NAME

theseus − Maximum likelihood, multiple simultaneous superpositions with statistical analysis

SYNOPSIS

theseus [−aAbBcCdDeEfFgGhHiIjklLmMnNoOpPqQrRsStTuvVwWxXyYZ] *pdbfile1 [pdbfile2 ...]*

and

theseus_align [−aAbBcCdDeEfFgGhHiIjklLmMnNoOpPqQrRsStTuvVwWxXyYZ] −f *pdbfile1 [pdbfile2 ...]*

Default usage is equivalent to:

theseus −a0 −e2 −g1 −i200 −k-1 −p1e-7 −r theseus −v −P0 *your.pdb*

DESCRIPTION

Theseus superpositions a set of macromolecular structures simultaneously using the method of maximum likelihood (ML), rather than the conventional least-squares criterion. **Theseus** assumes that the structures are distributed according to a matrix Gaussian distribution and that the eigenvalues of the atomic covariance matrix are hierarchically distributed according to an inverse gamma distribution. This ML superpositioning model produces much more accurate results by essentially downweighting variable regions of the structures and by correcting for correlations among atoms.

Theseus operates in two main modes, a mode for superimposing structures with identical sequences and a mode for structures with different sequences but similar structures:

(1) A mode for superpositioning macromolecules with identical sequences and numbers of residues, for instance, multiple models in an NMR family or multiple structures from different crystal forms of the same protein. In this mode, **Theseus** will read every model in every file on the command line and superposition them.

Example:

theseus *1s40.pdb*

In the above example, *1s40.pdb* is a pdb file of 10 NMR models.

(2) An "alignment" mode for superpositioning structures with different sequences, for example, multiple structures of the cytochrome c protein from different species or multiple mutated structures of hen egg white lysozyme. This mode requires the user to supply a sequence alignment file of the structures being superpositioned (see option **−A** and "FILE FORMATS" below). Additionally, it may be necessary to supply a mapfile that tells **theseus** which PDB structure files correspond to which sequences in the alignment (see option **−M** and "FILE FORMATS" below). When superpositioning based on a seqeunce alignment, **theseus** uses a novel maximum likelihood algorithm for superpositioning multiple structures that include arbitrary gaps and insertions relative to each other. Unlike other algorithms for simultaneous superpositioning of multiple structures, our Expectation-Maximization algorithm uses all available data by including all residues aligned with gaps in the calculations. In this mode, if there are multiple structural models in a PDB file, **theseus** only reads the first model in each file on the command line. In other words, **theseus** treats the files on the command line as if there were only one structure per file.

Example 1:

theseus −A cytc.aln −M cytc.filemap d1cih__.pdb d1csu__.pdb d1kyow_.pdb

In the above example, d1cih__.pdb, d1csu__.pdb, and d1kyow_.pdb are pdb files of cytochrome c domains from the SCOP database.

Example 2:

theseus_align −f d1cih__.pdb d1csu__.pdb d1kyow_.pdb

In this example, the **theseus_align** script is called to do the hard work for you. It will calculate a sequence alignment and then superimpose based on that alignment. The script **theseus_align** takes the same options as the **theseus** program. Note, the first few lines of this script must be modified for your system, since it calls an external multiple sequence alignment program to do the alignment. See the **examples/** directory for more details, including example files.

OPTIONS

Algorithmic options, defaults in {brackets}:

−a [*selection***]**

Atoms to include in the superposition. This option takes two types of arguments, either (1) a number specifying a preselected set of atom types, or (2) an explict PDB-style, colon-delimited list of the atoms to include.

For the preselected atom type subsets, the following integer options are available:

- 0, alpha carbons for proteins, C1´ atoms for nucleic acids
- 1, backbone
- 2, all
- 3, alpha and beta carbons
- 4, all heavy atoms (no hydrogens)

Note, only the **−a0** option is available when superpositioning structures with different sequences.

To custom select an explicit set of atom types, the atom types must be specified exactly as given in the PDB file field, including spaces, and the atom-types must encapsulated in quotation marks. Multiple atom types must be delimited by a colon. For example,

−a' N : CA : C : O '

would specify the atom types in the peptide backbone.

−c Use ML atomic covariance weighting (fit correlations, much slower)

Unless you have many different structures with few residues, fitting the correlation matrix is likely unwarranted statistically due to a plethora of parameters and a paucity of data.

