SECOST: sequence–conformation–structure database for amino acid residues in proteins

Oleg Shats¹, Iosif I. Vaisman², Aelita Shats¹ and Simon Sherman¹

¹Epley Institute for Research in Cancer and Allied Diseases, University of Nebraska Medical Center, 986805 Nebraska Medical Center, Omaha, NE 68198-6805 and ²University of North Carolina at Chapel Hill, Chapel Hill, NC 2759-7360, USA

Received on December 18, 1998; revised on March 4, 1999; accepted on March 16, 1999

Abstract
Summary: The sequence–conformation–structure database for amino acid residues contains information on 114,828 individual residues derived from the spatial structures of 473 high-quality non-homologous proteins. The information in the database is obtained using a variety of different methods and can be used in various protein modeling applications.

Availability: The database is accessible at http://bims.unmc.edu/secost.htm

Contact: ssherm@unmc.edu

Supplementary information: A list of the selected proteins is presented in Table 1 (http://bims.unmc.edu/ECCC5/cdb_t1.htm).

Knowledge of intrinsic conformational propensities of amino acid residues within polypeptide chains is necessary to improve the accuracy and precision of protein structure determination from X-ray or NMR data, to solve problems of protein folding and protein engineering, to predict protein and peptide structures from incomplete experimental data, to analyze the structure–function relationship of polypeptide molecules, and for many other applications. To help solve these problems, we developed a sequence–conformation–structure database (SECOST) for amino acid residues from 473 high-quality non-homologous proteins.

To create SECOST, protein structures with <20% sequence identity, with a resolution of 2 Å and better, and R-factors better than 0.225 were selected from the Brookhaven PDB [the list of proteins was derived from the all-against-all BLAST sequence comparison (Dunbrack, 1998) and is presented in supplementary information]. The conformational characteristics of amino acid residues of these proteins were determined and entered into the database. The records for each residue in the database contain the following information: four-letter code of the corresponding protein in Brookhaven PDB; one-letter chain identifier (if the protein has more than one chain); three-letter code of the amino acid residue and its sequential number in the polypeptide chain; one-letter code for the following residue; values of the dihedral angles; an area of the residue accessible by water molecules (fractional area); energy by residue (i.e. energy of interaction of the atoms of the residue with the atoms of other residues in the protein); number of other residues found in the 8 Å sphere around the residue; residue participation in a secondary structure; name of the conformational cluster (to which conformational state the residue was assigned); its prior and posterior probabilities; and number of appearances of the residue in five different types of Delaunay simplices.

The values of the dihedral angles, energy of interaction of the residue’s atoms with the atoms from the other residues, and the number of residues within the 8 Å sphere around the given residue were calculated by the PROTABLE module from the SYBYL 6.4 software package (TRIPOS Associates Inc., St Louis, MO). The conformations of the residues were assigned to the maximum probability conformational cluster using the ConCoder program (Sherman et al., 1998). The fractional area of each residue was calculated using the Lee and Richards (1971) approach. Involvement of the residues in the formation of elements of secondary structure was determined by the Kabash and Sander (1983) procedure. Residues unassigned by this procedure to any secondary structure element, but exposed to the solvent and assigned to the ‘P’ conformation cluster (Sherman et al., 1998), were considered as polyproline II elements. Proteins in the database were tessellated using the Delaunay algorithm (Singh et al., 1996). A single point (Cα) representation per residue was used for the tessellation. For each residue, the number of Delaunay simplices with this residue as a vertex were calculated. The Delaunay simplices were classified according to the sequential proximity of the participating residues (Singh et al., 1996), and the number of simplices in each of five classes was entered into the database for each residue. Currently, the tessellation data are presented for a subset of the database including only proteins that do not contain modified amino acid residues, gaps in sequence and other irregularities.

