© 2001 Oxford University Press
Nucleic Acids Research, 2001, Vol. 29, No. 1 329–331

# CKAAPs DB: a conserved key amino acid positions database

**Wilfred W. Li, Boojala V. B. Reddy, Ilya N. Shindyalov and Philip E. Bourne[1,*]**

San Diego Supercomputer Center and [1]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA

Received September 12, 2000; Accepted October 17, 2000

## ABSTRACT

**The Conserved Key Amino Acid Positions DataBase (CKAAPs DB) provides access to an analysis of structurally similar proteins with dissimilar sequences where key residues within a common fold are identified. The derivation and significance of CKAAPs starting from pairwise structure alignments is described fully in Reddy *et al.* [Reddy,B.V.B., Li,W.W., Shindyalov,I.N. and Bourne,P.E. (2000) *Proteins*, in press]. The CKAAPs identified from this theoretical analysis are provided to experimentalists and theoreticians for potential use in protein engineering and modeling. It has been suggested that CKAAPs may be crucial features for protein folding, structural stability and function. Over 170 substructures, as defined by the Combinatorial Extension (CE) database, which are found in approximately 3000 representative polypeptide chains have been analyzed and are available in the CKAAPs DB. CKAAPs DB also provides CKAAPs of the representative set of proteins derived from the CE and FSSP databases. Thus the database contains over 5000 representative polypeptide chains, covering all known structures in the PDB. A web interface to a relational database permits fast retrieval of structure-sequence alignments, CKAAPs and associated statistics. Users may query by PDB ID, protein name, function and Enzyme Classification number. Users may also submit protein alignments of their own to obtain CKAAPs. An interface to display CKAAPs on each structure from a web browser is also being implemented. CKAAPs DB is maintained by the San Diego Supercomputer Center and accessible at the URL http://ckaaps.sdsc.edu.**

## BACKGROUND

Among the proteins whose 3-D structures are known, there exist many whose sequence similarities extend beyond the twilight zone into the midnight zone (1), yet their structural similarity is well established (2). While there are several hypotheses regarding the evolutionary relationship between protein sequences and structures, there has yet to be a consensus (3–5). Conserved Key Amino Acid Position (CKAAP) analysis (6) is one recent attempt to better understand the relationship between protein sequence and structure evolution; examples of other related work can be found in Mirny and Shakhnovich (7) and Hamill *et al.* (8).

The CKAAP algorithm is described fully elsewhere (6). In summary, the CKAAP algorithm analyzes the sequence relationship among representative substructures derived from the Combinatorial Extension (CE) structural alignment database (9). A substructure is a super secondary structure forming a domain or part thereof, as classified, for example, in the CATH database (10). Each substructure is represented by a reference structure formed by a polypeptide of at least 60 amino acids (master sequence). Unrelated polypeptides (subsequences with <25% identity) adopting a similar substructure are aligned to the master sequence based on the structural alignment. The sequence space is expanded by obtaining the homologs of each subsequence from SWALL (SWISS-PROT, TrEMBL, TrEMBL New) using FASTA3 (11). Using a weighted scoring scheme, the CKAAP algorithm attempts to provide an unbiased examination of the conservation of amino acid positions based on amino acid identities and property groups (6,12).

The importance of the CKAAPs identified for a number of common folds such as the immunoglobulin fold (IgFF) is well supported by existing experimental or computational literature (6,7,13). CKAAPs are found not only within the expected hydrophobic core of proteins, but also in loops and turns. CKAAPs identify the majority of the nucleation/stabilization centers predicted by other methods (14,15). In cases where systematic mutation studies are available (16,17), CKAAPs are substantiated and may provide guidance for future studies. CKAAP analysis is extensible beyond substructures to full-length polypeptide chain alignment if greater than five chains can be structurally aligned.