−e [*n***]** Embedding algorithm for initializing the average structure
- 0 = none; use randomly chosen model
- {2} = {ML embedded structure}

−f Only read the first model of a multi-model PDB file

−g [*n***]** Hierarchical model for variances
- 0 = none (may not converge)
- {1} = inverse gamma distribution

−h Help/usage

−i [*nnn***]**
Maximum iterations, {200}

−k [*n***]** constant minimum variance {−1} {if set to negative value, the minimum variance is determined empirically}

−p [*precision***]**
Requested relative precision for convergence, {1e−7}

−r [*root name***]**
Root name to be used in naming the output files, {theseus}

−s [*n−n:...***]**
Residue selection (e.g. −s15−45:50−55), {all}

−S [*n−n:...***]**
Residues to exclude (e.g. −S15-45:50−55) {none}

The previous two options have the same format. Residue (or alignment column) ranges are indicated by beginning and end separated by a dash. Multiple ranges, in any arbitrary order, are separated by a colon. Chains may also be selected by giving the chain ID immediately preceding the residue range. For example, **−sA1−20:A40−71** will only include residues 1 through 20 and 40 through 70 in chain A. Chains cannot be specified when superpositioning structures with different sequences.

−v use ML variance weighting (no correlations) {default}

Input/output options:

−A [*sequence alignment file***]**
Sequence alignment file to use as a guide (CLUSTAL or A2M format)

For use when superpositioning structures with different sequences. See "FILE FORMATS" below.

−E Print expert options

−F Print FASTA files of the sequences in PDB files and quit

A useful option when superpositioning structures with different sequences. The files output with this option can be aligned with a multiple sequence alignment program such as CLUSTAL or MUSCLE, and the resulting output alignment file used as **theseus** input with the **−A** option.

−h Help/usage

−I Just calculate statistics for input file; don't superposition

−M [*mapfile***]**
File that maps PDB files to sequences in the alignment.

A simple two-column formatted file; see "FILE FORMATS" below. Used with mode 2.

−n Don't write transformed pdb file

−o [*reference structure***]**
Reference file to superposition on, all rotations are relative to the first model in this file

For example, 'theseus −o cytc1.pdb cytc1.pdb cytc2.pdb cytc3.pdb' will superposition the structures and rotate the entire final superposition so that the structure from cytc1.pdb is in the same orientation as the structure in the original cytc1.pdb PDB file.

−O Olve's segID file

Useful output when superpositioning structures with different sequences (mode 2). In 'theseus_sup.pdb', the main output superposition PDB file, the segID field now holds the number of the sequence alignment column that it belongs to. This number, divided by 100, is also echoed in the B-factor field. When using **O** (or any other capable molecular visualization program), one can then color by B-factor ranges and immediately see in the superposition which regions of the structure are aligned in the sequence alignment file. An additional file is also output, called 'theseus_olve.pdb' which only contains the very atoms that were included in the ML superposition calculation. That is, it will only contain alpha carbons or phosphorous atoms, and it will only contain atoms from the columns selected with the **−s** or **"−S"** options. Requested by Olve Peersen of Colorado State University.

−V Version

Principal components analysis:

−C Use covariance matrix for PCA (correlation matrix is default)

−P [*nnn***]**
Number of principal components to calculate {0}

In both of the above, the corresponding principal component is written in the B-factor field of the output PDB file. Usually only the first few PCs are of any interest (maybe up to six).

EXAMPLES **theseus** *2sdf.pdb*

theseus −l −r new2sdf *2sdf.pdb*

theseus −s15−45 −P3 *2sdf.pdb*

theseus −A *cytc.aln* −M *cytc.mapfile* −o *cytc1.pdb* −s1−40 *cytc1.pdb cytc2.pdb cytc3.pdb cytc4.pdb*

ENVIRONMENT

You can set the environment variable 'PDBDIR' to your PDB file directory and **theseus** will look there after the present working directory. For example, in the C shell (tcsh or csh), you can put something akin to this in your .cshrc file:

setenv PDBDIR '/usr/share/pdbs/'

FILE FORMATS

Theseus will read standard PDB formatted files (see <http://www.rcsb.org/pdb/>). Every effort has been made for the program to accept nonstandard CNS and X-PLOR file formats also.

Two other files deserve mention, a sequence alignment file and a mapfile.

Sequence alignment file

When superpositioning structures with different residue identities (where the lengths of each the macromolecules in terms of residues are not necessarily equal), a sequence alignment file must be included for **theseus** to use as a guide (specified by the **−A** option). **Theseus** accepts both CLUSTAL and A2M (FASTA) formatted multiple sequence alignment files.

