User Manual for ereg

Michael J. Dudek
Hexadecapole

June 2024

1 INTRODUCTION

ereg is a software package for modeling of proteins and nucleic acids. The energy functions, which include a careful treatment of electrostatic energy, provide an alternative to standard models.

The initial functionality of the package was molecular mechanics-based prediction of protein structure. Over a period of several years, the energy functions and the algorithms for search through the space of conformations evolved under the selective pressure of success in this initial application. To support the core molecular mechanics machinery, functionalities were added to build homology models, to detect defects in models, to predict folds (ab initio) for single-domain proteins, and to predict docked configurations for pairs of structures. Sequence design functionality was obtained from the initial structure prediction functionality by adding reference states (for example the unfolded state for design of thermodynamic stability, or the unbound conformation for design of binding affinity) together with search through residue sequence space. The same methods added to enable sequence design also enable prediction of $\mathrm{pK_a}$ for ionizable residues. By introducing a generalization of the concept of residue, much of the protein functionality was extended to oligonucleotides. Some of the more common nucleotide chemical modifications are included in the residue data set.

The user interface is designed with a goal of creating an easily-usable tool set, providing functionality in factored units of utility that should be combinable to accomplish a range of modeling studies.

2 COMPUTATIONAL REQUIREMENTS

The program, which consists of roughly 124,000 lines of C++ code, was developed for a macOS (or linux) workstation with a requirement of 8, and preferably more, Gigabytes of memory. The source code is compiled using "gcc". The calculations are computationally intensive.

3 INSTALLATION

1) From the company website, download the current version of the program: "ereg_jun2024.tar.gz".

2) Restore the directory structure.

```
%  gunzip ereg_jun2024.tar.gz
%  tar xvf ereg_jun2024.tar
```

The top level directory, "ereg_jun2024", can be renamed as desired.

3) Activate LAPACK options.

```
%  cd ereg_jun2024/src
```

For one source code file "med_sig.cc", better performance can be achieved by interfacing to LAPACK. The default version of this file interfaces correctly to the LAPACK available with Xcode installed on a macOS system. So for a macOS system, continue to step 4.

For a linux system, the script "zzcodebase/zzlapack_linux" replaces file "med_sig.cc" with a version that should interface correctly to LAPACK installed on a linux system.

From directory "src/zzcodebase"

```
%  zzlapack_linux
```

Alternatively, script "zzcodebase/zzlpack_noblas" uses equivalent (but slower) linear algebra routines contained within the ereg codebase.

From directory "src/zzcodebase"

```
%  zzlapack_noblas
```

4) Compile the codebase.
From directory "src"

```
%  zzmake
```

5) Compile the program commands.
From directory "src/str"

```
%  make compile
```

If no errors are encountered, the program has installed successfully.

4 DIRECTORY STRUCTURE

The top level directory contains the following 8 subdirectories.

The "src" directory tree contains the C++ source code, the object files created by compilation of the source code, and all permanent data files. Following compilation, the executable user commands are contained in subdirectory "src/str".

The "man" directory contains the user manual in PDF format.

The "fam" directory tree contains the path structure required by ereg commands for locating user-supplied and program-generated data files. A new user project is initiated by copying and renaming this directory tree structure. The subdirectories "fam/arg", "fam/car", "fam/dgn", "fam/exp", "fam/seq", "fam/stp", and "fam/tor" organize all data files associated with a single user project.

The "test" directory is a user project populated with input files for each of the example test cases included in this manual. For each ereg command, the corresponding example demonstrates the functionality, verifies proper execution, and benchmarks computational requirements.

The "TESTCASES" directory is a backup of the "test" directory following execution of all test cases. Proper execution can be verified by comparison of data files generated in directory "test" with the corresponding files in directory "TESTCASES".

The directories "up0", "up1", and "up2" contain tutorial user projects that demonstrate the use of pipelines of ereg commands to accomplish real world modeling tasks.

The "up0" tutorial demonstrates prediction of structure for 1pga, a small protein.

The "up1" tutorial demonstrates generation of a homology model for herceptin Fv domains, an antibody fragment.

The "up2" tutorial demonstrates prediction of relative binding affinities to ribonuclease H1 catalytic domain for three nucleic acid duplex compositions: dna-dna, dna-rna, and rna-rna.

5 USER INTERFACE

At the expense of a graphical interface, development resources have been focused on energy functions, on algorithms for search of conformation space, and on a handful of supporting utilities. The ereg package interfaces to programs (such as "pymol") for graphical output display through pdb-format files. The user interface consists of 12 easily learned commands typed to the macOS prompt. For computers running macOS, ereg includes a simple viewing app (written in Apple's Swift and Metal languages) for graphical display of program generated structures.

The following command line arguments, each replaced by the user with a meaningful string of characters, specify the associated objects.

FAM A project name (i.e. test, up0). The name can be any sequence of characters for which a directory tree has been created usually by copying and renaming directory "fam"

MOL A molecule or set of molecules that compose a system (i.e. 1pga, 1lni).

CNF A structure or conformation of the system.

SUB A subset of the set of rigid-geometry degrees of freedom of the system.

GRP A collection of templates used in homology model building.

6 CENTRAL COMMANDS

The functionality of the package is accessed through the following 10 commands.

6.1 greg

FUNCTIONALITY
  geometry regularization

SYNTAX

greg FAM MOL

INPUT FILES	OUTPUT FILES
/FAM/exp/MOL.pdb	/FAM/dgn/greg.MOL
	/FAM/seq/seq.MOL
	/FAM/tor/tor.MOL.00
	/FAM/car/MOL.00.pdb

SUMMARY

1) inputs a pdb-format file, "FAM/exp/MOL.pdb", usually an experimental structure

2) regularizes geometry, meaning bond lengths, bond angles, and some torsion angles are adjusted to standard values with minimal movement of atom coordinates

3) outputs, in file "FAM/seq/seq.MOL", residue sequence and disulfide crosslinks, a specification of the covalent connectivity of the mechanical system independent of conformation

4) outputs 2 alternative specifications of the geometry regularized structure, cartesian coordinates in pdb format in file "FAM/car/MOL.00.pdb", and torsion angle coordinates in file "FAM/tor/tor.MOL.00"

5) outputs, in file "FAM/dgn/greg.MOL", a diagnostic summary of the execution of the command

NOTES

The energy surface is defined for a rigid-geometry model. To access the energy surface for structure prediction, a generalized structure of a molecule or system of molecules must be moved into the subspace of structures consistent with regularized geometry.

The primary use of this command is, for a collection of templates in preparation for homology model building, to convert experimental structures into geometry regularized structures. A second, less common, use is geometry regularization of large structures, as an alternative to the "ereg" command, in preparation for application of structure prediction to localized regions.

Preparation for the command consists of entering a pdb-format file "MOL.pdb" into directory "FAM/exp", and possibly editing "MOL.pdb" to include only the chains or fragments of interest. The only diagnostic information associated with the command is the heavy atom RMSD between the final geometry regularized structure and the initial structure. The conformation name ’00’ is assigned by the program to the conformation that results from initial geometry regularization of an experimental structure.

The "greg" command does not access the energy surface and requires only a few seconds of computation time.

EXAMPLE TEST CASE

File "1lni.pdb", a crystal structure of a 96-residue ribonuclease from streptomyces aureofaciens, has been entered into directory "test/exp".

To execute the command, open a window to directory "src/str", and type the following line to the macOS (or linux) prompt.

```
%  greg test 1lni
```

The program outputs files "test/seq/seq.1lni", "test/tor/tor.1lni.00", "test/car/1lni.00.pdb", and "test/dgn/greg.1lni". As a test of proper execution, each output file can be compared to the corresponding file in the directory "TESTCASES".

The heavy atom RMSD between experimental and geometry regularized structures is .170 Å.

6.2 ereg

FUNCTIONALITY
local energy minimization

SYNTAX
ereg FAM MOL

INPUT FILES	OUTPUT FILES
/FAM/exp/MOL.pdb	/FAM/dgn/ereg.MOL
	/FAM/seq/seq.MOL
	/FAM/tor/tor.MOL.CNF
	/FAM/car/MOL.CNF.pdb

SUMMARY

1) inputs a pdb-format file, "FAM/exp/MOL.pdb", usually an experimental structure

2) regularizes geometry, meaning bond lengths, bond angles, and some torsion angles are adjusted to standard values with minimal movement of atom coordinates

3) starting from the initial geometry regularized structure, minimizes energy locally with respect to all torsion angle degrees of freedom

4) to prevent movement out of the well of the initial conformation, minimization on the energy surface is accomplished by generating a sequence of up to 9 local minimization trajectories, gradually reducing to zero the weighting coefficient for a set of harmonic distance constraints taken from the initial structure

5) outputs, in file "FAM/seq/seq.MOL", a specification of residue sequence and disulfide crosslinks

6) outputs, using CNF=00 to denote the initial geometry regularized structure, and CNF=$\{01, 02, 03, ..., 09\}$ to denote the energy minima corresponding to the endpoints of the sequence of local minimization trajectories, 2 alternative specifications of conformation, cartesian coordinates in pdb-format file "FAM/car/MOL.CNF.pdb", and torsion angle coordinates in file "FAM/tor/tor.MOL.CNF"

7) outputs, in file "FAM/dgn/ereg.MOL", a diagnostic summary of the execution of the command, including a compact analysis of the single point energy and RMSD evaluations at the end points of the sequence of local minimizations

NOTES

The "ereg" command is a better alternative to the "greg" command for bringing an experimental structure into regularized geometry in preparation for sequence design, prediction of pKa values of ionizable groups, or structure prediction of localized regions. Since this command accesses the energy surface, it is also useful for scoring structures based on evaluation of full energy.

Preparation for the command consists of entering a pdb-format file "MOL.pdb" into directory "FAM/exp", and possibly editing "MOL.pdb" to include only the chains or fragments of interest. Consistent with the "greg" command, the conformation name ’00’ is assigned by the program to the conformation that results from initial geometry regularization. If the input structure, "FAM/exp/MOL.pdb", was created by the ereg program, then geometry regularization should not alter the input structure, and the CNF=00 structure is identical to the input structure.

Table I. Components of Total Energy.

Component	Name	Functional Form
repulsion+dispersion	F_r	3-parameter buf14-7
electrostatic	F_e	multipole expansion
disulfide crosslink	F_s	harmonic
intrinsic torsional	F_t	1D, 2D, or 3D fourier series
ring closure crosslink[a]	F_b	harmonic
distance constraint	F_c	harmonic
hydration entropy	F_h	gaussian volume
dielectric medium	F_m	boundary element solution to Poisson equation
conformation entropy[b]	F_g	step function
polarization[c]	F_p	

[a] The F_b component, which contributes to the energy of the proline ring and to the energies of the ribose and deoxyribose rings, could logically be considered a part of F_t. It is currently being split out as a separate component to assist in analysis of the parameterization of nucleotide residues.

[b] The F_g component is an exploratory term used to penalize chain conformations associated with loss of entropy. It has not yet been described in publication.

[c] The F_p component is an exploratory term used to model energies associated with electron polarization. It has not yet been described in publication.

The notation used to refer to components of the total energy is defined in Table I. Energy minimization is carried out on the (Fr+Fe+Fs+Ft+Fb+Fc+Fh+Fw+Fp) energy surface, with charges on ionized groups scaled to 9/64 of full values. This approximation to the full (Fr+Fe+Fs+Ft+Fb+Fc+Fh+Fm+Fg+Fp) energy surface excludes Fg and replaces Fm with Fw, a crude hydration shell model for which calculation of analytic 1st and 2nd derivatives is more tractable. At the endpoint of local minimization, charges on ionized groups are restored to full values, and Fe is recalculated along with (Fm+Fg) as a single point evaluation. The sequence of local minimizations on the approximate energy surface (Fr+Fe+Fs+Ft+Fb+Fc+Fh+Fw+Fp), using reduced charges on ionized groups, and gradually reducing to zero the weighting coefficient for Fc, is ended if the single point evaluation of full energy (excluding Fc) increases.

The "ereg" command accesses the energy surface. Using an Apple M1 Pro processor, local energy minimization of the 56 residue protein 1pga requires computation time of about 11 minutes.

For systems in which the number of torsion angle degrees of freedom is greater than 1700, the command chooses a collection of subsets of degrees of freedom, each containing less than 1700 degrees of freedom, such that the union spans the entire molecule or system of molecules. Local energy minimization is accomplished by cycling through this collection, minimizing locally with respect to each subset before moving to the next. For energy minimizations with respect to subsets of the full set of rigid geometry degrees of freedom, the function minimized is a partial sum over the full set of terms that contribute to the total energy of the system. Those terms that depend only on degrees of freedom outside of the subset are not included. Because partial energies are of limited utility for predicting the relative stability of 2 conformations, the single point evaluation of (Fe+Fm+Fg) at the endpoint of local minimization is not included. For accomplishing geometry regularization of very large structures, the "greg" command can be used as a fast alternative to the "ereg" command.

EXAMPLE TEST CASE

File "1pga.pdb", a crystal structure of a small 56 residue protein, has been entered into directory "test/exp".

To execute the command, open a window to directory "src/str", and type the following line to the macOS (or linux) prompt.

```
%  ereg test 1pga
```

The program outputs files "test/seq/seq.1pga", "test/tor/tor.1pga.CNF" for CNF=$\{00, 01, 02, ..., 05\}$, "test/car/1pga.CNF.pdb" for CNF=$\{00, 01, 02, ..., 05\}$, and "test/dgn/ereg.1pga".

In the following abbreviated listing, frame boxes enclose descriptions of the output data.

---

file="test/dgn/ereg.1pga"

Diagnostic file created by the command:
% ereg test 1pga

```
REGULARIZE GEOMETRY
 heavy atom RMSD=   0.124
```

For CNF=00, created by geometry regularization with no consideration of energy.

```
CONSTRUCT MECHANICAL SYSTEM
 subset name=a00
```

Variable "subset name" is a name assigned by the program to the subset of degrees of freedom with respect to which energy is being minimized. For a protein chain composed of less than about 150 residues, a single subset is sufficient to contain all torsion angle degrees of freedom.

```
   nZ0   nS0
     1     0
   cQ1 cQ1bb   cU1   cJ1   cY1
   322   168   154   205   145
   cQ2 cQ2bb
   322   168
   cF1   cG2   cB2   cH1   cG1   cB1
  3970  3938    32  1750  3938    32
   cQ3   cX2
   322   321
   cE0   cC1
 51681  2126
```

Diagnostic set sizes. For program design utility, set names consist of a 2-character (capital letter+number) combination. Set name definitions can be found in file "src/str/dimensions". For example, "Z0" is the set of chains in the system, "Q1" is the set of torsion degrees of freedom. This information, which has utility for algorithm development, is best ignored for application use.

```
 Z0sub
  0
 iR0 R0aa Q0sub
   1 eMET  111111
   2 THR   111111
   3 TYR   111111
   4 LYS   11111111
   5 LEU   1111111
   6 ILE   1111111
  .
  .
  .
  51 THR   111111
  52 PHE   11111
  53 THR   111111
  54 VAL   111111
  55 THR   111111
  56 GLUe  111111
```

A profile of the set of torsion angle degrees of freedom. Here, chain translation +rotation is not used, all torsion degrees of freedom are variable. For each residue, the order of torsion angles used in mechanical system objects can be found in data file "src/dat/residue_mappings", input to variable "Q0tor".

```
CONTRACT MECHANICAL SYSTEM
 subset name=a00
   cQ2 cQ2bb   cF2   cG3   cB3   cH2   cJ2   cY2   cE1   cC1
   322   168  3870  3841    29  1750   205   145 51681  2126
```

Diagnostic set sizes following mechanical system contraction, meaning energy contributions that can not change using this subset of degrees of freedom are removed.

```
MINIMIZE RESTBCHMGP LOCALLY
 distance constraint coeff index= 0
```

Affecting energy component Fc, the weighting coefficient for the sum of harmonic distance constraints is set at 1.00 kcal/(mol-bohr^2). Weights corresponding to other indexes can be found in data file "src/dat/energy_params", input to program variable "W0a".

```
 subset name=a00
 steps remaining=189 steps taken= 67 endstate=0
```

Variable "steps remaining" is a counter, initiated as the maximum number of steps, and decremented following each step. A trajectory ends either when the RMS of the gradient falls below a threshold value, or when "steps remaining"=0.