CKAAP analysis is based on structural alignment, the quality of the alignment is critical yet it is well known that different methods of structural alignment do not present a unique solution (18). For example, FSSP and CE offer both unique and overlapping structural alignments (18,19). Hence, CKAAPs have been calculated for both methods of structural alignment.

There is no unique structure alignment and hence no rigorous statistical treatment in defining CKAAPs. Nevertheless, the significance of CKAAPs is linked to the number of structures

*To whom correspondence should be addressed at: San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA. Tel: +1 858 534 8301; Fax: +1 858 822 0873; Email: bourne@sdsc.edu

that can be aligned, hence, CKAAP analysis is timely on the eve of structural genomics, in which new structures and folds are targeted and solved on an industrial scale (20–22). CKAAPs DB will be updated as new structure data are made available and hence provides a means to easily retrieve and review an updated list of those amino acid positions believed to be most crucial in a biological and structural role.

## DATABASE

CKAAPs DB uses the Oracle 8i object relational database (http://ckaaps.sdsc.edu). It takes advantage of Oracle WebDB/ Portal features to provide a query interface. Our approach in CKAAPs DB is to recalculate CKAAPs as we improve our own alignment methodology (23) and to incorporate the alignment methodology of others, especially FSSP (24).

At the time of writing the database comprises the features described below.

### Content

The database contains over 170 substructures determined by CE as recurring subdomains in PDB representative chains. Each substructure is similar to 5–100 or more other substructures with <25% sequence identity. The criteria for inclusion are Z score >3.7 and r.m.s.d. <4 Å (6,9).

It has over 5000 polypeptide chains, which are structural representatives determined by CE and/or FSSP. Each representative is similar to 5–100 or more other representatives with <25% sequence identity. The criteria for inclusion for CE are the same as above. The criteria for inclusion for FSSP are based solely on a Z score > 2.0 (25).

### Display features

The number of CKAAPs is by default set to be 20% of the representative sequence length. The combined score from amino acid identity and property group conservation is used to rank the amino acids within this cutoff. The rank 'a' is the highest scoring position, 'b–z' ranks lower in that order; 'A–Z' is used for ranks lower than 26, and so on. A lookup table of a rank and its respective character designation is provided.

A confidence level is calculated based on the random withdrawal of 20% of the represented sequences. Currently 500 iterations are performed and the CKAAPs that are present 100% of the time are given a confidence level of 9, with a range down to 0 (present <20% of the time). Such iterations are useful because of the arbitrary cutoff of 20%. During the iterative process, the rankings may exhibit variations due to the removal of a particular set of sequences. Therefore, a cumulative rank score is calculated such that a rank of 'a' is representative of the whole iterative process, and not just a single run. For example, position 45 may rank 'a' in one run, and get a score of 19 (an arbitrary scale relative to the total number of CKAAPs, 19, in this example, for the particular reference structure); yet only a 'c' in the second run, and a score of 17. Position 45 would get the 'a' rank only if its overall score is the highest over the total number of runs.

A profile or log odds matrix is also provided with each structure–sequence alignment. The log odds matrix provides a complete picture of the potential residues that may occur at each position. This allows the user to combine the information from the property grouping to determine which group of amino acids is most likely to occur at each position.

A rendering of the CKAAPs using MolScript (26) is currently being implemented. It will provide a spatial context for assessment of CKAAPs.

### Query and report features

Queries may be made based on the following options: representative PDB ID from CE or FSSP, PDB ID, protein names, protein function, protein classification, enzyme classification number. The user may specify the source of structural alignments as one or a combination of CE substructures, CE representatives or FSSP representatives.

Dynamically generated reports, which highlight the CKAAPs according to their confidence level, are available. This ensures that the most updated information is accessible as soon as it is in the database. Reports may contain annotations and links to the PDB query browser (27), which is a portal to other comprehensive information.

### Maintenance

Updates are performed bi-monthly using the most recent non-redundant SWALL database.

Currently database searches are performed using FASTA3 parallel-programmed for high performance computing (11).

