pcazip –i trajfile.traj –o pczfile.pcz –n 760 –q 95 –mask maskfile.pdb

Only include atoms present in *maskfile.pdb* in the compression procedure, and include enough eigenvectors to capture 95% of the variance. Be sure the atom numbers in *maskfile.pdb* are correct, and note that the '*-n*' argument specifies the number of atoms in the original *trajfile.traj*, not the number in the selected subset.

AUTHOR

Charlie Laughton 2006, 2008,2011

PCAUNZIP

NAME

pcaunzip – uncompress MD trajectory files that have been compressed with *pcazip*

SYNOPSIS

pcaunzip –i pczfile [–o trajfile] [-format fmt] [-iv1 v1 [-iv2 v2]]

DESCRIPTION

Pcaunzip decompresses trajectory files that have been compressed using *pcazip*. Note that because the compression process is (adjustably) 'lossy', uncompressed files will not be identical to the original trajectory file.

REQUIRED

-i pczfile

The compressed ('pcz format') MD trajectory file.

OPTIONS

-o trajfile

The uncompressed trajectory file. Without this option, the file is written to standard output. The default format is Amber *mdcrd*, but Charmm *dcd* is also supported (see below).

-format fmt

If fmt is "charmm" the output is a *dcd* format file, instead of *mdcrd*.

-iv1 v1 [-iv2 v2]

Selects a subset of the available eigenvectors for the uncompression.

SEE ALSO

pcazip, pczdump, pczcomp, pczformat, quickmask

DIAGNOSTICS/BUGS

Pcaunzip will complain about missing or wrong format input files, or overwriting existing output files. But the error catching is far from complete, so expect more-or-less obvious system error messages…

EXAMPLES

pcaunzip –i pczfile.pcz –o trajfile.traj

Uncompress *pczfile.pcz* to *trajfile.traj.*

pcaunzip –i pczfile.pcz > trajfile.binpos

Equivalent to the above.

pcaunzip –i pczfile –iv1 1 –format charmm -o trajfile_e1.dcd

Only uses the first eigenvector, and output is in dcd format

pcaunzip –i pczfile –iv1 1 -iv2 8 -o trajfile_e1-8.dcd

Only uses the first eight eigenvectors.

AUTHOR

Charlie Laughton 2006, 2008, 2011, Yiming Chen (2011)

PCZDUMP

NAME

pczdump – extract various data from a MD trajectory file compressed using pcazip

SYNOPSIS

pczdump –i pczfile [-o outfile] [-info] [-avg [-pdb pdbref]] [-evals] [-evec iv] [-proj iv] [-fluc iv] [-anim iv [-pdb pdbref]][-rms iref] [-maha nv] [-coll] [-help]

DESCRIPTION

Pczdump extracts various pieces of information from MD trajectory files compressed using pcazip.

REQUIRED

-i pczfile

The compressed (‘pcz format’) MD trajectory file to be analysed

OPTIONS

-o outfile

Write output to *outfile*. Without this option, data are written to standard output. The format of the information in *outfile* will depend on the choice of options (below).

-info

Print basic information about the data in *pczfile*. This includes the title in the original MD trajectory file, the numbers of atoms, frames and eigenvectors, and the quality (%age of variance captured).

-avg [-pdb pdbref]

Output the time averaged structure, in AMBER/g86 trajectory file format, or in PDB format if the optional *–pdb* argument is present. In this case *pdbfile* is used as a template for the output – i.e. it is important that the identity and ordering of the ATOM records in this is correct, but the coordinates themselves are ignored.

-evals

Print out the eigenvalues associated with each eigenvector present in *pczfile*, in order of decreasing magnitude.

-evec iv

Print out the *iv*th eigenvector. The most important (largest eigenvalue) eigenvector is the first.

-proj iv

Print out the projections of the *iv*th eigenvector (1 value for each frame in the trajectory)

-fluc iv

Print out the atomic fluctuations associated with the *iv*th eigenvector.

-anim iv [-pdb pdbref]

Produce a file animating the motion of the molecule along the *iv*th eigenvector/principal component. This will be in AMBER/g86 trajectory format, unless the optional *–pdb* argument is given as well, in which case it will be in multi-model pdb format. See the *–avg* option above for details regarding how *pdbref* is used.

-rms iref

Print out the rmsd of each frame in *pczfile* from the structure in snapshot *iref* (specified as an integer in the range 0 to the number of frames). Note that depending on the quality setting used in the original compression, values calculated this way will not exactly match those obtained from calculations done on the original data, but for typical quality settings (90% or greater) the difference is small. As a special case, if *iref* is set to 0, rmsds from the time-average structure will be output.

-maha nv

Prints out the Mahalanobis distance of each frame in the trajectory from the time-averaged structure, as calculated over modes 1 to *nv*.

-coll

Prints out the collectivity metric, κ, for each eigenvector (see Brunschweiler et al, *J. Chem. Phys.*, 1995, **102**, 3396-3403). Modes that produce the most collective motion in the system will have high κ values, while modes that, for example, originate from large displacements of small substructures, will show low κ values.