For variable "endstate", a nonzero value indicates abnormal termination.

```
  i        F          z          b       lam2         s2         d2       delF
  0-7.99571e+02 4.0746e+01 4.4800e-02 7.1332e+01 4.4806e-02 4.4806e-02-8.58701e+02
  1-8.46142e+02 5.7753e+00 5.6000e-02 2.9667e+01 5.6242e-02 9.3246e-02-8.81492e+02
  2-8.77677e+02 3.3079e+00 4.7040e-02 3.3128e+01 4.7050e-02 1.3580e-01-8.97303e+02
  3-8.93028e+02 1.5844e+00 2.2760e-02 6.4990e+01 2.2761e-02 1.5579e-01-9.02125e+02
  4-8.99371e+02 1.3630e+00 4.2675e-02 2.5215e+01 4.2862e-02 1.8335e-01-9.12201e+02
  5-9.09322e+02 9.5596e-01 2.1508e-02 3.8589e+01 2.1513e-02 2.0177e-01-9.14709e+02
```

```
 6-9.08121e+02 2.2529e+00 2.7100e-02 3.3359e+01 2.7186e-02 2.0850e-01-9.16485e+02
 7-9.09428e+02 8.5044e-01 8.0000e-02 8.1993e+00 8.0002e-02 8.0002e-02-9.26029e+02
 8-9.25396e+02 2.5275e+00 8.0000e-02 6.1683e+00 8.0001e-02 1.5482e-01-9.38978e+02
 .
 .
 .
60-9.79878e+02 5.4086e-03 1.6334e-03 0.0000e+00 1.5834e-03 1.5834e-03-9.79878e+02
61-9.79878e+02 4.0550e-03 1.3720e-03 5.6838e-02 1.3721e-03 7.9961e-04-9.79878e+02
62-9.79878e+02 3.1983e-03 1.2544e-03 0.0000e+00 1.1613e-03 1.1613e-03-9.79878e+02
63-9.79878e+02 2.2008e-03 1.0537e-03 4.7840e-02 1.0537e-03 6.4566e-04-9.79878e+02
64-9.79878e+02 2.0083e-03 9.6341e-04 0.0000e+00 9.4775e-04 9.4775e-04-9.79878e+02
65-9.79878e+02 1.5516e-03 8.0926e-04 5.7892e-02 8.0928e-04 5.4109e-04-9.79878e+02
66-9.79878e+02 1.2253e-03 7.3990e-04 0.0000e+00 6.7416e-04 6.7416e-04-9.79878e+02
67-9.79878e+02 8.7011e-04
```

At each point of the trajectory, the program calculates energy, 1st derivatives, and 2nd derivatives. These quantities define a harmonic approximation to the actual energy surface. Steps are calculated using Newton's method to minimize the harmonic surface within a trust region.

For the above compact characterization of the minimization trajectory, column labels are defined as follows:

i Minimization step.

F Total energy.

z RMS of the gradient.

b Radius of the trust region, or equivalently RMS of a step to the boundary of the trust region.

lam2 Lagrange multiplier used to enforce the condition that the step minimize the harmonic surface on the boundary of the trust region. A value of zero indicates the minimum of the harmonic surface is inside of the boundary of the trust region.

s2 RMS of the calculated step.

delF Value of F estimated by the harmonic surface at the position of the calculated step.

At each step i>0, the radius of the trust region, b(i), is adjusted based on a comparison of the actual change in energy, F(i)−F(i−1), to the decrease in energy, delF(i−1)−F(i−1), predicted by the harmonic surface of the previous step.

```
      F
 -979.88
    Fr       Fe       Fs       Ft       Fb       Fc       Fh       Fw       Fg       Fp
  335.72 -1195.62     0.00    -4.33     0.00    44.49   -81.71   -43.66     0.00   -34.78
```

Decomposition of energy at the endpoint of local minimization.

Variable F is the sum of components (Fr+Fe+Fs+Ft+Fb+Fc+Fh+Fw+Fp), where Fw is a greatly simplified estimate of Fm based on volume exclusion from a hydration shell.

Affecting component Fe, charges on ionized groups are scaled to 9/64 of full values.

```
SINGLE POINT RESTBCHMGP EVALUATION
```

```
  nA5   nB5   nC5   nH5   nV5   nE5   nE6   nF5   nF6   nF7
    6    10     6    14    18    18    18     6     9     5
 nDOT   nD5 nHIT
  740    57   10
  nA5   nB5   nC5   nH5   nV5   nE5   nE6   nF5   nF6   nF7
 1180 40364   638  1856  1914  1914  1896   638   948   318
 nDOT   nD5 nHIT
39868  3060  920
```

Diagnostic set sizes, following generation of a molecular surface, and partitioning of the surface into elements.

```
     F
-1562.12
    Fr       Fe       Fs       Ft       Fb       Fc       Fh       Fm       Fg       Fp
  335.72 -1407.63     0.00    -4.33     0.00    44.49   -81.71  -395.04    28.69   -37.82
```

Decomposition of full energy at the endpoint of local minimization.

Charges on ionized groups are restored to full values. F is the sum of all energy components excluding Fc.

Single point evaluation of Fm. When placed in a dielectric continuum, the charge distribution of the protein (or, more generally, the system of molecules) induces a charge distribution on the boundary surface between the protein and the dielectric continuum. Fps is the interaction energy of the charge distribution of the protein with the charge distribution of the boundary surface. Fss is the interaction energy of the charge distribution of the boundary surface with itself. Fm is calculated as a linear combination of components Fps and Fss.

```
heavy atom RMSD=   0.371
```

For CNF="01", the endpoint of local minimization.

```
distance constraint coeff index=-1
```

The weighting coefficient for the sum of harmonic distance constraints is assigned successively smaller values for minimization trajectories CNF=$\{02, 03, 04, \ldots, 09\}$.

```
 .
 .
 .
     F
-1602.75
    Fr       Fe       Fs       Ft       Fb       Fc       Fh       Fm       Fg       Fp
  326.28 -1405.12     0.00    -4.24     0.00    28.23   -83.21  -394.97    13.54   -55.03
heavy atom RMSD=   0.420
```

For CNF=02.

```
 .
 .
 .
     F
-1620.00
    Fr       Fe       Fs       Ft       Fb       Fc       Fh       Fm       Fg       Fp
  321.04 -1397.13     0.00    -2.85     0.00    17.87   -85.75  -399.19    13.54   -69.66
heavy atom RMSD=   0.522
```

For CNF=03.

```
  .
  .
  .
      F
 -1628.57
     Fr       Fe       Fs       Ft       Fb       Fc       Fh       Fm       Fg       Fp
   318.88 -1394.16     0.00    -1.24     0.00     7.64   -87.33  -395.01     5.39   -75.09
 heavy atom RMSD=   0.581
```

For CNF=04.

```
  .
  .
  .
      F
 -1630.58
     Fr       Fe       Fs       Ft       Fb       Fc       Fh       Fm       Fg       Fp
   316.11 -1391.82     0.00    -1.65     0.00     4.95   -90.45  -397.97     7.39   -72.19
 heavy atom RMSD=   0.788
```

For CNF=05.

```
  .
  .
  .
      F
 -1623.81
     Fr       Fe       Fs       Ft       Fb       Fc       Fh       Fm       Fg       Fp
   318.80 -1394.10     0.00    -0.97     0.00     2.61   -93.36  -399.97    19.94   -74.15
```

Because the full energy (excluding Fc) increases, the sequence of local minimization trajectories is ended, and the endpoint of this final minimization is not accepted.

```
COMPACT ANALYSIS
 W0     F     RMSD
  0 -1562.12   0.37
 -1 -1602.75   0.42
 -2 -1620.00   0.52
 -3 -1628.57   0.58
 -4 -1630.58   0.79
    --------
      -68.47
 W0     Fr       Fe       Fs       Ft       Fb       Fc       Fh       Fm       Fg       Fp
  0   335.72 -1407.63     0.00    -4.33     0.00    44.49   -81.71  -395.04    28.69   -37.82
 -1   326.28 -1405.12     0.00    -4.24     0.00    28.23   -83.21  -394.97    13.54   -55.03
 -2   321.04 -1397.13     0.00    -2.85     0.00    17.87   -85.75  -399.19    13.54   -69.66
 -3   318.88 -1394.16     0.00    -1.24     0.00     7.64   -87.33  -395.01     5.39   -75.09
 -4   316.11 -1391.82     0.00    -1.65     0.00     4.95   -90.45  -397.97     7.39   -72.19
    -------- -------- -------- -------- -------- -------- -------- -------- -------- --------
      -19.62    15.82     0.00     2.67     0.00   -39.54    -8.75    -2.92   -21.30   -34.37
```

A compact summary of how the full energy components and RMSD change with minimization.

The local energy minimum conformation is characterized by a heavy atom RMSD from the experimental structure of .7 Å, and by a total energy of -1630.58 kcal/mol.

6.3 estp

FUNCTIONALITY
segment structure prediction +sequence design

SYNTAX
estp FAM MOL CNF SUB

INPUT FILES	OUTPUT FILES
/FAM/seq/seq.MOL	/FAM/dgn/estp.MOL.CNF.SUB
/FAM/car/MOL.CNF.pdb	/FAM/car/MOL.CNF_SUB_???.pdb
/FAM/stp/stp.SUB.MOL	/FAM/tor/tor.MOL.CNF_SUB_???
	/FAM/car/MOL.v???.pdb

SUMMARY

1) inputs a sequence, "FAM/seq/seq.MOL"

2) inputs an initial conformation in pdb format, "FAM/car/MOL.CNF.pdb"

3) inputs a compact file, "FAM/stp/stp.SUB.MOL", specifying the subset of degrees of freedom to be searched and, at each sequence position, the set of residues to be substituted

4) for each sequence variant, generates a collection of local minima on the energy surface by minimizing globally with respect to a subset of torsion angle degrees of freedom concentrated in one or more chain segments

5) for each local energy minimum conformation, here using ??? to denote 3 digits specifying the order of the minimum based on energy, outputs cartesian coordinates in pdb-format file "FAM/car/MOL.CNF_SUB_???.pdb", and torsion angle coordinates in file "FAM/tor/tor.MOL.CNF_SUB_???"

6) outputs, in file "FAM/dgn/estp.MOL.CNF.SUB", a diagnostic summary of the execution of the command

7) for each of the most stable sequences found in the search through sequence space, here using ??? to denote the order of the sequence variant based on calculated free energy of folding, outputs the lowest-energy conformation in pdb-format file "FAM/car/MOL.v???.pdb"

NOTES

Two of the most useful functionalities of the ereg program are accessed using the "estp" command. Structure prediction for segments of proteins is achieved by search through conformation space. Sequence design of thermodynamic stability is achieved by additional search through sequence space.

Using conformation CNF as a starting structure, the "estp" command searches for the conformation that minimizes the full energy function, globally, within a subspace of the full space of motion for the rigid geometry system MOL. The subspace SUB of conformations to be searched consists of from 1 to 8 segments, each segment 7 to 13 residues in length, plus a collection of side chains.

The search through the specified subspace of conformations consists of the following sequence of calculations. A segment, or segments, of the initial backbone structure is deformed, analytically, resulting in a large collection of alternative backbone structures. Local energy minimization is applied, resulting in a collection of optimized backbone structures. Interactions involving side chains that will be adjusted later are not yet included in the energy. For each optimized backbone structure, one optimized side chain structure is added, the result of a trajectory of local minima search through the sub-subspace of side chain motion. Local energy minimization is applied to the backbone plus side chain combination, resulting in a collection of local minima for the deformed segment, or segments.

To achieve efficiency in applications to protein surface loops, local energy minimization uses distance constraints between rigid segments to maintain fixed the structure of the region outside of the segments specified as variable. Efficiency is enhanced further by original algorithms for deforming segments of the protein chain, and for removing overlaps from starting conformations.

In the default mode of execution, the subset of degrees of freedom specified in file "test/stp/stp.SUB.MOL" is expanded, following input to the program, to include side chain torsions for all side chains that can contact a deformable segment. This automation of subset expansion has been found preferable in comparison to manual subset expansion, which requires of the user graphical examination of the initial structure. Also specified in file "test/stp/stp.SUB.MOL" is the set of residue types allowed at each sequence position. In the default mode of execution, the space of sequences searched consists of all possible combinations. For example, 5 allowed residue types at 3 sequence positions produces 125 possible combinations. The 1st character of command line argument SUB is used to control 3 variant modes of execution. This mechanism of control has been found more convenient in comparison to addition of a command line argument. For 1st character ’l’ or ’h’, automated subset expansion is skipped, thus allowing complete subset specification by manual preparation of file "test/stp/stp.SUB.MOL". For 1st character ’h’ or ’i’, the combinatorial search through sequence space skips over all but the initial and final sequences, thus allowing a fast comparison of the stability of 2 sequences. A common use is for evaluation of the stability of a mutant sequence in comparison to native.

For each sequence of the specified space of sequences, the program searches through the specified space of conformations. The lowest-energy conformation found is used to evaluate dG, an estimate of free energy of folding. Because different models are used to represent the folded state of the protein and the reference unfolded state, dG is not, for a single sequence, a physically meaningful measure of stability. However, ddG=(dG(sequence1) $-$dG(sequence0)), the change in dG with sequence, does provide a meaningful measure of relative stability.

The free energy of the folded state is calculated as (c_{pp}(Fr +Fe +Fs +Ft +Fb +Fc +Fg +Fp) $+c_h$Fh $++c_{ps}$Fps $+c_{ss}$Fss), where { c_{pp}, c_h, c_{ps}, c_{ss}} are coefficients optimized such that calculated ddG matches observed ddG over a large dataset of experimental measurements, and where energy components are evaluated for the lowest-energy conformation found in the search through conformation space. The free energy of the unfolded state, Gunf, is estimated as the free energy of an Ising model representation of the unfolded state.