The developed interface for SECOST allows the users to request specific information from the database using the following parameters: Clean-up, Resolution, R-factor, Fractional area, Configuration, Pattern for amino acid sequence, Pattern for secondary structure and Pattern for conformational sequence. The meaning of each of these parameters is described below.
After receiving a query, the program extracts all the records from a subset of proteins determined with the specified resolution and crystallographic R-factors. These records are processed to select those that match (for the current residue) both the given interval for fractional area and the type of peptide group configuration. Moreover, these records must match simultaneously three patterns (sequential, secondary structure and conformational) involving the current and its four N- and four C-terminal residues.

**Clean-up.** In some rare cases, the information on particular amino acid residues can be incomplete or unrealistic. To exclude such ‘artifact’ data, the clean-up parameter can be used.

**Resolution.** SECOST contains information on amino acid residues in proteins determined by X-ray methods with a resolution of 2 Å and better. In the queries, users can limit the search of the information for proteins within smaller intervals (e.g. within the interval from 1.2 to 1.8 Å).

**R-factor.** Analogous to the resolution, the information in the database corresponds to proteins determined with R-factors better than 0.225. In queries, one can specify smaller intervals for R-factors.

**Fractional area.** In most of the cases, the calculated fractional areas are to be within the interval from 0.0 to 1.0. However, on N- and C-terminal parts of the polypeptide chains, the calculated fractional areas can be bigger than 1.0. Therefore, by default, there is no upper limit for fractional area. In queries, users can select amino acid residues having fractional areas within user-specified intervals. For instance, the interval from 0.0 to 0.3 (for buried residues) can be specified.

**Configuration.** One can select a particular (cis or trans) configuration of the peptide group. The default configuration is ‘any’ (i.e. without discriminating cis- or trans-configuration).

**Pattern for amino acid sequence.** The interface has a special tool allowing users to select residues within subsequences containing up to nine residues: one ‘current’, four ‘preceding’ and four ‘following’ residues. The default parameter for type of amino acid residue(s) is ‘any’. Users can specify one or several types of residue, given by three-letter code. It is also possible to choose a group of residues with common physicochemical features (such as polar, non-polar, charged, aromatic and small).

**Pattern for secondary structure.** Similarly, a requested pattern for secondary structure can be formed by the following letters: ‘H’ or ‘h’ (α-helix), ‘B’ or ‘b’ (isolated β strand), ‘E’ or ‘e’ (hydrogen-bonded β strand), ‘G’ or ‘g’ (3_10 helix), ‘I’ or ‘i’ (π helix), ‘T’ or ‘t’ (hydrogen-bonded β turn) and ‘P’ or ‘p’ (polyproline II elements). The symbol ‘-’ can be used to select cases without secondary structure.

**Pattern for conformational sequence.** The database can also be searched by a conformational pattern containing sequence names of conformational clusters for ‘current’, four ‘preceding’ and four ‘following’ residues. Conformational clusters for amino acid residues are coded by B (β-strand-like conformations), P (polyproline II-like conformations), R (right-hand α-helix-like conformations), L (left-hand α-helix-like conformations) and G (all the conformations not included in previous ones). The symbol ‘@’ is used to code both the conformational states with the cis-configuration of the peptide group and the conformations of the residues with incomplete information on φ, ψ or ω angles.

**Fields for the output file.** Users can specify particular fields to be presented in the output file. For instance, one can request ID of the PDB file, resolution and R-factor from which the records were taken to be present in the output file. Users can request the following information about the current residue: one-letter chain identifier (if there is more than one chain); three-letter code of the residue and its sequential number in the polypeptide chain; one-letter code for the following residue; values of the dihedral angles; fractional area; energy by residue; number of other residues in the 8 Å sphere; information about residue participation in a secondary structure; name of the conformational cluster; and its prior and posterior probabilities. Finally, users can request information on the number of times the residue appears in five different types of Delaunay simplices.

**Acknowledgements**

We acknowledge the use of the Molecular Modeling Core Facility supported by Cancer Center Support grant P30 CA36727 to the UNMC Epplsey Cancer Center. I.I.V. was supported, in part, by a Faculty Development Award in Bioinformatics from PhRMA Foundation.

**References**