NOTE 1: The residue sequence in the alignment must match exactly the residue sequence given in the coordinates of the PDB file. That is, there can be no missing or extra residues that do not correspond to the sequence in the PDB file. An easy way to ensure that your sequences exactly match the PDB files is to generate the sequences using **theseus' −F** option, which writes out a FASTA formatted sequence file of the chain(s) in the PDB files. The files output with this option can then be aligned with a multiple sequence alignment program such as CLUSTAL or MUSCLE, and the resulting output alignment file used as **theseus** input with the **−A** option.

NOTE 2: Every PDB file must have a corresponding sequence in the alignment. However, not every sequence in the alignment needs to have a corresponding PDB file. That is, there can be extra sequences in the alignment that are not used for guiding the superposition.

PDB −> Sequence mapfile

If the names of the PDB files and the names of the corresponding sequences in the alignemnt are identical, the mapfile may be omitted. Otherwise, **Theseus** needs to know which sequences in the alignment file correspond to which PDB structure files. This information is included in a mapfile with a very simple format (specified with the **−M** option). There are only two columns separated by whitespace: the

first column lists the names of the PDB structure files, while the second column lists the corresponding sequence names exactly as given in the multiple sequence alignment file.

An example of the mapfile:

```
cytc1.pdb   seq1
cytc2.pdb   seq2
cytc3.pdb   seq3
```

SCREEN OUTPUT

Theseus provides output describing both the progress of the superpositioning and several statistics for the final result:

Least-squares <sigma>:
The standard deviation for the superposition, based on the conventional assumption of no correlation and equal variances. Basically equal to the RMSD from the average structure.

Classical LS pairwise <RMSD>:
The conventional RMSD for the superposition, the average RMSD for all pairwise combinations of structures in the ensemble.

Maximum Likelihood <sigma>:
The ML analog of the standard deviation for the superposition. When assuming that the correlations are zero (a diagonal covariance matrix), this is equal to the square root of the harmonic average of the variances for each atom. In contrast, the 'Least-squares <sigma>' given above reports the square root of the arithmetic average of the variances. The harmonic average is always less than the arithmetic average, and the harmonic average downweights large values proportional to their magnitude. This makes sense statistically, because when combining values one should weight them by the reciprocal of their variance (which is in fact what the ML superpositioning method does).

Log Likelihood:
The final log likelihood of the superposition, assuming the matrix Gaussian distribution of the structures and the hierarchical inverse gamma distribution of the eigenvalues of the covariance matrix.

AIC: The Akaike Information Criterion for the final superposition. This is an important statistic in likelihood analysis and model selection theory. It allows an objective comparison of multiple theoretical models with different numbers of parameters. In this case, the higher the number the better. There is a tradeoff between fit to the data and the number of parameters being fit. Increasing the number of parameters in a model will always give a better fit to the data, but it also increases the uncertainty of the estimated values. The AIC criterion finds the best combination by (1) maximizing the fit to the data while (2) minimizing the uncertainty due to the number of parameters. In the superposition case, one can compare the least squares superposition to the maximum likelihood superposition. The method (or model) with the higher AIC is preferred. A difference in the AIC of 2 or more is considered strong statistical evidence for the better model.

BIC: The Bayesian Information Criterion. Similar to the AIC, but with a Bayesian emphasis.

Rotational, translational, covar chiˆ2:
The reduced chi-squared statistic for the fit of the structures to the model. With a good fit it should be close to 1.0, which indicates a perfect fit of the data to the statistical model. In the case of least-squares, the assumed model is a matrix Gaussian distribution of the structures with equal variances and no correlations. For the ML fits, the assumed models can either be (1) unequal variances and no correlations, as calculated with the **−v** option [default] or (2) unequal variances and correlations, as calculated with the **−c** option. This statistic is for the

superposition only, and does not include the fit of the covariance matrix eigenvalues to an inverse gamma distribution. See 'Omnibus chi^2' below.

Hierarchical minimum var:

The hierarchical fit of the inverse gamma distribution constrains the variances of the atoms by making large ones smaller and small ones larger. This statistic reports the minimum possible variance given the inferred inverse gamma parameters.

Hierarchical var (alpha, gamma) chi^2:

The reduced chi-squared for the inverse gamma fit of the covariance matrix eigenvalues. As before, it should ideally be close to 1.0. The two values in the parentheses are the ML estimates of the scale and shape parameters, respectively, for the inverse gamma distribtuion.