EXAMPLE TEST CASE

File "stp.lp0.1pga", a subset of torsion angle degrees of freedom corresponding to a surface segment of 1pga, has been created by hand in directory "test/stp". Files "test/seq/seq.1pga" and "test/car/1pga.05.pdb" were created by the previous execution of the test case for the "ereg" command.

To execute the command, open a window to directory "src/str", and type the following line to the macOS (or linux) prompt.

```
%  estp test 1pga 05 lp0
```

The lowest-energy structure found in the search recovers the backbone conformation of the crystal structure. The side chain conformation changes at residues VAL 39 and ASP 40.

Using an Apple M1 Pro processor, this application of global energy minimization to the 7-residue surface segment (36–42) of 1pga requires computation time of about 5.4 hours.

In the following abbreviated listing, frame boxes enclose descriptions of the output data.

---

file="test/dgn/estp.1pga.05.lp0"

Diagnostic file created by the command:
% estp test 1pga 05 lp0

```
REGULARIZE GEOMETRY
 heavy atom RMSD=   0.000
```

In this example, the input conformation "test/car/1pga.05.pdb" is already regularized. The small RMSD results from a truncation of atom positions imposed by storage of the conformation as a pdb format file.

```
SEARCH COMBINATORIAL SEQUENCE SPACE
```

This output line marks entry into the program’s outer loop, which controls search through sequence space. In this example, the functionality being accessed is structure prediction, as opposed to the more general functionality of sequence design. File "test/stp/stp.lp0.1pga" directs search through conformation space for a single 7-residue segment. Because the input sequence "test/seq/seq.1pga" is not changed, the program outer loop reduces to a single iteration.

```
MINIMIZE RESTBCHMGP GLOBALLY
```

For each iteration of the outer loop, this output line marks entry into search through conformation space.

```
CONSTRUCT MECHANICAL SYSTEM
 subset name=lp0
```

The fourth command line argument is used as the name for this initial subset of degrees of freedom.

```
   nZ0   nS0
     1     0
   cQ1 cQ1bb   cU1   cJ1   cY1
   322   168   154   205   145
   cQ2 cQ2bb
    31    22
   cF1   cG2   cB2   cH1   cG1   cB1
  3970  1533  2437  1750  3938    32
   cQ3   cX2
    79    79
   cE0   cC1
  1864    24
```

Diagnostic set sizes.

```
 Z0sub
  0
 iR0 R0aa Q0sub
   1 eMET  000000
   2 THR   000000
   3 TYR   000000
 4 LYS   00000000
 5 LEU   0000000
 6 ILE   0000000
 7 LEU   0000000
 8 ASN   000000
 9 GLY   000
10 LYS   00000000
11 THR   000000
12 LEU   0010000
13 LYS   00000000
14 GLY   000
15 GLU   000000
16 THR   000000
17 THR   000000
18 THR   000000
19 GLU   000000
20 ALA   0000
21 VAL   000000
22 ASP   00000
23 ALA   0000
24 ALA   0000
25 THR   000000
26 ALA   0000
27 GLU   000000
28 LYS   00000000
29 VAL   000000
30 PHE   00000
31 LYS   00000000
32 GLN   0000000
33 TYR   000000
34 ALA   0000
35 ASN   001000
36 ASP   11101
37 ASN   111001
38 GLY   111
39 VAL   111001
40 ASP   11101
41 GLY   111
42 GLU   111001
43 TRP   00000
44 THR   000000
45 TYR   000000
46 ASP   00000
47 ASP   00000
48 ALA   0000
49 THR   000000
50 LYS   00000000
51 THR   000000
52 PHE   00000
53 THR   000000
54 VAL   001000
55 THR   000000
56 GLUe  001000
```

A profile of the set of torsion angle degrees of freedom. Chain translation +rotation is not used. Torsion degrees of freedom include all backbone torsions for the 7-residue segment {36 ASP–42 GLU}. Also included are the χ_1 torsions for residues of the segment, and for residues {12 LEU, 35 ASN, 54 VAL, 56 GLUe} whose side chains contact the segment. Inclusion of the χ_1 torsions provides a mechanism for neglect of side chain interactions, beyond the C$^\beta$ atom, in the search over segment backbone conformations.

```
 subset name=lp0
   cQ2 cQ2bb   cF2   cG3   cB3   cH2   cJ2   cY2   cE1   cC1
    31    22  3718  1345  2373  1645    17     5  1190    24
 iJ3=    0  [ 0]
```

Diagnostic output following mechanical system contraction.

```
BACKBONE DEFORMATIONS, segment= 0
  def -kT*ln(p)  dchi     cnf0
```

Column headings:

def — index of deformation in order of probability, the initial undeformed conformation is assigned index 0

-kT*ln(p) — estimate of energy of deformation based on probabilities of occurrence of dipeptide conformations

dchi — RMS difference (in degrees) of ϕ and ψ torsion angles from nondeformed conformation

cnf0 — sequence of single residue conformational regions

```
    0    10.14    0.0  A A A*E E E E
    1     9.20   25.4  A C E*E A E*E
    2     9.62   22.4  A C E*X E E E
    3     9.72   24.6  A A E*X*X E*E
  .
  .
  .
  220    14.19   27.9  E A*E*X X E E
```

In this example, a single segment is deformed. Deformations are exact meaning coordinates for adjacent nondeformed segments are unchanged. Coverage of the space of segment deformations is approximately uniform. The initial conformation is included as the first deformation before ordering. For each new deformation, RMS distance from all previously accepted deformations is required to be greater than 9 degrees.

```
BACKBONE OVERLAP REMOVAL, SCORE= 0.500[restbhw] +SCREEN
  iD2     tot   nE3      S1     S2     S3     S4     S5      Z2iD1
 2048     151.2   0    -82.9   10.1  -35.5  234.1   25.4     0
  cnf1=[A A A*E E E E ]
  cnf0=[A A A*E E E E ]
 2049     180.9   0    -63.0    9.2  -19.6  228.4   25.9     1
  cnf1=[A E E*E X E*E ]
  cnf0=[A C E*E A E*E ]
 2050     177.8   0    -65.1    9.6  -19.1  231.1   21.2     2
  cnf1=[A E E*X E E E ]
  cnf0=[A C E*X E E E ]
 2051     218.5   0    -69.1    9.7  -24.6  249.7   52.8     3
  cnf1=[A A E*X*X E*E ]
  cnf0=[A A E*X*X E*E ]
  .
  .
  .
 2267     178.7   0    -58.6   14.2  -31.4  236.5   18.0   220
  cnf1=[E A*E*X X E*E ]
  cnf0=[E A*E*X X E E ]
```

tot=(S1+S2+S3+S4+S5) is a score evaluated at the endpoint of overlap removal for the purpose of ordering deformations, enabling selection of those likely to be most productive.

S1 .5(Fr+Fe+Fs+Ft+Fb+Fh+Fw), using charges on ionized groups scaled to 9/64 of full values

Components S2-S5 are a fast screen.

S2 energy contributed by local conformation
S3 energy contributed by residue-residue contacts
S4 energy contributed by ionized groups
S5 energy contributed by hydrophobic groups

nE3 number of overlapping atom pairs remaining following overlap removal
cnf1 sequence of conformational regions at the endpoint of overlap removal

Deformations for which overlaps can not be removed are excluded.

For consistency with the more general case of multiple segment deformations, the index used here for "iD2" is a temporary value, indexing scratch locations of "D2", the set of combined backbone deformations. For the case of multiple segment deformations, search through all combinations of individual segment deformations produces large numbers of combined backbone deformations. These combined deformations are processed, using overlap removal and evaluation of screen score, in groups of 2048, before being assimilated into the ordered set of "D2" locations.

```
SORTED ORDER
  iD2     tot       S1     S2     S3     S4     S5      Z2iD1
    0    151.22   -82.9   10.1  -35.5  234.1   25.4     0
  cnf1=[A A A*E E E E ]
    1    161.51   -74.5   12.3  -38.7  232.6   29.9   104
  cnf1=[X X E*E E X E ]
    2    174.66   -80.4   13.3  -38.2  232.9   47.0   177
  cnf1=[X A A*E E*X E ]
    3    177.78   -72.0   12.4  -34.2  237.7   33.9   116
  cnf1=[X X E E X*E*E ]
  .
  .
  .
  190    252.14   -55.7   12.7  -22.9  262.7   55.4   134
  cnf1=[A A E*E A A*E ]
```

The number of overlap free deformations has been reduced slightly by clustering. For single segment searches, the number of deformations passed to the next stage of local minimization on the energy surface is limited to 512.

```
BACKBONE LOCAL ENERGY MINIMIZATION,400 STEPS, SCORE= 0.750[restbchm] +SCREEN
 iD2=    0 Z2iD1=   0
 steps taken=351 rms of gradient= 8.9523e-04
   S      S1     S2     S3     S4     S5
 -133.8 -253.6    5.1  -19.1  120.1   13.7
   Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
   23.3 -175.5    0.0    5.6    0.0    0.0   -9.7 -181.8    0.0    0.0
  cnf2=[A A A*E E E*E ]
  cnf0=[A A A*E E E E ]
 iD2=    1 Z2iD1= 104
 steps taken=304 rms of gradient= 1.3855e-03
   S      S1     S2     S3     S4     S5
 -143.5 -258.2    6.1  -22.6  115.6   15.5
   Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
   18.6 -167.8    0.0    6.1    0.0    0.0  -16.9 -184.2    0.0    0.0
  cnf2=[C X A*E E E E ]
  cnf0=[X A E*X E A E ]
 iD2=    2 Z2iD1= 177
 steps taken=335 rms of gradient= 4.9517e-03
   S      S1     S2     S3     S4     S5
 -143.0 -258.2    6.6  -22.6  115.6   15.5
   Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
   18.6 -167.8    0.0    6.1    0.0    0.0  -16.9 -184.2    0.0    0.0
  cnf2=[C X A*E E E E ]
  cnf0=[X X A*E E*A E ]
  .
  .
  .
 iD2=  190 Z2iD1= 134
 steps taken=406 rms of gradient= 6.5578e-01
   S      S1     S2     S3     S4     S5
 -102.5 -247.1    6.3  -12.8  118.5   32.6
   Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
   24.1 -161.9    0.0    5.3    0.0    0.0  -10.4 -186.7    0.0    0.0
  cnf2=[A A E*E*X E*E ]
  cnf0=[A A E*E*C X*E ]
```

A local minimization trajectory, limited to approximately 400 steps, is generated on the [restbchw] energy surface with charges on ionized groups scaled to 9/64 of full values. At the endpoint of local minimization, (Fe +Fm) are evaluated with charges on ionized groups restored to full values. Variable "rms of gradient" provides a measure of the extent of convergence.

S=(S1+S2+S3+S4+S5) is a score, evaluated at the endpoint of the local minimization trajectory, enabling selection of those conformations most likely to be productive.

S1= .75×(Fr+Fe+Fs+Ft+Fb+Fc+Fh+Fm).

Components S2-S5 are described above.

cnf2= sequence of conformational regions following energy minimization.

Deformation 53 minimizes to a sequence of conformational regions identical to the initial conformation, indicating that the extent of coverage of backbone deformations is probably sufficient.

For single segment searches, the number of backbone conformations passed on to complete minimization is limited to 128.

```
BACKBONE LOCAL ENERGY MINIMIZATION,800 STEPS, SCORE= 1.000[restbchm] +SCREEN
 iD2=    0 Z2iD1= 104
 steps taken=  0 rms of gradient= 1.3056e-03
   S      S1     S2     S3     S4     S5
 -229.6 -344.2    6.1  -22.6  115.6   15.5
   Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
   18.6 -167.8    0.0    6.1    0.0    0.0  -16.9 -184.2    0.0    0.0
  cnf2=[C X A*E E E E ]
  cnf0=[X A E*X E A E ]
 iD2=    1 Z2iD1= 220
 steps taken= 23 rms of gradient= 3.8844e-04
   S      S1     S2     S3     S4     S5
 -226.8 -336.7    7.1  -22.4  114.0   11.2
   Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
   18.3 -160.5    0.0    6.6    0.0    0.0  -16.0 -185.0    0.0    0.0
  cnf2=[C X E*X E E E ]
  cnf0=[E A*E*X X E E ]
 iD2=    2 Z2iD1=  16
 steps taken=100 rms of gradient= 3.0336e-04
   S      S1     S2     S3     S4     S5
 -225.1 -347.0    5.3  -21.3  115.8   22.1
   Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
   19.8 -168.9    0.0    4.3    0.0    0.0  -17.2 -184.9    0.0    0.0
  cnf2=[A A A*E E E*E ]
  cnf0=[A A E*X*X*A E ]
  .
  .
  .
 iD2=   87 Z2iD1= 195
 steps taken=693 rms of gradient= 4.0857e-04
   S      S1     S2     S3     S4     S5
 -203.0 -340.2    6.8  -24.0  125.0   29.5
   Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
   20.8 -173.6    0.0    8.5    0.0    0.0  -16.3 -179.5    0.0    0.0
  cnf2=[C X A*X E A E ]
  cnf0=[E A*E*E E A E ]
```

The number of backbone conformations passed on to side chain conformational search is limited to 64.

```
DEFORMATIONS FOR MODE=bb
Number of conformations=  56
 iD0=   0   S=?? initial    cnf=[A A A*E E E E ]
 S1=??        S2=??        S3=??        S4=??        S5=??
LEU      -68.76 116.28 172.49-169.88 175.40  60.00  60.00
ASN      -52.76 -51.31-176.53 -87.92-143.39-177.53
ASP  A   -53.16 -39.85-170.00 -77.09  -9.50
ASN  A  -101.45  27.45 166.59 -73.93 -92.96-179.34
GLY  A*   68.69  32.87 169.21
VAL  E  -118.40 104.78 176.89-174.31  60.00  60.00
ASP  E  -154.68 109.49 175.89-176.04  -5.91
GLY  E  -155.35-103.25-179.32
GLU  E   -98.54 124.86-178.49-118.19-177.98 -61.52
VAL     -120.53 120.67-176.38  53.05  60.00  60.00
```

```
GLUe     -99.80 114.75        -68.39-147.06 -97.31
 iD0=   1   S=-2.29553e+02  cnf=[C X A*E E E E ]
 S1=  -344.20 S2=     6.14 S3=   -22.59 S4=   115.63 S5=    15.47
LEU      -68.76 116.28 172.49-169.88 175.40  60.00  60.00
ASN      -52.76 -51.31-179.40 -87.92-143.39-177.53
ASP  C   -90.34  62.65-175.11 -77.09  -9.50
ASN  X  -174.72 -47.64 171.71 -73.93 -92.96-179.34
GLY  A*  111.93  48.87-178.75
VAL  E  -153.50  57.55 179.96-174.31  60.00  60.00
ASP  E  -146.63 140.92 177.25-176.04  -5.91
GLY  E  -168.51 -97.94 164.51
GLU  E   -72.94 123.03-178.49-118.19-177.98 -61.52
VAL     -120.53 120.67-176.38  53.05  60.00  60.00
GLUe     -99.80 114.75        -68.39-147.06 -97.31
 iD0=   2   S=-2.26769e+02  cnf=[C X E*X E E E ]
 S1=  -336.66 S2=     7.09 S3=   -22.43 S4=   113.99 S5=    11.23
LEU      -68.76 116.28 172.49-169.88 175.40  60.00  60.00
ASN      -52.76 -51.31-179.10 -87.92-143.39-177.53
ASP  C   -88.67  62.83-174.40 -77.09  -9.50
ASN  X  -173.96 -51.33 175.83 -73.93 -92.96-179.34
GLY  E*  116.45  72.54-162.11
VAL  X  -158.15 -48.65-168.95-174.31  60.00  60.00
ASP  E   -61.13 133.16 171.95-176.04  -5.91
GLY  E  -150.15-113.60 161.23
GLU  E   -65.18 124.70-178.49-118.19-177.98 -61.52
VAL     -120.53 120.67-176.38  53.05  60.00  60.00
GLUe     -99.80 114.75        -68.39-147.06 -97.31
  .
  .
  .
 iD0=  55   S=-1.82400e+02  cnf=[A A E*C C*E*E ]
 S1=  -327.14 S2=     7.07 S3=   -14.65 S4=   123.95 S5=    28.36
LEU      -68.76 116.28 172.49-169.88 175.40  60.00  60.00
ASN      -52.76 -51.31 179.28 -87.92-143.39-177.53
ASP  A   -39.82 -46.04-168.02 -77.09  -9.50
ASN  A  -116.16  26.08-179.93 -73.93 -92.96-179.34
GLY  E*   89.05 100.40-164.02
VAL  C   -97.72  54.10 171.02-174.31  60.00  60.00
ASP  C*   54.56-106.61-177.84-176.04  -5.91
GLY  E*  164.90 -45.99 175.66
GLU  E  -101.84 122.83-178.49-118.19-177.98 -61.52
VAL     -120.53 120.67-176.38  53.05  60.00  60.00
GLUe     -99.80 114.75        -68.39-147.06 -97.31
```

