

## NAME

**esseb** - Constrained Enumeration of Protein SSE Conformations

## VERSION

Version 0.1

Molecule Library Version 0.1

## SYNOPSIS

**esseb** input\_pdb [-a max\_error] [-b empty\_count empty\_length] [-c] [-d database\_file] [-e error\_adjust] [-f cross\_over closest\_approach axis\_ends] [-g] [-h] [-i drs constraints range] [-j rms\_lower] [-k] [-l log\_file] [-m] [-n notice\_level] [-o err\_scale] [-p tau\_max dist\_min dist\_max dist\_adj\_max] [-q min\_closest max\_closest min\_end max\_end] [-r rms\_upper] [-s] [-t d\_num omg\_num theta\_num psi\_num phi\_num] [-u] [-v] [-x] [-y axis\_num dump\_type] [-z]

## DESCRIPTION

ESSEB performs constrained enumeration of secondary structure element conformations representing protein structures. A protein is considered as a set of Secondary Structure Elements (*alpha helices, beta-strands, etc.*). Each SSE is represented by an axis whose position is described by a quaternion and a center of mass. The conformational space of the set of SSEs is divided up into *cells*. Each cell represents a *range* of theta, phi, and psi angular orientations and a volume representing the center of mass *range*. The program generates an enumeration of the possible SSE orientations that satisfy *constraints* including tau angle limitations, minimum and maximum distance of closest approach between SSE axes, distance between SSE axis endpoints, axis crossover, minimum and maximum axis center distances, and pairwise atomic distance restraints.

The program is currently limited to integral membrane proteins such that the enumeration occurs in a plane where each SSE axis center must lie. The angle of the SSE axis with respect to the plane is restricted, and each SSE axis is flipped with respect to the previous one so that the sequence number at endpoints is consistent with respect to the plane. See the **USAGE** section for a more detailed description.

## PARAMETERS

**-a max\_error**

Set the maximum internal error for acceptance of a conformation during local search. *If max\_error* is set to zero, NO local search is performed. The maximum error is set *per* SSE and therefore the maximum error in a distance restraint can be twice this value.

**-b empty\_count empty\_length**

If the input file is set to *empty*, *empty\_count* SSEs axes of length *empty\_length* angstroms are used for the enumeration. The default values are *empty\_count=5* and *empty\_length=30*. This parameter has no effect if an *input\_pdb* is used for enumeration.

**-c** Perform NO database output. This option is used only to get a count of the conformations found to satisfy the specified restraints. See also **-u**

**-d database\_file**

Specify the name of the ESSEB database file. The default is "**esseb.db**"

**-e** *error\_adjust*

Adjust the experimental error for every distance restraint. The minimum distance is decreased by *error\_adjust* and the maximum distance is increased by the same amount. Signed values are allowed for *error\_adjust*.

**-f** *cross\_over closest\_approach axis\_ends*

Specify enumeration *constraints* to be used. If *cross\_over* is set to zero, the axis crossover constraint is ignored. If *closest\_approach* is set to zero, the distance of closest approach between two SSE axes is not constrained. If *axis\_ends* is set to zero, the distance between axis end points is not constrained. The default is **-f 1 1 1** where all *constraints* are enforced. Use **-f 0 0 0** to ignore the above restraints.

**-g** Do not allow internal error tolerance for conformation acceptance. By default, a conformation is accepted once the maximum internal error is below the specified threshold and distance restraints are satisfied with a tolerance for the internal error. If this option is set, local search is performed until maximum internal error is below the threshold, however, a conformation is only accepted at this point if the distance restraints are satisfied with no internal error tolerance.

**-h** Print out the man page for help

**-i** *drs constraints range*

Determine how internal error is handled.

*drs*

When nonzero, internal error is added to the distance restraints. If zero, Local search and adaptive knockout are not performed. Conformations are accepted only if they satisfy the error present in the distance restraints.

*constraints*

When nonzero, *constraints* for closest approach, axis crossover, and axis end points are not violated if the internal error allows for some conformation within the cell to pass

*range*

When nonzero, the minimum, maximum, and maximum adjacent axis center *ranges* with respect to a previous SSE are added to the axis center spacing of the previous SSE to increase the *range*.

Use **-i 0 0 0** to ignore internal error.

**-j** *rms\_lower*

Set the division count for each SSE such that the enumerated protein conformations will have a RMS deviations no lower than *rms\_lower* from each other. The division counts for all SSEs are set to 1 and iteratively increased until the RMS cannot be decreased. Set **-n 30** to see the resulting division counts for each SSE. *Note: If 'blank' SSEs are used, the RMSD\_HEL is used instead of the all atom RMS.* See also **-r**.

