Structural bioinformatics

THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures

Douglas L. Theobald* and Deborah S. Wuttke

Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, CO 80309-0215, USA

Received on May 2, 2006; accepted on June 12, 2006
Advance Access publication June 15, 2006
Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: THESEUS is a command line program for performing maximum likelihood (ML) superpositions and analysis of macromolecular structures. While conventional superpositioning methods use ordinary least-squares (LS) as the optimization criterion, ML superpositions provide substantially improved accuracy by down-weighting variable structural regions and by correcting for correlations among atoms. ML superpositioning is robust and insensitive to the specific atoms included in the analysis, and thus it does not require subjective pruning of selected variable atomic coordinates. Output includes both likelihood-based and frequentist statistics for accurate evaluation of the adequacy of a superposition and for reliable analysis of structural similarities and differences. THESEUS performs principal components analysis for analyzing the complex correlations found among atoms within a structural ensemble.

Availability: ANSI C source code and selected binaries for various computing platforms are available under the GNU open source license from http://monkshood.colorado.edu/theseus/ or http://www.theseus3d.org
Contact: douglas.theobald@colorado.edu
Supplementary Information: Supplementary data including details of the ML superpositioning algorithm are available at Bioinformatics online.

1 INTRODUCTION

Superpositioning macromolecular structures is an essential tool in structural bioinformatics and is used routinely in the fields of NMR, X-ray crystallography, protein folding, molecular dynamics, rational drug design and structural evolution (Bourne and Shindyalov, 2003; Flower 1999). Superpositioning allows comparison of structures by fitting their atomic coordinates to each other as closely as possible. The valid interpretation of a superposition relies upon the quality of the estimated orientations of the molecules, and thus reliable and robust superpositioning tools are a critical component of structural analysis and comparison.

The structural superposition problem has classically been solved with the standard statistical optimization method of least-squares (LS) (Flower, 1999). The LS objective is to find the rotations and translations that minimize the squared distances among corresponding atoms in the observed structures. A fundamental justifying assumption of LS (as given in the Gauss–Markov theorem) requires that the errors have equal variance (Seber and Wild, 1989). When this assumption does not hold, a condition known in statistics as heteroscedasticity, LS can provide misleading and inaccurate results. However, the requirement for homogeneous variances is generally violated with macromolecular superpositions. For example, in reported superpositions of multiple NMR protein models the backbone variances commonly range over three orders of magnitude. Similarly, in comparisons of different protein domains belonging to the same fold, the structures deviate from each other with varying degrees of local precision: some atoms ‘superimpose well’ and others do not. LS further requires that the variances be uncorrelated. However, this assumption is also violated in the case of macromolecular superpositions. The variance for each atom is highly correlated with the variances of proximal atoms, owing to linkage resulting from inter-atomic chemical bonds and physical interactions.

To correct for these shortcomings of LS, we have applied the principle of maximum likelihood (ML) to the superposition problem by assuming a Gaussian distribution of the structures in the analysis (Theobald and Wuttke, 2006). ML is widely considered to be fundamental in statistical modeling and parameter estimation (Pawitan, 2001). ML superpositioning requires solving for four types of unknowns: a global covariance matrix describing the variance and correlations for each atom in the structures, a mean structure, and, for each structure in the analysis, a rotation matrix and a translation vector. In the present case, the ML method accounts for uneven variances and correlations in the structures by weighting by the inverse of the atomic covariance matrix. The unknowns are interdependent and cannot be solved analytically. For simultaneous estimation, we use an iterative numerical algorithm for maximizing the joint likelihood (see Supplementary data).