Local energy minima conformations for the segment backbone. The initial conformation is included, associated with index 0. No screen score has been evaluated for the initial conformation. Other conformations are ordered based on screen score.

```
SIDE CHAIN GLOBAL ENERGY MINIMIZATION, SCORE= [resthm]
```

For each backbone conformation, a single side chain conformation is generated by minimization of the energy function (Fr+Fe+Fs+Ft+Fb+Fh+Fm) with respect to side chain torsions for residues of the deformable segment and for the 4 side chains that contact the segment.

If, for a given backbone conformation, the set of variable side chains can be partitioned into independent packing units, then the search through the combinatorial space of side chain conformations is directed as a sequence of smaller searches though the packing units. Packing units are assigned based on a clustering of side chain ellipsoids, with a maximum size limitation of 12 side chains.

For each packing unit, a collection of combined rotamer conformations is generated using a dead-end elimination algorithm to minimize energy over a multidimensional lattice.

A better scoring of rotamer combinations is obtained by moving off lattice, using local energy minimization with respect to variable side chain torsions. Energy components are evaluated for end points of local minimization trajectories, starting from the combined rotamer conformations of the collection.

For each backbone conformation and packing unit, the choice of a rotamer combination to be carried forward is selected randomly from a Boltzmann probability distribution over the collection of local minima.

```
  .
  .
  .
CONSTRUCT MECHANICAL SYSTEM
 subset name=lp0
   nZ0   nS0
     1     0
   cQ1 cQ1bb   cU1   cJ1   cY1
   322   168   154   205   145
   cQ2 cQ2bb
    48    22
   cF1   cG2   cB2   cH1   cG1   cB1
  3970  1533  2437  1750  3938    32
   cQ3   cX2
    96    96
   cE0   cC1
  3361    24
 Z0sub
  0
 iR0 R0aa Q0sub
   1 eMET  000000
   2 THR   000000
   3 TYR   000000
   4 LYS   00000000
   5 LEU   0000000
   6 ILE   0000000
   7 LEU   0000000
   8 ASN   000000
   9 GLY   000
  10 LYS   00000000
  11 THR   000000
  12 LEU   0011110
  13 LYS   00000000
  14 GLY   000
  15 GLU   000000
  16 THR   000000
  17 THR   000000
  18 THR   000000
  19 GLU   000000
  20 ALA   0000
  21 VAL   000000
  22 ASP   00000
  23 ALA   0000
  24 ALA   0000
  25 THR   000000
  26 ALA   0000
  27 GLU   000000
  28 LYS   00000000
  29 VAL   000000
  30 PHE   00000
  31 LYS   00000000
  32 GLN   0000000
  33 TYR   000000
  34 ALA   0000
  35 ASN   001110
  36 ASP   11111
  37 ASN   111111
  38 GLY   111
  39 VAL   111111
  40 ASP   11111
  41 GLY   111
  42 GLU   111111
  43 TRP   00000
  44 THR   000000
  45 TYR   000000
  46 ASP   00000
  47 ASP   00000
  48 ALA   0000
  49 THR   000000
  50 LYS   00000000
  51 THR   000000
  52 PHE   00000
  53 THR   000000
  54 VAL   001110
  55 THR   000000
  56 GLUe  001110
```

The subset of degrees of freedom includes both backbone and side chain torsions.

```
 subset name=lp0
   cQ2 cQ2bb   cF2   cG3   cB3   cH2   cJ2   cY2   cE1   cC1
    48    22  3870  1497  2373  1750    26    15  3361    24
BACKBONE +SIDE CHAIN LOCAL ENERGY MINIMIZATION,800 STEPS, SCORE= [restbchm]
 iD0=   0
 steps taken= 21 rms of gradient= 1.6399e-03
   tot    Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
 -637.0   47.1 -197.1    0.0   -2.9    0.0    0.0  -26.5 -398.7   13.5  -72.3
 iD0=   1
 steps taken=330 rms of gradient= 2.3532e-03
   tot    Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
 -634.1   44.8 -185.8    0.0    1.9    0.0    0.0  -28.1 -397.0    4.5  -74.3
 iD0=   2
 steps taken=136 rms of gradient= 4.8191e-04
   tot    Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
 -632.1   48.0 -176.0    0.0   -1.1    0.0    0.0  -29.5 -407.9    4.9  -70.6
  .
  .
  .
 iD0=  55
 steps taken=251 rms of gradient= 3.4196e-03
   tot    Fr     Fe     Fs     Ft     Fc     Fb     Fh     Fm     Fg     Fp
 -595.5   75.0 -150.8    0.0   -7.0    0.0    0.0  -18.0 -438.5    1.6  -57.8
```

tot=(Fr+Fe+Fs+Ft+Fb+Fc+Fh+Fm+Fg+Fp).

Energy evaluation is at the end point of local minimization on the simplified energy surface [restbchwp].

```
DEFORMATIONS FOR MODE=lp
Number of conformations=  14
 iD0=   0   F=-6.47972e+02  cnf=[A B A*E D D E ]
    Fr      Fe      Fs      Ft      Fc      Fb      Fh      Fm      Fg      Fp
   47.03 -173.08    0.00   -3.83    0.00    0.00  -29.64 -422.55    6.64  -72.54
LEU      -68.76 116.28 172.49-172.19 172.44  51.04  81.76
ASN      -52.76 -51.31-176.44 -93.49 -99.70-178.73
ASP  A   -56.60 -46.02-170.88-175.93  -1.86
ASN  B   -92.63  28.00 163.49 -73.75 -94.19-179.00
GLY  A*   65.23  37.13 167.42
VAL  E  -135.44 118.49 172.14  39.65  57.44  52.11
ASP  D  -176.05 130.26 177.52  56.47 101.57
GLY  D  -162.89 -85.46 175.20
GLU  E   -98.65 122.47-178.49-156.22 176.24-163.85
VAL     -120.53 120.67-176.38  50.61  61.55  59.70
GLUe     -99.80 114.75        -69.55-131.59  65.01
 iD0=   1   F=-6.36988e+02  cnf=[A B A*E D D E ]
    Fr      Fe      Fs      Ft      Fc      Fb      Fh      Fm      Fg      Fp
   47.10 -197.14    0.00   -2.92    0.00    0.00  -26.54 -398.73   13.54  -72.29
LEU      -68.76 116.28 172.49-171.42 175.13  51.44  81.49
ASN      -52.76 -51.31-177.56 -87.69-143.06-177.47
ASP  A   -52.73 -41.06-169.13 -76.98  -9.58
ASN  B  -102.07  28.58 166.23 -73.76 -92.79-179.61
GLY  A*   67.27  33.90 167.75
VAL  E  -119.44 105.10 177.83-173.22  54.85  71.84
ASP  D  -156.34 108.41 175.35-176.33  -5.60
GLY  D  -155.37-101.60-179.55
GLU  E   -98.36 124.68-178.49-118.79-177.96 -61.38
VAL     -120.53 120.67-176.38  53.32  60.06  61.48
GLUe     -99.80 114.75        -67.91-147.38 -97.69
 iD0=   2   F=-6.36037e+02  cnf=[A A G*E E D E ]
    Fr      Fe      Fs      Ft      Fc      Fb      Fh      Fm      Fg      Fp
   43.10 -181.16    0.00    0.59    0.00    0.00  -26.31 -404.08    5.78  -73.95
LEU      -68.76 116.28 172.49-173.04 176.33  49.82  81.73
ASN      -52.76 -51.31-175.92 173.80  88.51 178.15
ASP  A   -54.45 -51.92-171.79 -76.86  -7.16
ASN  A   -66.86 -43.24 178.26 -84.31 -92.76 176.88
GLY  G*  130.56  37.78 173.79
VAL  E  -120.47  94.10 174.43-169.55  55.85  69.30
ASP  E  -133.48 121.26 177.54 -78.44 -20.03
GLY  D  -170.67-110.17-177.73
GLU  E   -98.13 124.74-178.49-118.80-177.97 -61.38
VAL     -120.53 120.67-176.38  54.45  59.85  63.59
GLUe     -99.80 114.75        -70.24-150.56-100.01
```

```
  .
  .
  .
 iD0=  13   F=-6.16447e+02  cnf=[A A D D*F*E*E ]
    Fr      Fe      Fs      Ft      Fc      Fb      Fh      Fm      Fg      Fp
   52.20 -181.82    0.00   -4.12    0.00    0.00  -12.02 -402.09    6.54  -75.13
LEU      -68.76 116.28 172.49 170.30  57.20  53.89  51.58
ASN      -52.76 -51.31-175.78 -82.49 -18.73 179.39
ASP  A   -51.07 -44.86-176.95 174.53  60.56
ASN  A   -71.06 -50.25-175.68 -81.82 -95.90 178.91
GLY  D  -126.52-108.49 175.42
VAL  D*  178.71 128.34 155.31  98.83  64.31  59.29
ASP  F*   59.98-146.95 178.50-157.71  27.66
GLY  E*  116.95-165.81 168.66
GLU  E   -93.76 124.64-178.49-155.89 176.57-164.05
VAL     -120.53 120.67-176.38  49.72  62.08  58.55
GLUe     -99.80 114.75        -96.63 174.52 124.01
```

For the final collection of low-energy local minima, conformations are characterized by energy, by sequence of conformational regions, and by torsion angle values for the set of residues specified in file "test/stp/stp.lp0.1pga" to contain torsion angles allowed to vary in the search.

The conformation having the 2nd lowest energy is the crystal structure conformation. The conformations having the 1st, 4th, and 10th lowest energies have the same backbone conformation as the crystal structure with differing side chain conformations. The conformation having the 10th lowest energy results from side chain placement onto the undeformed backbone.

```
    dG
  669.09
    Fr      Fe      Fs      Ft      Fc      Fb      Fh      Fm      Fg      Fp
   47.03 -173.08    0.00   -3.83    0.00    0.00  -29.64 -422.55    6.64  -72.54
  Gunf    Sunf
 -911.32 2360.89
```

dG=(G$-$Gu) is an estimate of free energy of folding that is not meaningful for a single sequence.

Gunf and Sunf are the free energy and entropy, respectively, of an Ising model representation of the unfolded state.

In a search through sequence space, ddG=(dG(sequence1) $-$dG(sequence0)) enables meaningful comparison of stability between different sequences.

dG is calculated as the free energy of the folded state c_{pp}(Fr+Fe+Fs+Ft+Fb+Fc+Fg+Fp) $+c_h$Fh $+c_{ps}$Fps $+c_{ss}$Fss) minus the free energy of the unfolded state (Gunf), where { c_{pp}, c_h, c_{ps}, c_{ss}} are coefficients optimized such that calculated ddG matches observed ddG over a large dataset of experimental measurements.

```
COMPACT ANALYSIS
 energy minima in sequence-structure space
 iM1    Fr      Fe      Fs      Ft      Fb      Fh      Fm      Fg      Fp      Fc
   0   47.03 -173.08    0.00   -3.83    0.00  -29.64 -422.55    6.64  -72.54    0.00
 iM1 G -Gu      Gu      Qsol
   0  669.09 -911.32   -0.29
 iM1  sequence substitutions
```

```
      12  35  36  37  38  39  40  41  42  54  56
   0  LEU ASN ASP ASN GLY VAL ASP GLY GLU VAL GLUe
```

The collection of most stable sequences found in order of stability.

As a second example, input file "test/stp/stp.c2.1pga" directs a search through sequence space. At 3 residue positions { 5 LEU,34 ALA,43 TRP}, each of which contributes to the hydrophobic core of 1pga, a small set of alternative side chains is substituted.

To execute the command, open a window to directory "src/str", and type the following line to the macOS (or linux) prompt.

```
%  estp test 1pga 05 c2
```

File "test/stp/stp.c2.1pga" specifies 10 positions that contain torsion angles variable in the search. Backbone is variable but not searched for residue positions { 5, 7,34,43,54}. Because backbone variability includes the omega torsion of the previous residue, in addition to ϕ and ψ of the residue marked as variable, residue positions { 4, 6,33,42,53} must also be included as residues containing 1 or more torsions that will vary with conformational search. The side chain is searched for residue positions { 5, 7,34,43,54}.

Automated expansion of the search subspace, to include all side chains that can contact a side chain initially marked as searchable, adds 8 side chains {12,16,30,31,33,39,40,52} to the searchable collection. To avoid automated expansion of the search subspace, a name must be chosen for the subset of degrees of freedom such that the 1st character is either ’l’ or ’h’.

The collection of most stable sequences found in the search is listed, in order of stability, at the end of the diagnostic output file "test/dgn/estp.1pga.05.c2". For the most stable sequence found in the search through sequence space, the corresponding lowest-energy conformation found in the search through conformation space is output, in pdb format, to file "test/car/MOL.v000.pdb".

In this example, the native sequence is found to be the most stable of the collection examined.

Using an Apple M1 Pro processor, this small search through sequence space, in which the sequences of 1pga and 63 variants are examined, requires computation time of about 8.7 hours.

In the following abbreviated listing, frame boxes enclose descriptions of the output data.

file="test/dgn/estp.1pga.05.c2"

Diagnostic file created by the command:
% estp test 1pga 05 c2

Input file "test/stp/stp.c2.1pga" directs search through 64 sequences. At 3 positions that contribute to the hydrophobic core { 5 LEU, 34 ALA, 43 TRP}, alternative hydrophobic side chains are substituted.