-help

Prints a quick summary of options, and exits.

SEE ALSO

Pcazip, pcaunzip, pczcomp, pczformat, quickmask

DIAGNOSTICS/BUGS

Many and varied. *Pczdump* will complain about missing or wrong format input files, overwriting existing output files, etc. But the error catching is far from complete, so expect more-or-less obvious system error messages…

EXAMPLES

pczdump –i pczfile.pcz –info

Output (to the screen) basic information about the contents of *pczfile.pcz.*

pczdump –i pczfile.pcz –o average.pdb –avg –pdb ref.pdb

Extract the time average structure from *pczfile.pcz*, writing to *average.pdb* in pdb format, using *ref.pdb* as a template.

pczdump –i pczfile.pczj –o rms_from_avg.dat –rms 0

Calculate the rmsd between each frame in the MD trajectory and the time-averaged structure.

pczdump –i pczfile.pczj –o rms_from_first.dat –rms 1

Calculate the rmsd between each frame in the MD trajectory and the first frame.

Pczdump –i pczfile.pcz –anim 2 –pdb ref.pdb –o pc2_animation.pdb

Produce a multi-model pdb format file *pc2_animation.pdb* that animates the motion of the structure, about its time averaged position, along the second principal component.

AUTHOR

Charlie Laughton 2006, 2008, 2011

PCZCOMP

NAME

pczcomp – compare two MD trajectories, compressed using pcazip

SYNOPSIS

pczcomp -x pczfile1 –y pczfile2 [-nv eigenvectors]

DESCRIPTION

A very simple utility to do a basic comparison between two MD trajectory files that have been compressed using *pcazip*. The output contains the following information:

1. The RMSD between the two time-average structures contained in *pczfile1* and *pczfile2*.
2. The Mahalanobis distance of the time-averaged structure in *pczfile2* from that in *pczfile1*, and vice versa (for a discussion of Mahalanobis distances, see http://en.wikipedia.org/wiki/Mahalanobis_distance).
3. The dot product matrix of the top *n* eigenvectors in *pczfile1* (x-axis), with those in *pczfile2* (y-axis). By default *n* is 10, but this can be overridden.
4. The subspace overlap.

REQUIRED

-x pczfile1

The first compressed (‘pcz format’) MD trajectory file.

-y pczfile2

The second compressed (‘pcz format’) MD trajectory file.

OPTIONS

-nv eigenvectors

Use *eigenvectors* (specified as an integer less than or equal to the number of eigenvectors stored in *pczfile1* or *pczfile2*, whichever is the smaller) eigenvectors in the calculation of Mahalanobis distances, the dot product matrix, and the subspace overlap.

SEE ALSO

pcazip, pcaunzip, pczdump, pczformat, quickmask

DIAGNOSTICS/BUGS

Pczcomp will check that the two .pcz files have the same number of atoms per snapshot, but otherwise no test is made as to whether the comparison is going to be meaningful or not. *Pczcomp* will complain about missing or wrong format input files, but the error catching is far from complete, so expect more-or-less obvious system error messages…

EXAMPLES

Pczcomp -x pczfile1.pcz –y pczfile2.pcz –o compare.dat –nv 6

Compare *pczfile1.pcz* with *pczfile2.pcz*, using only six eigenvectors from each. Write the results to *compare.dat*.

AUTHOR

Charlie Laughton 2006, 2008, 2011

QUICKMASK

NAME

quickmask – utility to process trajectory files

SYNOPSIS

quickmask –i oldtrajfile –o newtrajfile –n natoms [-mask maskfile][-v]

OR:

quickmask –a albfile –o newtrajfile –n natoms [-mask maskfile][-v]

DESCRIPTION

Quickmask is a little utility to concatenate and/or strip down trajectory files. All the same functionality is available within *pcazip*, except the ability to just write out the resulting trajectory file, so *quickmask* is a (maybe) useful extra rather than a necessary component of the package.

REQUIRED

Either:

-i oldtrajfile

The existing trajectory file (any VMD molfile-supported format)

Or:

-a albfile

An ‘album’ is simply a text file listing, one per line, the names of trajectory files to include in the analysis. Mixtures of different format files are permitted, as long as each contains the same number of atoms in the same order. The filenames can optionally be followed by a frame selection. So if one of the files listed in *albfile* is called “traj1.x”, then if the line containing it reads “traj1.x(23:50)” then just the 23rd through 50th snapshots in traj1.x will be used. Other acceptable syntaxes are:

traj1.x(n)	selects just snapshot *n*
traj1.x(:e)	selects snapshots 1-*e*
traj1.x(b:)	selects snapshots from *b* to the end of the file
traj1.x(b:e:s)	selects every *s*-th snapshot from *b* to *e*
traj1.x(::s)	selects every *s*-th snapshot in the file

Note that this should work for any trajectory file format

Plus:

-o newtrajfile

The processed trajectory file. Scripps *binpos* format.