2 IMPLEMENTATION

Our numerical algorithm for calculating ML superpositions is implemented in the command-line UNIX program THESEUS. Rendered output is shown in Figure 1, where a comparison with the LS method clearly shows the increased accuracy of ML superpositions when including all atoms in the calculation. THESEUS works in two modes: (1) a mode for superpositioning structures with identical atoms and (2) an ‘alignment mode’ which can superposition homologous structures with different residues. Note that THESEUS is not a tool for structure-based sequence alignment, which is a separate bioinformatic challenge (Bourne and Shindyalov, 2003). Thus, like all structural superposition methods, THESEUS requires an a priori one-to-one mapping among the atoms/residues in the structures under consideration. When superpositioning multiple
conformations of the same protein (e.g. NMR models or different crystal structures of identical proteins), the one-to-one mapping is trivial. However, when superpositioning different proteins, the user must supply a sequence alignment of the proteins for THESEUS to use as a guide. THESEUS accepts sequence alignments in both CLUSTAL and A2M (FASTA) formats.

There is no limit on the number of structures that THESEUS will superposition (aside from that mandated by the operating system and memory capability). Via simple command line options, users can choose to superposition with the conventional LS method, to select residues (or alignment columns) for inclusion or exclusion from the calculation, and, when superpositioning structures of identical residues (mode 1), to select atom types (e.g. only $\alpha$-carbons or only backbone atoms). THESEUS writes two PDB format files, one of the final superposition and one of the estimate of the mean structure. For easy visualization, the estimated variance for each atom is converted to a ‘pseudo-B-factor’ and written in the temperature factor field of the mean structure file.

In addition to estimating the optimal superposition of multiple structures, THESEUS calculates various frequentist and likelihood-based statistics for evaluating the fit and quality of the superposition, including the conventional least-squares RMSD$_{LS}$, the maximum likelihood RMSD$_{ML}$, and the reduced $\chi^2$ for the overall superposition. The overall absolute likelihood is produced, as well as likelihoodist model selection measures such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Burnham and Anderson, 1998).

Finally, THESEUS will calculate the principal components of the covariance and correlation matrices for analysis of the major modes of correlated conformational differences within a superposition. Each principal component is written into the temperature factor field of two additional files: a superposition of all structures and the estimate of the mean structure. Principal components can then be visualized readily using software that colors the structures according to values in the temperature factor field (Fig. 1C).

When assuming a diagonal covariance matrix (i.e. assuming no correlations), the calculation usually converges in a fraction of a second on modern personal computers for moderate-sized problems (e.g. 50 structures, 100 $\alpha$-carbons) and in a few seconds for larger problems (e.g. 100 structures, 500 $\alpha$-carbons). Calculation of the full atomic covariance matrix can take up to a few minutes for larger problems, as each iteration requires a matrix inversion.

ACKNOWLEDGEMENTS

The authors are grateful to Olve Peersen for extensive bug-testing of THESEUS. The authors thank the NIH for funding (GM59414). D.L.T. is supported by Postdoctoral Fellowship Grant #PF-04-118-01-GMC from the American Cancer Society.

Conflict of Interest: none declared.

REFERENCES


Fig. 1. A conventional LS superposition versus the ML superposition (A and C) of 30 NMR models of the 71 amino acid Kunz domain 2 of Tissue Factor Pathway Inhibitor (PDB ID: 1adz). All C$_\alpha$s were included in the calculations. For the LS superposition, RMSD$_{LS}$ = 4.37, overall reduced $\chi^2$ = 27.9, absolute log likelihood = −9067.0 and AIC = −9139.9. For the ML superposition, RMSD$_{ML}$ = 0.113, the overall reduced $\chi^2$ = 1.01, absolute log likelihood = −1459.3 and AIC = −1906.5. Relative to the ML superposition, AAIC = 7177.8, indicating that the ML model is preferred by a large margin as judged by likelihoodist model selection criteria (P $\leq$ 0.0, Vuong likelihood ratio test) (Burnham and Anderson, 1998; Vuong, 1989). Qualitatively similar results are seen with pairwise superpositions. The first principal component of the ML correlation matrix plotted on the Kunz domain 2 family superposition. The red-colored loops at lower right indicate regions that are strongly correlated within the family, whereas the light blue $\beta$-strands at middle left are modestly anti-correlated with the red regions.