**-k** Do not use adaptive knockout.

**-l** *log\_file*

Redirect output to a log file. Errors and warnings are also displayed on the screen. Use **-n 30** for full output.

**-m** Do not merge continuous SSEs into a single axis. By default an SSE is merged with a second SSE if the terminal residue of one is the starting residue of another, or if the terminal residue of one comes just before the starting residue of another in terms of residue sequence numbers. *Note: if this option is set, the ending residue for one SSE may NOT be the starting residue for another.*

**-n** *notice\_level*

Set the degree of program output. Use:

- n 0** No output
- n 10** Normal program output
- n 20** Parameters useful for reproducing the results
- n 30** All output

**-o** *err\_scale*

Scale the internal error by *err\_scale*. This can be used to increase or decrease the calculated internal error.

**-p** *tau\_max dist\_min dist\_max dist\_adj\_max*

Set the enumeration *constraints* for maximum tau and axis center distances.

*tau\_max* is the maximum angle of SSE axis with the Z-axis (which is normal to the bilayer plane). The default is 40 degrees. *The angle should be specified in degrees*

*dist\_min* is the minimum distance between two SSE axis centers. The default is 6 angstroms.

*dist\_max* is the maximum distance between two SSE axis centers. The default is 40 angstroms.

*dist\_adj\_max* Maximum distance between two SSE axis centers for SSEs that are adjacent with respect to their occurrence within the primary sequence. Default is 13.4 Angstroms

**-q** *min\_closest max\_closest min\_end max\_end*

Set *constraints* for SSE axis closest approach and end-point distances.

*min\_closest* is the minimum distance of closest approach between two axes (default 6)

*max\_closest* is the maximum distance of closest approach between two SSEs adjacent in terms of location in primary sequence (default 13.4)

*min\_end* is the minimum distance between SSE axis end points (default 7)

*max\_end* is the maximum distance between adjacent SSEs on each side of the bilayer plane (default 22).

**-r** *rms\_upper*

Set the division count for each SSE such that the upper bound for the all atom RMS is not greater than *rms\_upper*. The division counts for all SSEs are set to 1 and the count corresponding to the highest internal error is incremented iteratively until the RMS is below *rms\_upper*. Set **-n** 30 to see the resulting division counts for each SSE. *Note: If 'blank' SSEs are used, the RMSD\_HEL is used for the upper bound instead of the all atom RMS.* See also **-j**.

**-s** Do not sort the enumeration order by the number of distance restraints. This option greatly reduces the internal error caused by discrete center of mass placement at a constant division count because each SSE is placed relative to an adjacent SSE and therefore has tighter constraints on the center of mass distances. If the number of distance restraints for each SSE is relatively invariant, or if there are a large number of distance restraints for each SSE, then *this option should be set*. The first SSE is the middle SSE with the most distance restraints. Each next SSE is the SSE with the most distance restraints that is adjacent to a previous SSE.

**-t** *d\_num omg\_num theta\_num psi\_num phi\_num*

Set the division counts used to divide the conformational space into cells for all SSE axes.

*d\_num* is the number of distances from a previous helix. This parameter applies to the cell division between the minimum and maximum axis center distance between two SSEs. Because this *range* is different between adjacent SSEs in integral membrane proteins, the internal error created by the cell division at the maximum values is calculated and the division at other *ranges* is calculated to maintain a constant internal error.

*omg\_num* is the number of center of mass enumerations at one distance. Similar to *d\_num*, this division number occurs at the maximum enumerated distance from a previous SSE. The division number at other distances is to maintain error. The *range* of axis centers covered by one cell is determined by *d\_num* and *omg\_num*. If both values are set to one. Each cell covers the entire *range* of possible axis centers.

*theta\_num* describes the number of theta angles.

*psi\_num* describes the number of psi angles. Similar to *d\_num* and *omg\_num*, this division occurs at the theta closest to  $\pi/2$ , and all other divisions are calculated to maintain error.

*phi\_num* describes the number of axis spins.

**-u** Do not perform the enumeration, but instead report the upper bound for the number of possible enumerated protein conformations. This upper bound considers only the tau angle restraint and the axis center restraints. For upper bounds including other restraints, the global enumeration should be run with *empty* axes or using an input file with no distance restraints.

**-v** Output to the database using vector format. See **essebdb(1)** for a description of the format. *Protein structures cannot be mapped onto conformations specified in this format.*

**-x** **DEBUGGING:** Output every accepted conformation during local search. Each conformation is output as a database record which can later be converted to PDB formatted files.

