**Bioinformatics Advance Access published March 3, 2005**

# CoC: A database of universally conserved residues in protein folds.

Jason E. Donald [†], Isaac A. Hubner [†], Veronica M. Rotemberg[†], Eugene I. Shakhnovich[†], and Leonid A. Mirny[‡*]

[†]Department of Chemistry and Chemical Biology
Harvard University
12 Oxford Street
Cambridge, MA 02138

[‡]Harvard-MIT Division of Health Sciences and Technology
Massachusetts Institute of Technology
77 Massachusetts Avenue, 16-343
Cambridge, MA 02139

[*]corresponding author
email: leonid@mit.edu
Tel: (617) 452-4862
Fax: (617) 253-2514

## ABSTRACT

**Summary:** The Conservatism of Conservatism (CoC) database presents statistically analyzed information about the conservation of residue positions in folds across protein families.

**Availability:** On the web at http://kulibin.mit.edu/coc/

**Contact:** leonid@mit.edu

**Supplementary information:** The website details the method and contains a FAQ and documentation.

## INTRODUCTION

The Conservatism of Conservatism (CoC) database presents the conservation of residue positions in folds across protein families. Residues with high CoC are universally conserved in every family of homologous proteins that acquire a particular fold. Such residues can be different in non-homologous proteins (analogues) that exhibit the same fold. We calculate and present the statistical significance of such conservation and outline residues that are more conserved than expected given the residue's solvent accessibility. Such high CoC residues have been shown to be crucial for the kinetics and/or thermodynamics of protein folding, are involved in folding nucleation, and are often identified in positions of functional importance such as "super-sites"(1-3).

The database contains 3081 proteins, which cover all known protein structures in the *Protein data bank(4)* (PDB), through representation by all members of the *Families of structurally similar proteins(5)* (FSSP). Convenient access is provided by a search function that accepts queries in the form of a PDB ID, Swiss-Prot name, or FASTA amino acid sequence. A search produces a list of exact and close matches, which link to each protein's information page. Results present residues of high CoC, their sequence conservation, and corresponding p-values. A user can obtain this information in text format, view graph of Z-score and p-value, and render a PDB interactive image that highlights residues of user specified CoC and p-value cutoffs. A color-coded multiple alignment with marked CoC positions is also available to aid in visualizing the results. A compressed archive of all data files may be downloaded for further analysis.

The CoC database may be used to identify amino acids that are important for protein function and folding; and presents data sufficient for performing a stand-alone study, or suggesting key functionally and/or structurally residues to be studied via experiment. Additionally, CoC complements the analysis of experiment and simulation, and encourages understanding of protein structure in an evolutionary context.

## THE SERVER

**Queries and results:** From the search page, queries are made by PDB ID, Swiss-Prot name, or FASTA amino acid sequence. A compressed directory containing the data for all proteins may also be downloaded. Results are listed as "exact matches," which include all chains corresponding to a given PDB that match a FSSP file exactly, and also "related structures" which include PDBs identified through structural alignment. Clicking on any PDB ID in the search results will take a user to that protein's results page. The results page consists of three sections described below.

The top section gives the protein name, including domain and chain, along with a Raster3D(7) image of the protein structure. The number of sequences and structures used in the calculation is listed along with links to the tabulated data and a structural alignment of multiple representative sequences. The table of raw data lists the PDB file residue number, residue type, solvent accessibility, sequence entropy (*S(l)*), Z-score, and p-value for both six and twenty amino acid types. The multiple alignments link leads to a separate page where sequences of representative proteins from each analogous family are structurally aligned with the query sequence. FASTA query sequences are also aligned by BLAST and displayed with the structural alignments. Residues are colored by residue type (hydrophobic, polar, *etc.*) The background of individual positions is shaded by the sequence entropy obtained in the individual families. Such representation makes the concept of CoC clear: positions that are conserved in each family (appear as dark vertical stripes in the alignment) correspond to CoC residues, which are marked on the top of the alignment. Each sequence is labeled with its corresponding PDB ID, which links to that protein's main page. This offers an alternative presentation of the data and an intuitive way to picture both the conservation of a position across homologous structures and the identity of conserved positions.