-n natoms

Number of atoms per snapshot

OPTIONS

-mask maskfile

Only include the subset of atoms present in *maskfile*, not every atom in *oldtrajfile,* in the compressed file. *Maskfile* should be a PDB-format file. *Quickmask* will use the atom numbers in this – the second field in each ATOM record – as indices into the coordinate arrays in *oldtrajfile*. So it important that these are correct! The philosophy behind this approach is that most modelling packages will include utilities to create correctly-numbered PDB format files, and simple editing of these, retaining only the desired ATOM records, provides a simple way of selecting the required subset of atoms, and makes a visual check of this easy too.

-v

Verbose diagnostics.

SEE ALSO

pcazip, pcaunzip, pczdump, pczcomp, pczformat

DIAGNOSTICS/BUGS

Quickmask will complain about missing or wrong format input files, or overwriting existing output files. But the error catching is far from complete, so expect more-or-less obvious system error messages…

EXAMPLES

quickmask –i full.dcd –n 825 –o CAonly.binpos –mask CAonly.pdb

Selects a subset of atoms from *full.dcd*

quickmask –i "full.dcd(50:2000:5)" –n 825 –o CAonly.binpos –mask CAonly.pdb

Selects a subset of atoms from *full.dcd,* choosing every 5th frame from frames 50 to 2000. (note use of quotes).

quickmask –a album.alb –n 825 –o concatenated.binpos

Combines the trajectory files listed in *album.alb* (which may be a mixture of formats) and writes a single output trajectory file.

AUTHOR

Charlie Laughton 2011

PCZFORMAT

NAME

Pczformat – the format and contents of the files produced by *pcazi*p

DESCRIPTION

There are three pcazip file formats (as of pcazip version 3.0):
a) PCZ0: a simple formatted format, produced by *pcazip* with the '*-formatted*' flag.
b) PCZ2: the binary format for previous versions of *pcazip*, now obsolete but still readable by *pcaunzip*, *pczdump* and *pczcomp*.
c) PCZ4: the default version produced by the current version of *pcazip* (4.0). This is a 'pure' binary stream format.

The PCZ2 file format

Files in PCZ2 format are fortran direct access binary files. They consist of one header block, followed by one data block for each eigenvector stored. All blocks are of equal size. Eigenvectors are stored in order of decreasing eigenvalue.

The header block is organised as follows:

a) The format identifier ('PCZ2') (4 bytes)
b) The title, taken from the trajectory file, (80 bytes)
c) The number of atoms in a snapshot (4 byte integer)
d) The number of snapshots/frames in the trajectory (4 byte integer)
e) The number of eigenvectors stored (4 byte integer)
f) The total of ALL the eigenvalues that are obtained when the original covariance matrix was diagonalised – not just the sum of those present in this file (this allows the quality of the file to be confirmed, even if it has become truncated) (4 byte float)
g) The time-average structure (3*no. of atoms: 4 byte floats)
h) Enough padding to extend the size of this block to that of the data blocks.

A data block is organised as follows:

a) The coefficients of the eigenvector (3*no. of atoms: 4 byte floats)
b) The eigenvalue (4 byte float)
c) The projections (one 4 byte float for each snapshot)

In the unlikely case that the length of a data block is less than that of the unpadded header block, the data blocks are padded instead.

Note that as these are fortran-compatible direct i/o files, each block (record) is actually 'topped and tailed' by a further byte that gives the record length.

The PCZ4 file format

Files in PCZ4 format are pure binary, (little-endian) sequential access files. They begin with a header as follows:

a) 4 byte string: the format identifier ('PCZ4')
b) 80 byte string: the title, taken from the trajectory file, (80 characters)
c) Three 4 byte integers: the number of atoms in a snapshot, snapshots (frames) in the trajectory, number of eigenvectors stored.
d) 4 byte float: the total of ALL the eigenvalues that were obtained when the original covariance matrix was diagonalised – not just the sum of those present in this file (this allows the quality of the file to be confirmed, even if it has become truncated).
e) Three 4-byte integers, reserved for future use

f) 4 byte integer: if >0, indicates that the file contains PDB-style information. If so, this is followed by one 16-byte block for each atom, each of which contains:
    i) 4 byte integer: the atom number
    ii) 4 byte string: the atom name
    iii) 4 byte integer: the residue number
    iv) 3 byte string: the residue name
    v) 1 byte character: the chain identifier

g) The time-average structure. Each of the 3*(no.of atoms) coordinates as a 4 byte float

h) The rest of the file consists of blocks, one per eigenvector stored. Each block contains the following data as 4 byte floats:

    i) The coefficients of the eigenvector (3*no. of atoms)
    ii) The eigenvalue
    iii) The projections of that eigenvector (one value for each snapshot)

SEE ALSO

pcazip, pcaunzip, pczdump, pczcomp,quickmask

AUTHOR

Charlie Laughton 2006, 2008, 2011