The update not only presents the latest information, it is also a self-consistency check for the CKAAPs as different releases are available for comparison.

In the future support for BLAST searches (28) and hence sequence similarity by expectation values is to be implemented. In addition, we plan to incorporate superfamily classifications according to SCOP (2).

## ACKNOWLEDGEMENTS

This work was supported by the National Biomedical Computation Resource (NIH P41 RR 08605-06), National Science Foundation grant DBI 9808706 and Microstructure Image-Based Collaboratory grant (NCRR 5 P41 RR04050-10S1).

## REFERENCES

1. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
2. Lo Conte,L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
3. Wood,T.C. and Pearson,W.R. (1999) Evolution of protein sequences and structures. *J. Mol. Biol.*, **291**, 977–995.
4. Lattman,E.E. and Rose,G.D. (1993) Protein folding–what's the question? *Proc. Natl Acad. Sci. USA*, **90**, 439–441.
5. Russell,R.B. and Barton,G.J. (1994) Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.*, **244**, 332–350.
6. Reddy,B.V.B., Li,W.W., Shindyalov,I.N. and Bourne,P.E. (2000) Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins. *Proteins*, in press.
7. Mirny,L.A. and Shakhnovich,E.I. (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.*, **291**, 177–196.
8. Hamill,S.J., Steward,A. and Clarke,J. (2000) The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.*, **297**, 165–178.

9. Shindyalov,I.N. and Bourne,P.E. (2000) An alternative view of protein fold space. *Proteins*, **38**, 247–260.
10. Orengo,C.A., Pearl,F.M., Bray,J.E., Todd,A.E., Martin,A.C., Lo Conte,L. and Thornton,J.M. (1999) The CATH database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 223–227.
11. Pearson,W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.*, **24**, 307–331.
12. Taylor,W.R. (1986) The classification of amino acid conservation. *J. Theor. Biol.*, **119**, 205–218.
13. Clarke,J., Cota,E., Fowler,S.B. and Hamill,S.J. (1999) Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway. *Structure Fold. Des.*, **7**, 1145–1153.
14. Demirel,M.C., Atilgan,A.R., Jernigan,R.L., Erman,B. and Bahar,I. (1998) Identification of kinetically hot residues in proteins. *Protein Sci.*, **7**, 2522–2532.
15. Michnick,S.W. and Shakhnovich,E. (1998) A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold. Des.*, **3**, 239–251.
16. Brown,B.M. and Sauer,R.T. (1999) Tolerance of Arc repressor to multiple-alanine substitutions. *Proc. Natl Acad. Sci. USA*, **96**, 1983–1988.
17. Dalal,S., Balasubramanian,S. and Regan,L. (1997) Transmuting alpha helices and beta sheets. *Fold. Des.*, **2**, R71–R79.
18. Godzik,A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, **5**, 1325–1338.
19. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
20. Brenner,S.E. and Levitt,M. (2000) Expectations from structural genomics. *Protein Sci.*, **9**, 197–200.
21. Skolnick,J., Fetrow,J.S. and Kolinski,A. (2000) Structural genomics and its importance for gene function analysis. *Nature Biotechnol.*, **18**, 283–287.
22. Shapiro,L. and Harris,T. (2000) Finding function through structural genomics. *Curr. Opin. Biotechnol.*, **11**, 31–35.
23. Guda,C., Scheeff,E., Bourne,P. and Shindyalov,I. (2001) A new algorithm for alignment of multiple protein structures using Monte Carlo optimization. *Pacific Symp. Biocomp.*, in press.
24. Holm,L. and Sander,C. (1996) The FSSP database: fold classification based on structure–structure alignment of proteins. *Nucleic Acids Res.*, **24**, 206–209.
25. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
26. Esnouf,R.M. (1999) Further additions to MolScript version 1.4, including reading and contouring of electron-density maps. *Acta. Crystallogr. D, Biol. Crystallogr.*, **55**, 938–940.
27. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 214–218.
28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.