The middle section of the results page contains tables displaying the number of residues identified at various CoC/p-value cutoff pairs for the data calculated using both six and twenty amino acid types. When a cutoff pair is selected, the user is brought to a separate page with an interactive cartoon image of the protein, created using jmol(8) with selected CoC residues in spacefill representation. This page also links to the structural alignment display, where a star indicates selected CoC residues. The third section of the results page presents plots of CoC and corresponding p-value. Gaps in the plots correspond to significant gaps (> 50% unaligned) in the structural alignments. Because sufficient statistics are lacking, meaningful CoC values cannot be calculated for these positions. Nevertheless, we show the conservation of these positions in the sequence alignment.

**Implication for protein function, stability, and kinetics:** There has been detailed discussion of interpreting CoC in the context of function, stability, and kinetics(1). Also, in a previous study of conservation in the protein folding nucleus(3), it was shown that residues in the nucleus are significantly more conserved than the rest of the protein. These, and other implications, have also been reviewed in the larger context of protein folding theory(2). One finds that when a position is highly conserved, but the conservation can be described by solvent accessibility, the CoC is most likely attributable to thermodynamic importance of the position. When a highly conserved position corresponds to a disulfide bond, the reason is likely thermodynamic. When most proteins of a given fold have active/binding site in the same location on the structure, such site is called a “super-site” (e.g. Rossman fold and TIM barrel). Residues of the super-site are conserved in proteins of the fold and, therefore, exhibit high CoC. When high CoC cannot be explained by any of the above rational, the cause of conservation is most likely kinetic. These high CoC residues are responsible for fast folding to the native structure and correspond to the “folding nucleus”(9). Assuming that topology determines the folding mechanism (and nucleus) of a protein, then one expects that nucleus residues to exhibit high CoC. The hypothesis of kinetic importance may also be cross-referenced with protein engineering experiment data where available.

**Concluding remarks:** The CoC database describes not only the evolutionary conservation of positions in protein folds, but serves as a source of predictions as to which residues may play an important role in protein folding, stability, and function. These data cover all known protein folds and may be used both as testable predictions for experiment or bioinformatics studies, as well as an aid in interpretation of experimental results. CoC is distinct from the simple analysis of conservation available elsewhere. It accounts for the probability that a position is conserved due to solvent accessibility and measures the conservation of a position regardless of its identity across structural homologues. We believe that CoC brings a unique and useful perspective to the analysis of evolutionary data in proteins.

## ACKNOWLEDGEMENTS

We would like to thank Grigory Kolesov for aid in development of the multiple alignment display and jmol/HTML integration, and Ivan Adzhubey for invaluable computer support. JED is supported by NSF graduate research fellowship and IAH is supported by an HHMI predoctoral fellowship. LM is an Alfred P. Sloan research fellow.

## REFERENCES

1. Mirny, L.A. and Shakhnovich, E.I. (1999) *J Mol Biol*, **291,** 177-196.
2. Mirny, L. and Shakhnovich, E. (2001) *Annu Rev Biophys Biomol Struct*, **30,** 361-396.
3. Mirny, L. and Shakhnovich, E. (2001) *J Mol Biol*, **308,** 123-129.
4. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Res*, **28,** 235-242.
5. Holm, L. and Sander, C. (1996) *Science*, **273,** 595-603.
6. Sander, C. and Schneider, R. (1991) *Proteins*, **9,** 56-68.
7. Merritt, E.A.a.B., David J. (1997) *Methods in Enzymology*, **277,** 505-524.
8. Murray-Rust, P., Rzepa, H.S., Williamson, M.J. and Willighagen, E.L. (2004) *J Chem Inf Comput Sci*, **44,** 462-469.
9. Fersht, A.R. (2000) *Proc Natl Acad Sci U S A*, **97,** 1525-1529.
10. Lopez-Hernandez, E. and Serrano, L. (1996) *Fold Des*, **1,** 43-55.