.
.
.

For each sequence, the program outputs a description of the search through conformational space. The lowest-energy conformation found is used to evaluate dG. Relative stability is calculated as ddG, the change in dG with sequence.

```
COMPACT ANALYSIS
 energy minima in sequence-structure space
 iM1    Fr      Fe      Fs      Ft      Fb      Fh      Fm      Fg      Fp      Fc
   0   24.48   14.38    0.00    2.40    0.00 -103.21 -418.14    7.39  -71.34    0.01
   1   50.26   -0.20    0.00    6.63    0.00 -116.19 -406.28    7.39  -72.43    0.01
   2   43.37    5.36    0.00    4.60    0.00 -105.63 -407.10    7.39  -71.68    0.01
   3   47.73   18.67    0.00    4.31    0.00 -113.01 -418.07    7.39  -71.59    0.01
  .
  .
  .
  63   70.03   35.48    0.00    3.81    0.00  -93.25 -423.39    7.39  -72.63    0.02
```

Energy components for sequences ordered by stability.

In rare instances, evaluation of Fm fails due to occurrence of an error in calculation of the molecular surface. In such cases, Fm is set to zero.

```
 iM1 G -Gu      Gu      Qsol
   0  696.97 -911.32   -0.29
   1  698.47 -911.39   -0.29
   2  699.41 -908.52   -0.30
   3  699.51 -908.99   -0.31
  .
  .
  .
  63  717.61 -902.08   -0.41
```

dG=(G$-$Gu) is an estimate of free energy of folding that is not meaningful for a single sequence.

Gu is the free energy of an Ising model representation of the unfolded state.

In a search through sequence space, ddG=(dG(sequence1) $-$dG(sequence0)) enables meaningful comparison of stability between different sequences.

dG is calculated as the free energy of the folded state c_{pp}(Fr+Fe+Fs+Ft+Fb+Fc+Fg+Fp) $+c_h$Fh $+c_{ps}$Fps $+c_{ss}$Fss) minus the free energy of the unfolded state (Gu), where { c_{pp}, c_h, c_{ps}, c_{ss}} are coefficients optimized such that calculated ddG matches observed ddG over a large dataset of experimental measurements.

```
 iM1  sequence substitutions
        4   5   6   7  12  16  30  31  33  34  39  40  42  43  52  53  54
   0  LYS LEU ILE LEU LEU THR PHE LYS TYR ALA VAL ASP GLU TRP PHE THR VAL
   1  LYS LEU ILE LEU LEU THR PHE LYS TYR VAL VAL ASP GLU TRP PHE THR VAL
   2  LYS LEU ILE LEU LEU THR PHE LYS TYR VAL VAL ASP GLU PHE PHE THR VAL
   3  LYS LEU ILE LEU LEU THR PHE LYS TYR ILE VAL ASP GLU PHE PHE THR VAL
  .
  .
  .
  63  LYS ILE ILE LEU LEU THR PHE LYS TYR MET VAL ASP GLU LEU PHE THR VAL
```

The collection of most stable sequences found in order of stability.

Input file "test/stp/stp.c2.1pga" specifies 10 positions that contain torsion angles variable in the search.

Backbone is variable but not searched for residue positions { 5, 7,34,43,54}. Because backbone variability includes the ω torsion of the previous residue in addition to the ϕ and ψ torsions of the residue specified as variable, residue positions { 4, 6,33,42,53} must also be included as residues containing 1 or more torsions that will vary with conformational search.

The side chain is searched for residue positions { 5, 7,34,43,54}. Automated expansion of the search subspace, to include side chains that can contact the side chains specified as searchable, adds 8 side chains { LEU12, THR16, PHE30, LYS31, TYR33, VAL39, ASP40, PHE52}.

To avoid automated expansion of the search subspace, choose a name for the subset of degrees of freedom such that the 1st character is ’l’ or ’h’.

The native sequence has LEU, ALA, and TRP at positions 5, 34, and 43, respectively. For the collection of sequences evaluated, the native sequence is predicted to be the most stable.

6.4 prof

FUNCTIONALITY
structure quality assessment

SYNTAX
prof FAM MOL CNF

INPUT FILES	OUTPUT FILES
/FAM/seq/seq.MOL	/FAM/dgn/prof.MOL.CNF
/FAM/tor/tor.MOL.CNF	/FAM/car/dprof.MOL.CNF
/FAM/car/MOL.CNF.pdb	

SUMMARY

1) inputs a sequence, file "FAM/seq/seq.MOL"

2) inputs a conformation in 2 alternative formats: torsion angle format in file "FAM/tor/tor.MOL.CNF", and pdb format in file "FAM/car/MOL.CNF.pdb"

3) identifies defects in a structure, associates with these defects differential free energies of folding, and accumulates a profile along the sequence of defect energy density

4) outputs the defect energy density profile in file "FAM/car/dprof.MOL.CNF"

5) outputs, in file "FAM/dgn/prof.MOL.CNF", additional characterization of identified defects

NOTES

Structure quality assessment identifies chain segments likely to be improved by applications of molecular mechanics-based structure prediction.

A common use of the "prof" command is in preparation for the "hlog" command, which uses defect energy densities of template structures in deciding the best alignment of a target sequence to a group of structurally aligned templates. For each template, the "hlog" command inputs, along with the geometry regularized structure, the defect energy density profile. Creation of these input files requires execution of the "prof" command following the "greg" command for each template structure.

The "prof" command does not access the energy surface and requires only a few seconds of computation time.

EXAMPLE TEST CASE

To execute the command, open a window to directory "src/str", and type the following line to the macOS (or linux) prompt.

```
%  prof test 1pga 05
```

Files needed as input for this command were created previously in the example test case for the "ereg" command.

Because the PDB database entry 1pga (resolution=2.07 Å) should be free of significant defects, profile "test/car/dprof.1pga.05" provides an example of a baseline level of defect energy density that can be expected for a good structure.

---

file="test/dgn/prof.1pga.05"

Diagnostic file created by the command:
% prof test 1pga 05

```
OVERLAP
  .
  .
  .
DISALLOWED (PHI,PSI,OMG,CHI)
  .
  .
  .
EXPOSED HYDROPHOBIC SURFACE
  .
  .
  .
CAVITY
  .
  .
  .
BURIED CHARGE
  .
  .
  .
UNPAIRED HBOND DONORS AND ACCEPTORS
  .
  .
  .
OUTLIER STATISTICAL CONTACT
  .
  .
  .
ELECTROSTATIC
```

.
.
.

Output consists of a listing of defects. For each defect, a mapping is included to residues of the polymer chains.

The "prof" command assigns energies to defects. By partitioning these energies onto the contributing residues, the command accumulates a defect energy density along the chains.

The calculated energy density is output as a compact profile in file "test/car/dprof.PRO.CNF".

6.5 rcyc

FUNCTIONALITY
energy refinement of a homology model

SYNTAX
rcyc FAM MOL CNF

INPUT FILES	OUTPUT FILES
/FAM/seq/seq.MOL	/FAM/dgn/rcyc.MOL
/FAM/tor/tor.MOL.CNF	/FAM/stp/stp.g???.MOL
	/FAM/car/MOL.g???.pdb
	/FAM/tor/tor.MOL.g???

SUMMARY

1) inputs a sequence, "FAM/seq/seq.MOL"

2) inputs an initial conformation in torsion angle format, "FAM/tor/tor.MOL.CNF"

3) generates a collection of 7-residue segments chosen such that the union spans the entire chain (or chains) and, using notation SUB=g??? where ? is a digit, outputs to files "FAM/stp/stp.g???.MOL" the corresponding subsets of degrees of freedom to be searched

4) cycles through this collection of segments, performing a limited conformational search with respect to each subset of degrees of freedom

5) outputs a sequence of conformations in 2 alternative formats, files "FAM/car/MOL.g???.pdb" and "FAM/tor/tor.MOL.g???", one for each subset of degrees of freedom, where CNF=g??? is the lowest energy conformation obtained in the search with respect to SUB=g???

6) outputs, in file "FAM/dgn/rcyc.MOL", a diagnostic summary of the execution of the command

NOTES

The "rcyc" command accesses the energy surface. Using an Apple M1 Pro processor, refinement search for the 56 residue 1pga requires about 3.5 hours of computation time.

EXAMPLE TEST CASE

In this example, following the local energy minimization of the experimental structure, the "rcyc" command is used to search the space of conformations in the neighborhood of the experimental structure, thereby establishing a more meaningful energy for the native state.

Files "test/seq/seq.1pga" and "test/tor/tor.1pga.05" were created by the previous execution of the test case for the "ereg" command.

To execute the commands, open a window to directory "src/str", and type the following lines to the macOS (or linux) prompt.

```
%  rcyc test 1pga 05
%  cp ../../test/car/1pga.g013.pdb ../../test/exp/1pga_r.pdb
%  ereg test 1pga_r
```

The "rcyc" command first creates 14 subsets of torsion angle degrees of freedom, SUB={ g000, g001, . . . , g013}, then cycles through these subsets, accomplishing a limited conformational search with respect to each subset.

Following this energy refinement of the experimental structure, the final conformation (MOL="1pga", CNF="g013") is renamed using the more descriptive MOL="1pga_r". The final command, directing local energy minimization with respect to the full set of torsion angle degrees of freedom, establishes an energy for the region of the experimental structure. We note that, because the subspace searches exclude energy terms that do not change with respect to the subset of degrees of freedom, the partial energies associated with a subset of degrees of freedom are not comparable between subsets, or to a full energy associated with the full set of torsion angle degrees of freedom. The energy of the refined structure is more relevant than the energy of the experimental structure as a point of reference in searches of the conformational space for regions having lower energy than the native conformation. Existence of such regions would indicate errors remaining in the energy functions.

From file "test/dgn/ereg.1pga", the total energy of the energy minimized crystal structure, "test/car/1pga.05.pdb", is -1630.58 kcal/mol. From file "test/dgn/ereg.1pga_r", the total energy of the energy refined native conformation, "test/car/1pga_r.05.pdb", is -1663.22 kcal/mol.

6.6 hlog

FUNCTIONALITY
homology model building

SYNTAX
hlog FAM MOL GRP

INPUT FILES	OUTPUT FILES
/FAM/arg/exptstructs.GRP	/FAM/dgn/hlog.MOL.GRP
/FAM/seq/seq.MOL	/FAM/car/TEM.GRP.pdb
/FAM/seq/seq.TEM	/FAM/tor/tor.MOL.GRP
/FAM/tor/tor.TEM.00	/FAM/car/MOL.GRP.pdb
/FAM/car/TEM.00.pdb	/FAM/car/dprof.MOL.GRP
/FAM/car/dprof.TEM.00	/FAM/stp/stp.hot.MOL

SUMMARY

1) inputs a target sequence, file "FAM/seq/seq.MOL"

2) inputs a group of templates, file "FAM/arg/exptstructs.GRP", to be used in building the homology model

3) for each template in group GRP; inputs a sequence, file "FAM/seq/seq.TEM"; 2 alternative specifications of the geometry regularized structure, files "FAM/car/TEM.00.pdb" and "FAM/tor/tor.TEM.00"; and the defect energy density profile of the geometry regularized structure, file "FAM/car/dprof.TEM.00"

4) aligns all template structures contained in group GRP and partitions this multiple structure alignment into structurally conserved and non-conserved regions

5) aligns the target sequence to the multiple structure alignment, selecting, within each region, the template that best matches the target

6) constructs a homology model by transferring coordinates for atoms aligned to a template structure, and by generating coordinates for unaligned residues and for substituted side chains

7) for each template in group GRP, outputs, in pdb format, file "FAM/car/TEM.GRP.pdb", the geometry regularized structure of the template applying the translation and rotation obtained from multiple structural alignment of the group of templates

8) outputs 2 alternative specifications of the homology model structure, files "FAM/car/MOL.GRP.pdb" and "FAM/tor/tor.MOL.GRP", and the corresponding defect energy density profile, file "FAM/car/dprof.MOL.GRP"

9) outputs, in file "FAM/stp/stp.hot.MOL", specification of a subset of degrees of freedom, in the format required for input to the "estp" command, chosen such that conformational search with respect to this subset is likely to reduce defects in the model structure

10) outputs, in file "FAM/dgn/hlog.MOL.GRP", a diagnostic summary of the execution of the command

NOTES

Homology model building leads to one of the major applications of molecular mechanics-based structure prediction, structure prediction of surface loops for which knowledge-based structure prediction may not be reliable.

At the expense of added preparation, geometry regulation of templates and subsequent defect energy density calculation have been separated from the "hlog" command. Effective use of the command often requires some experimentation and feedback to select a group of templates that align well structurally. By separating geometry regulation and defect energy density calculation, these computations can be performed only once for an entire family of templates.

The "hlog" command does not access the energy surface and requires only a few minutes of computation time.

EXAMPLE TEST CASE

In this example, a homology model is constructed for the sh3 domain of the human gene sequence grb2 (growth factor bound protein 2). Although the set of experimental structures for this family of gene sequences exceeds 60, we limit here the group of templates to 6 structures: SCOP domains 1ckaA_, 1bbzA_, 1cskA_, 1semA_, 1oebA_, and 1jo8A_. We assign to this group the name "sparse". For comparison, an NMR structure exists for this target sequence: pdb entry 1gbr.

File "test/arg/exptstructs.sparse", a listing of the templates for GRP="sparse", was created by hand editing. Files "test/exp/TEM.pdb", where TEM is the 6-character domain name, were created by copying and editing the corresponding 4-character PDB entry to extract the domain homologous to the target sequence. File "test/seq/seq.grb2" was created, starting from the SCOP definition of the domain which includes a 1-letter code specification of the amino acid sequence, using the utility command "seqformat". In preparation for the "seqformat" command, input file "test/exp/seq.grb2", the target residue sequence in 1-letter code format, was created by hand editing.

To execute the commands, open a window to directory "src/str", and type the following lines to the macOS (or linux) prompt.

```
%  seqformat test grb2
%  greg test 1ckaA_
%  greg test 1bbzA_
%  greg test 1cskA_
%  greg test 1semA_
%  greg test 1oebA_
%  greg test 1jo8A_
%  prof test 1ckaA_ 00
%  prof test 1bbzA_ 00
%  prof test 1cskA_ 00
%  prof test 1semA_ 00
%  prof test 1oebA_ 00
%  prof test 1jo8A_ 00
%  hlog test grb2 sparse
```

For each template TEM, input files "test/seq/seq.TEM", "test/tor/tor.TEM.00", "test/car/TEM.00.pdb", and "test/car/dprof.TEM.00" are created by executing the "greg" command followed by the "prof" command.

Graphical viewing of the homology model, file "test/car/grb2.sparse.pdb", shows a conformation with many stabilizing hydrophobic and H-bond interactions, but also with some destabilizing atom pair overlaps localized to segment 40–45, and with some questionable conformations of charged side chains. Supplementing the user’s eye, the defect energy density profile, file "test/car/dprof.grb2.sparse", identifies a peak over residues 40-45 originating from overlaps, disallowed (ϕ, ψ) values, and unpaired Hydrogen bond donors and acceptors.