**-y** *axis\_num dump\_type*

**DEBUGGING:** Writes out the cells from a local enumeration for the SSE axis *axis\_num* in PDB format. *axis\_num* refers to the zero-indexed **sorted** reference to an axis (See **-s**). If *dump\_type* is

set to *single* the enumeration is output as one file, if it is set to *multiple*, one file is output for each cell. The local enumeration can be visualized in PyMol with the script *show\_local\_enum*. Global enumeration is not performed with this option.

**-z**     **DEBUGGING:** Apply a random axis spin to each SSE axis before enumeration.

## USAGE

ESSEB performs constrained enumeration of protein conformations in terms of SSE axis orientations. The input for the program is a PDB file containing the atomic structure of each SSE and a list of distance restraints. SSEs within the PDB file are identified by HELIX records and distance restraints are specified with DSTRST records (see *essebdrs(1)* for a description of the format and information on managing distance restraints within PDB files). By default, SSEs continuous in terms of primary sequence are merged (see *essebdrs(1)* for details). This can be prevented using **-m**. Each SSE axis is calculated as the eigen vector corresponding to the minimum eigen value for the inertia matrix calculated for alpha-carbons in the SSE along with any atoms containing distance restraints. If a N-terminal nitrogen and a C-terminal carbon are present in the SSE, the projection of these atoms onto the SSE axis vector determines the endpoints. Otherwise, the first and last atoms in the SSE are used. For theoretical calculations, *empty* SSEs can be loaded (see **-b**).

The enumeration is performed by orienting an SSE, checking restraints, and then enumerating the next SSE based on the location of a previously enumerated SSE. There are two options for the order of SSE enumeration (see **-s**). The conformations considered for each SSE during enumeration are governed by the division count which can be specified using **-t**. Alternatively, an upper bound to the all-atom RMSD between a real conformation and an enumerated conformation can be specified to calculate the division counts (**-r**). If empty SSEs are used, this value reflects an upper bound to the RMSD\_HEL instead of the all-atom RMS. If desired, it can instead be specified that the enumerated protein conformations be at least a specified value apart in terms of all atom RMSD or RMSD\_HEL (**-j**). The *constraints* used for enumeration include axis *constraints* and distance restraints. The axis *constraints* include minimum and maximum axis center distances (**-p**), maximum tau angle (**-p**), distance of closest approach extremes (**-q**), extremes for distances between axis endpoints (**-q**), and an axis loop crossover constraint. The axis constraint checks can be turned off using **-f**. The distance restraints are read from the input file and can be adjusted using **-e**.

An upper bound for the number of possible enumerated conformations (neglecting distance of closest approach, end point, and crossover restraints) can be obtained using **-u**.

An internal error is associated with each SSE. By default a conformation is accepted if it satisfies the *constraints* with an allowance for the internal error. The internal error can be scaled linearly using **-o**. If the **-i** option is set, internal error is ignored for the specified restraints. Local search is performed for each SSE conformation until the maximum internal error is below a certain threshold specified by **-a**. Using **-a 0** turns off local search. Adaptive knockout describes a process by which local search is performed between SSE pairs before a full local search is performed. This can increase efficiency by throwing out conformations that do not meet the specified restraints with fewer evaluations. Adaptive knockout is turned off using **-k**. At the end of local search, a conformation is accepted if it satisfies the distance restraints with allowance for the specified maximum internal error. If instead, at the end of local search, a conformation should only be accepted if it satisfies the *constraints* with no tolerance for internal error, the **-g** option should be set.

During enumeration, the program displays a progress meter to standard out. The progress meter shows the percentage of the upper bound that has been evaluated, however, has NO meaning in terms of total runtime. Detailed output consisting of data for the individual SSEs and upper bounds can be generated using **-n 30**. To record this information to a log file, use **-l**.

Each acceptable conformation is written as a database record to a file specified by **-d**. See **-v** and *essebdb(1)* for details on database format. *essebdb* is used to convert database records into PDB files

following an enumeration. If the **-c** option is set, only a count of accepted conformations is generated and no database output is performed.

## EXAMPLES

**esseb** empty **-b** 5 30 **-r** 6 **-i** 0 0 0 **-c** **-n** 30

Ignore internal error. Get a count of the number of conformations that satisfy axis *constraints* for five 30 angstrom SSE axes with an RMSD\_HEL upperbound of 6 angstroms

**esseb** in.pdb **-r** 3 **-i** 0 0 0 **-n** 30

Enumerate all conformations satisfying *constraints* for the SSEs and distance restraints contained within *in.pdb*

**esseb** in.pdb **-a** 10 **-r** 25

Enumerate with an initial all atom RMS of 25, but perform local search and adaptive knockout until the maximum internal error for a distance restraint is 10 angstroms

## SEE ALSO

**essebdrs**(1) and **essebdb**(1)

## AUTHORS

ESSEB was developed by W. Michael Brown and Jean-Loup Faulon