Omnibus chi^2:

The overall reduced chi-squared statistic for the entire fit, including the rotations, translations, covariances, and the inverse gamma parameters. This is probably the most important statistic for the superposition. In some cases, the inverse gamma fit may be poor, yet the overall fit is still very good. Again, it should ideally be close to 1.0, which would indicate a perfect fit. However, if you think it is too large, make sure to compare it to the chi^2 for the least-squares fit; it's probably not that bad after all. A large chi^2 often indicates a violation of the assumptions of the model. The most common violation is when superpositioning two or more independent domains that can rotate relative to each other. If this is the case, then there will likely be not just one Gaussian distribution, but several mixed Gaussians, one for each domain. Then, it would be better to superposition each domain independently.

skewness, skewness Z-value, kurtosis & kurtosis Z-value:

The skewness and kurtosis of the residuals. Both should be 0.0 if the residuals fit a Gaussian distribution perfectly. They are followed by the P-value for the statistics. This is a very stringent test; residuals can be very non-Gaussian and yet the estimated rotations, translations, and covariance matrix may still be rather accurate.

FP error in transformed coordinates:

The empirically determined floating point error in the coordinates after rotation and translation.

Minimum RMSD error per atom:

The empirically determined minimum RMSD error per atom, based on the floating point error of the computer.

Data pts, Free params, D/P:

The total number of data points given all observed structures, the number of parameters being fit in the model, and the data-to-parameter ratio.

Median structure:

The structure that is overall most similar to the average structure. This can be considered to be the most "typical" structure in the ensemble.

Total rounds:

The number of iterations that the algorithm took to converge.

Fractional precision:

The actual precision that the algorithm converged to.

OUTPUT FILES

Theseus writes out the following files:

theseus_sup.pdb
The final superposition, rotated to the principle axes of the mean structure.

theseus_ave.pdb
The estimate of the mean structure.

theseus_cor.mat, theseus_cov.mat
The atomic correlation matrix and covariance matrices, based on the final superposition. The format is suitable for input to GNU's **octave.** These are the matrices used in the Principal Components Analysis.

theseus_embed_ave.pdb
The average structure as calculated by S. Lele's EDMA embedding algorithm, used as the starting point for the maximum likelihood iterations.

theseus_residuals.txt
The normalized residuals of the superposition. These can be analyzed for deviations from normality (whether they fit a standard Gaussian distribution). E.g., the chi^2, skewness, and kurtosis statistics are based on these values.

theseus_transf.txt
The final transformation rotation matrices and translation vectors.

theseus_variances.txt
The vector of estimated variances for each atom.

When Principal Components are calculated (with the **−P** option), the following files are also produced:

theseus_pcvecs.txt
The principal component vectors.

theseus_pcstats.txt
Simple statistics for each principle component (loadings, variance explained, etc.).

theseus_pcN_ave.pdb
The average structure with the Nth principal component written in the temperature factor field.

theseus_pcN.pdb
The final superposition with the Nth principal component written in the temperature factor field. This file is omitted when superpositioning molecules with different residue sequences (mode 2).

BUGS

Please send me (DLT) reports of all problems.

RESTRICTIONS

Theseus is *not* a structural alignment program. The structure-based alignment problem is completely different from the structural superposition problem. In order to do a structural superposition, there must be a 1-to-1 mapping that associates the atoms in one structure with the atoms in the other

structures. In the simplest case, this means that structures must have equivalent numbers of atoms, such as the models in an NMR PDB file. For structures with different numbers of residues/atoms, superpositioning is only possible when the sequences have been aligned previously. Finding the best sequence alignment based on only structural information is a difficult problem, and one for which there is currently no maximum likelihood approach. Extending **theseus** to address the structural alignment problem is an ongoing research project.

AUTHOR

Douglas L. Theobald
dtheobald@brandeis.edu
dtheobald@gmail.com

CITATION

When using **theseus** in publications please cite the following:

Douglas L. Theobald and Deborah S. Wuttke (2006)
"Empirical Bayes models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem."
PNAS 103(49):18521-18527

Douglas L. Theobald and Deborah S. Wuttke (2006)
"THESEUS: Maximum likelihood superpositioning and analysis of macromolecular structures."
Bioinformatics 22(17):2171-2172

Douglas L. Theobald and Deborah S. Wuttke (2008)
"Accurate structural correlations from maximum likelihood superpositions."
PLoS Computational Biology 4(2):e43

HISTORY

Long, tedious, and sordid.