The subset of torsion angle degrees of freedom SUB="hot", specified in file "test/stp/stp.hot.grb2", was selected by the program as that most likely to reduce defect energy. Alternatively, a "stp.SUB.MOL" file can be created by hand and conformational search with respect to this subset directed by the "estp" command.

As a test of proper execution, each output file can be compared to the corresponding file in the subdirectory "TESTCASES".

The diagnostic file contains the multiple structure alignment and the alignment of the target sequence to the multiple structure alignment. In the following abbreviated listing, frame boxes enclose descriptions of the output data.

file="test/dgn/hlog.grb2.sparse"

Diagnostic file created by the command:
% hlog test grb2 sparse

```
INPUT TEMPLATE STRUCTURES
```

```
ALIGN TEMPLATE STRUCTURES
structure-structure alignment of 1ckaA_ to 1oebA_
C-alpha RMSD of aligned substructures=   1.35
number of residues aligned=  44
GLU A 135    ARG A   2
TYR A 136    TRP A   3
VAL A 137    ALA A   4
  .
  .
  .
VAL A 187    VAL A  53
GLU A 188    ALA A  54
structure-structure alignment of 1bbzA_ to 1oebA_
C-alpha RMSD of aligned substructures=   1.25
number of residues aligned=  50
  .
  .
  .
structure-structure alignment of 1cskA_ to 1oebA_
C-alpha RMSD of aligned substructures=   1.63
number of residues aligned=  45
  .
  .
  .
structure-structure alignment of 1semA_ to 1oebA_
C-alpha RMSD of aligned substructures=   1.01
number of residues aligned=  53
  .
  .
  .
structure-structure alignment of 1jo8A_ to 1oebA_
C-alpha RMSD of aligned substructures=   0.93
number of residues aligned=  50
  .
  .
  .
```

A multiple structure alignment is generated for the templates of group GRP="sparse". The longest template is identified to be 1oebA_. For all remaining templates, the structure is aligned to the structure of the longest template. If one or a few templates fail to align over a large fraction of their length, then a better model will likely be obtained by reducing the size of the group of templates.

```
Number of Aligned Structures=  6
Number of Alignment Elements=  2
Alignment Element=  1 Length=  26
1ckaA_       ( 1:   2-  27) EYVRALFDFNGNDEEDLPFKKGDILR
1bbzA_       ( 1:   1-  26) NLFVALYDFVASGDNTLSITKGEKLR
1cskA_       ( 1:   2-  27) TECIAKYNFHGTAEQDLPFCKGDVLT
1semA_       ( 1:   3-  28) KFVQALFDFNPQESGELAFKRGDVIT
1oebA_       ( 1:   6-  31) RWARALYDFEALEEDELGFRSGEVVE
1jo8A_       ( 1:   1-  26) PWATAEYDYDAAEDNELTFVENDKII
Alignment Element=  2 Length=  13
1ckaA_       ( 1:  43-  55) EGKRGMIPVPYVE
1bbzA_       ( 1:  42-  54) KNGQGWVPSNYIT
1cskA_       ( 1:  44-  56) VGREGIIPANYVQ
```

```
1semA_       ( 1:  43-  55) NNRRGIFPSNYVC
1oebA_       ( 1:  46-  58) HNKLGLFPANYVA
1jo8A_       ( 1:  43-  55) DGSKGLFPSNYVS
```

Regions are identified for which structure is conserved over all templates. In the above listing, structurally conserved regions are referred to as alignment elements.

```
ALIGN TARGET SEQUENCE TO TEMPLATE STRUCTURES
Alignment of Target Sequence to Family of Templates
number of templates=  6
target sequence=grb2
%identical= 44.64  template sequence=1cskA_
 -MEAIAKYDF KATADDELSF KRGDILKVLN EECDQNWYKA E-LNGKDGFI PKNYIEMK
 GTECIAKYNF HGTAEQDLPF CKGDVLTIVA VTKDPNWYKA KNKVGREGII PANYVQKR
  ######### ########## #######                  ####### ######
  .
  .
  .
target sequence=grb2
%identical= 33.93  template sequence=1bbzA_
 MEAIAKYDFK ATADDELSFK RGDILKVLNE ECDQNWYKAE LNGKDGFIPK NYIEMK--
 NLFVALYDFV ASGDNTLSIT KGEKLRVLGY NHNGEWCEAQ TKNGQGWVPS NYITPVNS
 ########## ########## ######                 ######### ####
```

The target sequence grb2 is aligned to each template of the group. A pound character below the alignment indicates the above template residue is contained within a structurally conserved region.

```
Structurally Conserved Element=  0, Length=  26
grb2                          ( 1:   1-  26) MEAIAKYDFKATADDELSFKRGDILK
1cskA_                        ( 1:   2-  27) TECIAKYNFHGTAEQDLPFCKGDVLT
Preceding Structurally Conserved Element=  1, Length=  15
grb2         ( 1:  27-  41) VLNEECDQNWYKAEL
1bbzA_       ( 1:  27-  41) VLGYNHNGEWCEAQT
Structurally Conserved Element=  1, Length=  13
grb2                          ( 1:  42-  54) NGKDGFIPKNYIE
1cskA_                        ( 1:  44-  56) VGREGIIPANYVQ
Following Structurally Conserved Element=  1, Length=   2
grb2                          ( 1:  55-  56) MK
1cskA_                        ( 1:  57-  58) KR
```

For each region, structurally conserved and intervening, a template is selected to optimize alignment with the target sequence.

```
Segments Targeted for Backbone Structure Generation
```

For each target residue aligned to a template, backbone coordinates are copied from the aligned template residue. If insertions or deletions occur, constraints of chain connectivity and polypeptide geometry require backbone structure generation, including some searching of conformations, for segments of the chain surrounding an insertion or deletion. This section of output specifies those segments for which backbone structure has been generated as opposed to transferred from a template.

```
BUILD TARGET HOMOLOGY MODEL
  .
  .
  .
EVALUATE DEFECT PROFILE
  .
  .
  .
```

A listing of defects for the model structure.

As a second example using the "rcyc" command, we perform 1 cycle of energy refinement to the homology model characterized by MOL="grb2" and CNF="sparse". The following command executes the search.

```
%  rcyc test grb2 sparse
%  cp ../../test/car/grb2.g013.pdb ../../test/car/grb2.cyc1.pdb
%  cp ../../test/tor/tor.grb2.g013 ../../test/tor/tor.grb2.cyc1
```

The copy commands assign a more meaningful name to the final energy refined conformation.

Visual comparison of the energy refined conformation, "test/car/grb2.cyc1.pdb", with the initial homology derived conformation, "test/car/grb2.sparse.pdb", shows that the quality is much improved by the refinement. Although changes to the structure are localized, the packing of the hydrophobic core, the distribution of ionized groups on the protein surface, secondary structure elements, and the number of Hydrogen bonds all move noticeably toward patterns seen in native protein structures.

As another measure of the quality of the homology model, we can compare the energies of the homology model and the NMR structure. File "test/exp/1grbA1.pdb" was created by copying and hand editing the PDB database entry 1grb, to retain only the first model, and to remove all residues outside of the SCOP domain definition for the SH3 domain.

```
%  ereg test 1gbrA1
%  cp ../../test/car/grb2.g013.pdb ../../test/exp/grb2_c.pdb
%  ereg test grb2_c
```

The energy of the homology model is lower than the energy of the NMR structure, -1483.46 and -1420.24 kcal/mol, respectively.

As an alternative to the "rcyc" command,

```
%%  estp test grb2 sparse hot
```

uses energy-based structure prediction to patch the initial homology model. The altered prompt is meant to indicate that the above command is not a part of this test case.

6.7 igor

FUNCTIONALITY
ab initio fold prediction

SYNTAX
igor FAM MOL

INPUT FILES	OUTPUT FILES
/FAM/exp/seq.MOL	/FAM/dgn/igor.MOL
	/FAM/dgn/igodmp.MOL

/FAM/seq/seq.MOL_??i
/FAM/tor/tor.MOL_??i.igo
/FAM/car/MOL_??i.igo.pdb

SUMMARY

1) inputs a target sequence in 1-letter code format, file "FAM/exp/seq.MOL"

2) generates 16 ab initio folds by global search through the space of element compositions of the low-resolution igor model

3) forks the composition MOL into 16 identical compositions by appending "_??i" to MOL, where ?? denotes 2 digits specifying the order of the fold based on the igor model score

4) for each of the igor model folds, outputs a specification of residue sequence and 2 alternative specifications of structure in files "FAM/seq/seq.MOL_??i", "FAM/tor/tor.MOL_??i.igo", and "FAM/car/MOL_??i.igo.pdb"

5) outputs, in file "FAM/dgn/igor.MOL", a diagnostic summary of the execution of the command

6) outputs, in file "FAM/dgn/igodmp.MOL", a listing, for the top scoring folds, of the corresponding igor model scores and most probable chain states

NOTES

For each residue of a single polypeptide chain, the continuous space of conformations is replaced by 3 states {H,E,C}, abbreviations for the elements of secondary structure {helix, extended, coil} to which the residue can contribute. The set of chain states is the set of mappings from the set of residues into {H,E,C}. Letting n be the number of residues, the number of chain states is 3^n.

The set of element compositions is the set of mappings from the set of pairs of consecutive residues into a binary decision: either extension of the current string or transition between strings. The number of element compositions is $2^{(n-1)}$. For each chain state and pair of consecutive residues, association of states {HH,EE,CC} with extension of the current string, and states {HE,HC,EH,EC,CH,CE} with transition between strings, defines a mapping from the set of chain states to the set of element compositions.

For a single polypeptide chain, the igor model consists of the replacement of the continuous space of conformations with a discrete set of chain states, and a score function defined over the set of element compositions.

The igor model is a generalization of a residue Ising model. The residue Ising model is short range in the sense that residue i interacts directly only with neighboring residues in the range $\{i{-}5, \ldots, i{+}5\}$. The igor model extends the residue Ising model to include residue-residue interactions of intermediate and long range. Unlike the residue Ising model, for which integration over chain states has a fast analytic solution, the igor model requires use of a trajectory search over the space of element compositions to accomplish integration over chain states. For each element composition, integration over the subset of chain states consistent with the element composition has a fast analytic solution. The trajectory search through the space of element compositions enables isolation of the collection of element compositions that contribute the largest statistical weights to the integration.

The igor model score, defined over the set of $2^{(n-1)}$ element compositions, is formed as a sum of 7 components: SCORE=(SCO0 +SCO1 +SCO2 +SCO2b +SCO3 +SCO3b +SCO3c). The partial sums (SCO1+SCO2+SCO2b) and (SCO3+SCO3b+SCO3c) represent, respectively, residue-residue interactions of short+intermediate range and intermediate+long range.

Score components SCO1, SCO2, and SCO2b are evaluated using 3 distinct Ising models. In model 0, a residue Ising model is loaded with residue impulses. Residue i interacts with neighbor residues in residue range $[i-5, i+5]$. In model 2, an element Ising model is loaded with element impulses. Element i interacts with neighbor elements in element range $[i-5, i+5]$. In model 1, an element Ising model is loaded with element impulses and residue impulses. Element i interacts with neighbor elements in element range $[i-5, i+5]$, and residue i interacts with neighbor residues in residue range $[i-5, i+5]$.

Let gj and sj be the easily evaluated free energy and entropy, respectively, of Ising model j. Because the residue impulses loaded into model 0 depend only on residue sequence, g0 and s0 are constant over the space of element compositions. Because the element impulses loaded into models 2 and 1 vary with element composition, g2, s2, g1, and s1 are functions defined over the space of element compositions. The above functions over the space of element compositions are combined as ((g1$-$s1) $-$(g2$-$s2) $-$(g0$-$s0))= (SCO1+SCO2+SCO2b) to create a physically meaningful score for interactions of short and intermediate range. Igor model score components SCO1, SCO2, and SCO2b are defined as ((g1$-$s1) $-$(g0$-$s0)), -g2, and s2, respectively.

SCO0 is the natural logarithm of an estimate for the probability of occurrence of the element composition, independent of any calculated probability distribution over chain states.

A fold is a full-atom structure model generated for a sequence of helices, extended strands, and connecting coil segments by maximizing the (SCO3+SCO3b+SCO3c) component over the space of packed configurations. The search through packing space, a part of the igor model search through the space of element compositions, is accomplished as a sequence of 2 conceptually similar searches: the packing of strands into sheet configurations, followed by the packing of helices and sheet configurations into folds. The packing algorithm produces both the (SCO3+SCO3b+SCO3c) component, which enables more meaningful scoring of the element composition, and a full-atom structure model, which can be used as a starting point for global energy minimization using the "rcyc", "estp" or "ptra" commands.

A more complete description of the igor model is available in a publication on the company website.

Because the igor model score and components are expressed in units of $-kT$, where k is Boltzmann's constant and T is absolute temperature, the highest igor model score is the most probable fold. Igor model scores and most probable chain states are listed in output file "FAM/dgn/igodmp.MOL".

The "igor" command does not access the energy surface but can, for long target chains, be computationally intensive. Using an Apple M1 Pro processor, fold generation for the 56 residue protein 1pga requires computation time of about 69 minutes.

EXAMPLE TEST CASE

File "seq.1pga", the residue sequence in 1-letter code format, has been entered into directory "test/exp".

To execute the command, open a window to directory "src/str", and type the following line to the macOS (or linux) prompt.

```
%  igor test 1pga
```

In the following abbreviated listing, frame boxes enclose descriptions of the output data.

file="test/dgn/igodmp.1pga"

Diagnostic file created by the command:
% igor test 1pga

```
 iK7   SCORE    SCO0     SCO1     SCO1b    SCO2     SCO2b    SCO3     SCO3b    SCO3c
   0   83.025    1.612   -1.250    0.000   -2.155    3.569   52.711   20.067    8.470
  10
epeuidedet
EEEEEEEECCCEEEEEEECCCCHHHHHHHHHHHHHHHCCCCEEEEECCCEEEEEEC
   1   81.592    2.815   -0.071    0.000   -3.875    4.239   51.240   19.388    7.856
  10
euqqidepet
EEEEEEECCEEEEEEECCCCCCHHHHHHHHHHHHHHHCCCCEEEEECCCCEEEEEC
   2   81.259    2.875   -0.699    0.000   -2.360    2.834   49.628   20.100    8.880
  10
epeuiddpet
EEEEEEEECCCEEEEEEEEECCHHHHHHHHHHHHHHHCCCCEEEECCCCEEEEEEC
   3   79.501    2.661   -0.916    0.000   -1.278    2.349   45.995   21.909    8.781
  10
epeuiedpdu
EEEEEEECCCCEEEEEEECCCCHHHHHHHHHHHHHHHCCCCCCEEECCCCEEEECC
  .
  .
  .
```

A listing of high-scoring igor model element compositions is ordered by igor model score. Each element composition is characterized by the total score, a decomposition of the total score, and the most probable chain state. The most probable chain state is useful as a compact description of the corresponding model structure, output with CNF set equal to the order index for SCORE.

Igor model score components are defined as follows:

SCO0	ln(probability of element composition)
SCO1	((g1−s1) −(g0−s0))
SCO1b	not used
SCO2	−g2
SCO2b	s2
SCO3	side chain contacts in helix or sheet configuration formation
SCO3b	packing of helices and sheet configurations into folds
SCO3c	nonpolar side chains in electric field of ionized groups

Comparison of the model structures with the energy minimized crystal structure, "/test/car/1pga.05.pdb", shows that of the top 16 folds, 12 approximate the native conformation in secondary structure and topology. The stuctures with non-native topologies (04i, 08i, 14i, and 15i) all have a 3-stranded beta sheet. This result is also suggested by the listing in output file "test/dgn/igodmp.1pga". In user project 0, we use these models as starting points for global energy minimization in an attempt to either recover the native conformation or, by finding conformations with energies lower than the native energy, gain insight into deficiencies in the current energy function.

6.8 ptra

FUNCTIONALITY
guided trajectory search

SYNTAX
ptra FAM MOL CNF SUB

INPUT FILES	OUTPUT FILES
/FAM/seq/seq.MOL	/FAM/dgn/ptra.MOL.CNF.SUB
/FAM/tor/tor.MOL.CNF	/FAM/car/MOL.t???.pdb
/FAM/stp/stp.SUB.MOL	/FAM/tor/tor.MOL.t???

SUMMARY

1) inputs a sequence, "FAM/seq/seq.MOL"

2) inputs an initial conformation in torsion angle format, "FAM/tor/tor.MOL.CNF"

3) inputs a subset of degrees of freedom that will be allowed to vary, "FAM/stp/stp.SUB.MOL"

4) starting from the initial conformation, minimizes energy globally with respect to the subset of degrees of freedom

5) assigns CNF=t??? to the local minimum on the approximate energy surface generated by cycle ??? (where ? is a digit), and outputs 2 alternative specifications of conformation (in pdb and torsion angle formats)

6) outputs a diagnostic summary of the execution of the command, "FAM/dgn/ptra.MOL"

NOTES

Starting from a local minimum on the full energy surface, the fundamental unit of computation, referred to as a cycle, generates a neighboring local minimum on the full energy surface such that the targeted range for the RMSD from the starting local minimum is $[2, 4]$ Å. Global minimization is attempted as a sequence of cycles from which a trajectory of local minima is assembled by using a Monte Carlo criterion to decide acceptance or rejection for each new local minimum. The number of cycles directed by the "ptra" command is controlled by the parameter NCYCLES in file "src/str/thread_config". The default value of this parameter is 32.

Each cycle generates a trajectory on the simplified approximation to the full energy surface that was introduced in the description of the "ereg" command. Because of the complexity of component Fm, calculation of the full energy, and of 1st and 2nd derivatives, is too slow to support useful movement on the surface. The cycle trajectory, which is described more fully in a publication on the company website, attempts to minimize globally on the approximate energy surface, meaning the cycle trajectory attempts to pass over barriers into regions of the space of conformations containing local minima having lower energies. At the endpoint of each cycle, full energy is obtained as a single point evaluation, and a Monte Carlo acceptance decision is made based of full energy.

A cycle trajectory is a composite formed by linking trajectories generated by 3 distinct algorithms: the algorithm used in the "ereg" command to generate a local minimization trajectory, and 2 variants of this workhorse algorithm. The first variant generates an ascent trajectory that climbs barriers and passes through saddle regions. Directions of ascent are prioritized based on gradients along these directions of a subset of energy components that is targeted for reduction in the current cycle. In contrast to a physical trajectory, meaning a solution to the equations of motion for a constrained mechanical system, an ascent trajectory represents an attempt to sense by analysis of the local energy surface which directions of ascent will be most productive toward a goal of entering regions of conformation space associated with lower energy local minima. The second variant generates an inflation trajectory that deforms the approximate energy surface such that motion favorable to the Fe, Fh, and Fp components is facilitated. The following paragraphs present a more mathematical summary of the inflation trajectory.

Let p be atomic radii for a subset of atom types. And let x be coordinates for a subset of torsion angle degrees of freedom.

A trajectory through the combined space of parameters and torsion angles is generated by minimization of a target function $K(p) = (Ky(x(p)) + Kz(x(p)) + Kw(p))$ with respect to p. The torsion angles x also vary, but are dependent on p. The function $x(p)$ is determined by the constraint that $x(p)$ be a local minimum conformation on the approximate energy surface with respect to x. The approximate energy surface depends on p through the dependence of Fr(p, x), the repulsion +dispersion component, on atomic radii.

Components of the inflation target function have physical meaning as follows. Ky=(Fe +Fh +Fp). $-Kz$ is proportional to the displacement (meaning the root mean square) of x, the position in generalized coordinates, from the position at the start of the cycle. Kw drives inflation by decreasing in value as atomic radii expand. The combined effect of minimizing these target function components is expansion of atomic radii, pushed by decrease of $Kw(p)$, on a path through parameter space that is biased toward minimizing (Fe +Fh +Fp). The energy pumped into the mechanical system by component $Kw(p)$ is, in some ways, analogous to kinetic energy in the physical equations of motion. Both the inflation trajectory and the ascent trajectory represent an experiment in which energy is added to a mechanical system less randomly, and more strategically, in comparison to thermal energy in a physical trajectory. The inflation trajectory ends when either 1) the change in $Ky(x(p))$ becomes greater than some threshold value, or 2) the RMSD increase of atomic radii becomes greater than some threshold value.

At an inner level of nesting, the smallest repeatable unit of computation, referred to as a composite step of the cycle trajectory, is formed as 1) 128 steps of a local minimization trajectory, 2) 512 steps of an ascent trajectory, starting from the endpoint of the local minimization trajectory, and 3) 1 step of an inflation trajectory starting from the endpoint of the ascent trajectory. Each cycle trajectory consists of 1) an adaptable number of composite steps, 2) restoration of the original (physically meaningful) atomic radii, and 3) 128 steps of a local minimization trajectory. The number of composite steps adapts to maintain the RMSD between neighboring local minima produced in consecutive cycles in a targeted range of [2, 4] Å.

For each new cycle, the algorithm uses a different set of weighting coefficients for the components (Ky, Kz, and Kw) of the target function for the inflation trajectory. Because each cycle trajectory is calculated in a unique environment, a local minimum that is not accepted will not be repeated. For each ascent trajectory within a cycle, the subset of energy components that is targeted for reduction can be changed if the RMSD between conformations produced by consecutive composite steps falls below some threshold.

As a tool for search of conformation space, the "ptra" command complements the "estp" command. The "estp" command, by focusing computation on a spatially localized region of a larger structure, is more efficient when productive motions are concentrated within 1, or more, segments. The "ptra" command, by enabling full chain flexibility, facilitates unconstrained motions such as changes to the packing configuration of helices and sheets.

The "ptra" command accesses the energy surface. Using an Apple M1 Pro processor, a 32 cycle trajectory for the 56 residue protein 1pga requires computation time of about 92.4 hours.

EXAMPLE TEST CASE

We generate a trajectory of local minima, intended to accomplish global energy minimization, for one of the top scoring models produced in the test case for the "igor" command. As preparation for the "ptra" command, we first run 1 cycle of energy refinement using the "rcyc" command. The input files needed for the "rcyc" command have been created by the test case for the "igor" command. In preparation for the "ptra" command, input file "test/stp/stp.full.1pga" was created by hand editing.

In this example, file "stp.full.1pga" specifies that all torsion angles are variable. As a consequence, because no energy contributions are excluded in contraction of the mechanical system, energies can be meaningfully compared with energies evaluated using the "ereg" command. In contrast, for the collection of "stp.g???.1pga_02i" files that is generated by the "rcyc" command, each file specifies a different subset of the full set of torsion angles. In such cases, because different subsets of degrees of freedom exclude different energy contributions, energies can not be meaningfully compared between searches.

To execute the commands, open a window to directory "src/str", and type the following lines to the macOS (or linux) prompt.

```
%  rcyc test 1pga_02i igo
%  cp ../../test/car/1pga_02i.g013.pdb ../../test/car/1pga_02i.s000.pdb
%  cp ../../test/tor/tor.1pga_02i.g013 ../../test/tor/tor.1pga_02i.s000
%  cp ../../test/stp/stp.full.1pga ../../test/stp/stp.full.1pga_02i
%  ptra test 1pga_02i s000 full
```

---

file="test/dgn/ptra.1pga_02i.s000.full"

Diagnostic file created by the command:
% ptra test 1pga_02i s000 full

```
SINGLE POINT EVALUATION ON RESTBCHMGP
COMPACT ANALYSIS
W0    F        Fr       Fe       Fs       Ft       Fb       Fc       Fh       Fm       Fg       Fp    RMSD
-6 -1528.8   341.3 -1260.5     0.0      0.5      0.0      1.1    -70.4  -493.2     15.5    -62.0  1.42
```

Initial values of full energy and components.

```
INFLATE RESTBCHPW CYCLE
LOCAL MINIMIZATION TRAJECTORY
  .
  .
  .
      F       Fr        Fe        Fs        Ft        Fc        Fb        Fh        Fw        Fp
 -1019.25   333.66 -1180.50     0.00     1.61     0.00     0.00   -65.56   -43.47   -64.98
```

Energy and components for trajectory endpoint on approximate surface.

```
 heavy atom RMSD=   0.917
```

From start of local minimization trajectory.

```
GENTLE ASCENT TRAJECTORY
 iCYC=0[es  m ]
```

Energy components targeted for reduction in selection of ascent directions.

```
  .
  .
  .
      F       Fr       Fe       Fs       Ft       Fc       Fb       Fh       Fw       Fp
  -934.94   359.66 -1168.69     0.00    39.73     0.00     0.00   -62.54   -41.50   -61.60
```

Energy is raised by 85 kcal/mol. From file "test/dgn/ereg.1pga", the diagnostic output for the "ereg" command testcase, the mechanical system contains 322 torsion angle degrees of freedom. Here, the energy injected into the system is .26 kcal/mol per degree of freedom, less than .5 kT per degree of freedom.

```
 heavy atom RMSD=   1.352
```

From start of gentle ascent trajectory.

```
STEP IN PARAMETER SPACE
 change in RHO
 iM2     H00      H02      C08      C09      N11      O13
   0    0.037    0.001    0.007    0.001    0.009    0.007
```

The set of atom types for which atomic radii are allowed to vary in the current cycle.

Changes to atomic radii from original (physically meaningful) values to the current step of parameter inflation.

```
LOCAL MINIMIZATION TRAJECTORY
  .
  .
  .
      F       Fr       Fe       Fs       Ft       Fc       Fb       Fh       Fw       Fp
 -1024.33   339.07 -1187.62     0.00     0.20     0.00     0.00   -71.94   -42.42   -61.60
 heavy atom RMSD=   0.576
GENTLE ASCENT TRAJECTORY
 iCYC=0[es  m ]
  .
  .
  .
      F       Fr       Fe       Fs       Ft       Fc       Fb       Fh       Fw       Fp
  -929.60   348.19 -1163.37     0.00    47.09     0.00     0.00   -64.01   -44.52   -52.97
 heavy atom RMSD=   1.244
STEP IN PARAMETER SPACE
 change in RHO
 iM2     H00      H02      C08      C09      N11      O13
   1    0.142    0.127    0.103    0.098    0.091    0.095
LOCAL MINIMIZATION TRAJECTORY
  .
  .
  .
      F       Fr       Fe       Fs       Ft       Fc       Fb       Fh       Fw       Fp
  -968.81   343.04 -1163.46     0.00    17.36     0.00     0.00   -70.16   -45.19   -50.41
 heavy atom RMSD=   1.125
INFLATION TRAJECTORY
  .
  .
  .
```

At the completion of each inflation trajectory, a summary is output for the trajectory through parameter space. Because the target function changes with each inflation step, this segment of output is best ignored by application users.

```
       change in RHO
 H00        0.376
 H02        0.367
 C08        0.320
 C09        0.292
 N11        0.182
 O13        0.239
```

Final values for changes in atomic radii.

```
 heavy atom RMSD=   2.535
```

From the starting point of the inflation trajectory.

```
DEFLATE RESTBCHPW
  i        F          z          b       lam2         s2       delF
  0-9.69369e+02 5.5554e-01 2.0000e-02 7.5998e-01 2.0008e-02-9.69460e+02
  1-9.69406e+02 4.8727e-02 1.5000e-02 9.7935e-01 1.5010e-02-9.69452e+02
  2-9.69415e+02 7.5945e-02 1.1250e-02 8.2180e-01 1.1269e-02-9.69435e+02
  3-9.69416e+02 1.0196e-01 8.4375e-03 1.0002e+00 8.4388e-03-9.69431e+02
  .
  .
  .
 79-9.69427e+02 1.0297e-04 1.1869e-04 9.9845e-01 1.1880e-04-9.69427e+02
 80-9.69427e+02 9.6485e-05
      F       Fr       Fe       Fs       Ft       Fc       Fb       Fh       Fw       Fp
  -969.43   343.90 -1163.78     0.00    17.24     0.00     0.00   -70.13   -45.13   -51.53
```

Local minimization following restoration of atomic radii to physical values. The energy on the approximate surface has increased from -1019 to -969 kcal/mol.

```
SEGMENT GLOBAL SEARCH ON RESTBCHMGP
  .
  .
  .
```

A segment is chosen and searched using the algorithm of the "estp" command. The interspersing of a full chain trajectory with conformation searches by segment deformation is an attempt to catalyze some topological chain transitions for which large transition state energies might slow progression of a trajectory search.

```
SINGLE POINT EVALUATION ON RESTBCHMGP
COMPACT ANALYSIS
W0    F       Fr      Fe      Fs      Ft      Fb      Fc      Fh      Fm      Fg      Fp    RMSD
-4 -1537.2   330.4 -1228.4     0.0    -1.6     0.0     1.7   -67.6  -522.0    13.6   -61.7  0.46
```

The full energy has decreased from -1528 to -1537 kcal/mol.

```
 heavy atom RMSD=   1.757
```

From the previous accepted step. The RMSD following deflation and segment search indicates the global minimization cycle has passed through barriers to sample a neighboring local minima on the energy surface.

Based on a Monte Carlo acceptance criteria, this local minimum is added to the trajectory of local minima on the full energy surface. The algorithm proceeds to the next cycle using the last accepted conformation as the starting point.

```
 RMSD_inflate=   2.535 RMSD_deflate=   1.757 oM= 7
INFLATE RESTBCHPW CYCLE
  .
  .
  .
```

The global search continues for 32 cycles.

```
COMPACT ANALYSIS
     RMSD             Energy Decomposition                                         Acceptance
     Ri  oM  Rd     Ftot    Frsc      Fe      Ftb      Fh      Fm    Fgp  dtot fold   z    rn   A
  0 2.54  4 1.76 -1537.2   332.2 -1228.4    -1.6   -67.6  -522.0 -48.1  -8.5 0.54 0.000 0.000 2
  1 1.28  2 2.54 -1509.6   328.6 -1259.8    32.0   -75.0  -486.7 -48.8  27.6 0.54 0.000 0.685 1
  2 3.12  4 3.11 -1506.5   334.7 -1232.0    21.1   -66.7  -522.9 -40.9  30.7 0.54 0.000 0.416 1
  3 3.19  4 3.21 -1532.6   325.0 -1270.4    14.9   -65.3  -492.9 -44.0   4.6 0.54 0.171 0.435 1
  4 2.92  4 2.83 -1554.8   329.2 -1285.8    -3.0   -75.1  -470.4 -49.3 -17.6 0.51 0.000 0.000 2
  5 0.83  5 0.80 -1555.0   325.8 -1274.2    -1.6   -76.2  -478.7 -49.9  -0.2 0.52 0.000 0.000 2
  6 2.24  3 2.36 -1496.2   345.4 -1221.5    10.7   -75.5  -525.1 -30.1  58.8 0.52 0.000 0.781 1
  7 1.43  4 1.46 -1554.5   329.7 -1227.3    -1.7   -72.2  -538.8 -44.2   0.5 0.50 0.817 0.559 2
  8 1.89  3 1.92 -1497.1   342.0 -1215.0    25.0   -77.2  -524.0 -48.0  57.4 0.50 0.000 0.482 1
  9 2.83  5 2.81 -1517.3   330.3 -1260.0    19.2   -80.1  -490.3 -36.2  37.2 0.50 0.000 0.372 1
 10 2.59  5 2.59 -1533.9   333.6 -1259.9    12.3   -80.5  -491.9 -47.5  20.6 0.50 0.001 0.791 1
 11 2.97  6 3.07 -1543.0   333.1 -1301.1     3.8   -76.5  -450.6 -51.6  11.5 0.50 0.016 0.432 0
 12 1.66  5 1.66 -1583.8   332.9 -1243.8    -8.0   -83.7  -527.9 -53.3 -29.3 0.53 0.000 0.000 2
 13 2.24  2 2.33 -1556.5   332.1 -1254.5    -6.1   -92.3  -493.9 -41.8  27.3 0.53 0.000 0.787 1
 14 2.53  1 2.53 -1476.0   345.4 -1215.5    14.9   -93.7  -494.7 -32.3 107.8 0.53 0.000 0.152 0
 15 2.00  2 1.79 -1529.5   334.6 -1257.0    -1.9   -84.2  -484.7 -36.2  54.3 0.53 0.000 0.921 1
 16 2.45  4 2.45 -1539.2   337.4 -1257.1    -2.4   -90.9  -488.6 -37.5  44.6 0.53 0.000 0.418 1
 17 2.06  5 2.05 -1447.5   347.8 -1153.9    12.2   -82.1  -544.0 -27.5 136.3 0.53 0.000 0.786 0
 18 2.64  5 2.57 -1553.2   338.0 -1209.1     3.2   -87.1  -551.0 -47.2  30.6 0.53 0.000 0.222 1
 19 2.57  4 2.61 -1525.5   341.0 -1273.8    -5.2   -86.8  -466.0 -34.9  58.2 0.53 0.000 0.314 1
 20 2.52  4 2.55 -1590.4   327.5 -1303.7    -1.7   -91.7  -464.0 -56.8  -6.6 0.55 0.000 0.000 2
 21 1.75  7 1.75 -1588.1   335.0 -1267.8   -11.3   -84.3  -500.3 -59.5   2.2 0.55 0.420 0.640 1
 22 2.72  3 3.02 -1530.8   335.3 -1241.5    -9.6   -79.4  -489.6 -46.0  59.6 0.55 0.000 0.566 1
 23 3.39  5 3.29 -1547.5   337.8 -1274.6    -6.4   -78.3  -476.1 -49.8  42.9 0.55 0.000 0.943 1
 24 3.44  7 3.34 -1479.3   354.7 -1237.9    -3.2   -81.3  -480.5 -30.8 111.0 0.55 0.000 0.483 0
 25 1.10  4 1.09 -1544.3   315.9 -1249.3    22.7   -85.0  -502.0 -46.6  46.1 0.55 0.000 0.376 1
 26 3.51  7 3.52 -1501.9   331.6 -1182.2     0.2   -84.3  -532.8 -34.3  88.5 0.55 0.000 0.907 0
 27 2.40  6 2.40 -1525.8   347.3 -1205.7     5.3   -76.7  -548.7 -47.3  64.6 0.55 0.000 0.172 0
 28 1.32  3 1.32 -1528.9   330.4 -1263.1    14.3   -88.3  -479.6 -42.6  61.5 0.55 0.000 0.138 0
 29 1.35  4 1.57 -1616.3   312.7 -1291.6    -7.6   -80.4  -483.9 -65.5 -25.9 0.55 0.000 0.000 2
 30 0.77  1 0.87 -1581.4   326.5 -1284.8    -4.3   -82.9  -482.6 -53.3  34.8 0.55 0.000 0.174 1
 31 3.06  4 3.06 -1508.9   332.2 -1184.5    10.5   -83.2  -547.7 -36.0 107.4 0.55 0.000 0.700 0
```

For the above compact characterization of the global search trajectory, column headings are defined as follows:

Ri	RMSD for inflation trajectory
Rd	RMSD for cycle, including inflation, deflation, and segment search
oM	maximum number for steps in parameter space
Frsc	(Fr+Fs+Fc)
Ftb	(Ft+Fb)
Fgp	(Fg+Fp)
dtot	change in Ftot from previously accepted step
fold	measure of extent of folding, in range $[0,1]$
z	statistical weight calculated as $\exp(-\mathrm{f\,dtot/kT})$, where factor f, in range $[\,.125,\ .250]$, is dependent on fold
rn	random number in range $[0,1]$ used for Monte Carlo acceptance decision
A	acceptance decision: 0=rejected, 1=used as an intermediate point in a multicycle exploration of neighboring local minima, 2=accepted

A measure of extent of folding is defined as fold= (hd +ha)/(nd +na), where nd and na are the number of peptide H-bond donors and acceptors, and hd and ha are the number of peptide H-bond donors and acceptors involved in peptide-peptide H-bonds.

The algorithm targets a range of 2 to 4 Å for RMS deviations between local minima sampled in consecutive cycles.

As determined in the test case for the "rcyc" command, the energy of the refined native conformation (MOL=1pga_r, CNF=05) is -1663.22 kcal/mol. From file "test/dgn/ptra.1pga_02i.s000.full", the energy of the endpoint of the "ptra" trajectory (MOL=1pga_02i, CNF=t029) is -1616.3 kcal/mol, higher by 47 kcal/mol relative to the native conformation, but lower by 88 kcal/mol relative to the starting point of the trajectory (MOL=1pga_02i, CNF=s000). Comparison of structures between the start and end of the trajectory shows expansion of the β-sheet and attainment of the experimental twist.

On parallel computer systems, a strategy found to be effective is parallel trajectory searches using as starting points the 16 folds produced by the "igor" command, one fold per node, distributed over 16 nodes.

This test case demonstrates, perhaps counterintuitively, an ability to predict structure with modest computer power using the current energy function, which is expensive in comparison to energy functions more commonly used for modeling of proteins and nucleic acids.

Trajectories produced by the "ptra" command are chaotic in the sense that trajectories produced by different processors will quickly diverge if any difference occurs, even in the smallest bit, as a result of floating point operations or calls to math library functions. So while a trajectory produced by the "ptra" command is reproducible on an identical processor, or on different runs using the same processor, a trajectory produced by an Apple M1 processor will differ from a trajectory produced by an Intel Core i7 processor. This test case was run on an Apple M1 processor.

6.9 ionstate

FUNCTIONALITY
ionizable group pKa prediction

SYNTAX
ionstate FAM MOL

INPUT FILES	OUTPUT FILES
/FAM/exp/MOL.pdb	/FAM/dgn/ionstate.MOL
	/FAM/dgn/ionstate_titration.MOL
	/FAM/seq/seq.MOL???
	/FAM/tor/tor.MOL???.ph
	/FAM/car/MOL???.ph.pdb

SUMMARY

1) inputs a pdb-format file, "FAM/exp/MOL.pdb", most often a structure prepared by the "ereg" command and copied from directory "FAM/car"

2) regularizes geometry, meaning bond lengths, bond angles, and some torsion angles are adjusted to standard values with minimal movement of atom coordinates

3) for each pH ??.? (where ? is a digit of pH expressed to 1 decimal place), outputs (in file "FAM/seq/seq.MOL???") residue sequence and disulfide crosslinks for the most probable ionization state at pH=??.?

4) for each pH ??.?, outputs (using CNF="ph") 2 alternative specifications (in pdb and torsion angle formats) of the most probable conformation for the sequence of the most probable ionization state at pH=??.?

5) outputs a diagnostic summary of the execution of the command in 2 files: "FAM/dgn/ionstate.MOL" and "FAM/dgn/ionstate_titration.MOL"

NOTES

A common use of the "ionstate" command is to calculate the most probable ionization state for a specified pH, and to generate a most probable conformation consistent with this ionization state. This usage enables subsequent modeling of the most probable sequence, including protonation state, for a given pH.

The outer loop of the command is a titration over a wide range of pH values. For each pH value, search through the combinatorial space of whole-protein ionization states is used to isolate the collection of ionization states having the lowest free energies. For each whole-protein ionization state x, the free energy dG(x,pH) is calculated as a sum of free energies for 2 component processes: g(x,pH), the free energy required to create x in the absence of the protein environment, and dF(x), the energy contributed by the protein environment to stabilization of x. For each ionizable group, the titration data is used to calculate $\mathrm{pK_a}$ in the environment of the protein.

g(x,pH) is calculated as a sum over the subset of ionizable groups such that the ionization state in x differs form the ionization state observed for a model peptide at the current pH. The function being integrated is the experimental free energy required to change the state of the ionizable group in a model peptide at the current pH. dF(x) is calculated as the full protein model energy of ionization state x minus the full protein model energy of the fully protonated reference state minus the sum, over ionizable groups deprotonated in x, of the change in model energy for the process of deprotonation for the ionizable group in a model peptide. dG(x,pH) is calculated as g(x,pH) plus h(dF(x)), where the function h reduces differences in dF between conformations. The damping function h() is included to account for relaxation of structure with respect to degrees of freedom held fixed in the model. Details of the calculation are described in a publication available on the company website.

The "ionstate" command accesses the energy surface. Using an Apple M1 Pro processor, titration of the 96 residue protein 1lni requires computation time of about 20.0 hours.

EXAMPLE TEST CASE

File "1lni.pdb", a crystal structure of a 96-residue protein, ribonuclease from streptomyces aureofaciens, has been entered into directory "test/exp". To best model protonation of the carboxyl groups of ASP and GLU, the original PDB database file was modified as follows. The structure was viewed graphically. For each ASP (or GLU) residue, coordinates for the pair of atoms {OD1,OD2} of ASP (or {OE1,OE2} of GLU) are optionally exchanged such that the OD2 of ASP (or OE2 of GLU) is the atom of the pair more freely accessible to addition of a proton.

To execute the command, open a window to directory "src/str", and type the following lines to the macOS (or linux) prompt.

```
%  ereg test 1lni
%  cp ../../test/car/1lni.05.pdb ../../test/exp/x1lni.pdb
%  ionstate test x1lni
```

The output diagnostic file "/test/dgn/ionstate.x1lni" contains information that, while useful for algorithm development, is best ignored for application usage of the command. File "/test/dgn/ionstate_titration.x1lni" contains a compact summary of the titration.

In the following abbreviated listing, frame boxes enclose descriptions of the output data.

file="test/dgn/ionstate_titration.x1lni"

Diagnostic file created by the command:
% ionstate test x1lni

```
pKa OF IONIZABLE GROUPS
  res      pept   prot
   1 eASP [ 7.50]  8.70
   1 eASP [ 4.00]  3.10
  14 GLU  [ 4.40]  4.15
  17 ASP  [ 4.00]  4.25
  25 ASP  [ 4.00]  5.20
  30 TYR  [ 9.60] 10.45
  33 ASP  [ 4.00]  1.90
  41 GLU  [ 4.40]  3.55
  49 TYR  [ 9.60]  9.60
  51 TYR  [ 9.60] 10.95
  52 TYR  [ 9.60] 12.00
  53 HIS  [ 6.30]  6.65
  54 GLU  [ 4.40]  2.90
  55 TYR  [ 9.60] 10.60
  74 GLU  [ 4.40]  4.35
  78 GLU  [ 4.40]  3.65
  79 ASZ  [ 4.00]  6.65
  80 TYR  [ 9.60] 12.35
  81 TYR  [ 9.60] 10.45
  84 ASP  [ 4.00]  4.00
  85 HIS  [ 6.30]  4.40
  86 TYR  [ 9.60] 10.40
  93 ASP  [ 4.00]  6.05
  96 CYSe [ 3.80]  2.70
```

For each ionizable group, the above listing shows pKa calculated in the environment of the protein. For comparison, the value in brackets is experimental for the functional group in a peptide.

```
TITRATION
  pH    DDEDDYDEYYYHEYEEDYYDHYDC   dG     dF      g    p       pp      h       ps      ss
  0.00  000000000000000000000000 -18.98  -0.07   0.00 0.913 -112.93  -59.57 -400.85  152.30
  0.10  000000000000000000000000 -18.98  -0.07   0.00 0.893 -112.93  -59.57 -400.85  152.30
  0.20  000000000000000000000000 -18.98  -0.07   0.00 0.869 -112.93  -59.57 -400.85  152.30
   .
   .
   .
  3.40  010000100000100000001001 -23.44 -30.94   7.50 0.206 -275.16  -59.83 -173.12   59.30
  3.50  010000100000100000001001 -24.12 -30.94   6.82 0.170 -275.16  -59.83 -173.12   59.30
  3.60  010000100000100000001001 -24.80 -30.94   6.14 0.129 -275.16  -59.83 -173.12   59.30
  3.70  010000110000100100010001 -25.52 -29.75   4.23 0.098 -272.95  -59.80 -180.55   61.26
  3.80  010000110000100100010001 -26.48 -29.75   3.27 0.107 -272.95  -59.80 -180.55   61.26
   .
   .
   .
  6.40  011110110000101100011011 -24.03 -27.44   3.41 0.309 -191.08  -59.39 -313.14  118.91
  6.50  011110110000101100011011 -23.76 -27.44   3.68 0.285 -191.08  -59.39 -313.14  118.91
  6.60  011110110000101100011011 -23.49 -27.44   3.96 0.251 -191.08  -59.39 -313.14  118.91
  6.70  011110110001101110011011 -23.35 -21.46   0.00 0.269 -112.82  -59.43 -430.23  168.51
  6.80  011110110001101110011011 -23.35 -21.46   0.00 0.345 -112.82  -59.43 -430.23  168.51
   .
   .
   .
 10.20  111110111001101110011011 -15.91 -16.05   5.73 0.055  -42.31  -59.25 -552.81  221.14
 10.30  111111111001101110111111 -15.09   0.56   3.82 0.045   76.71  -58.90 -766.90  312.35
 10.40  111111111001101110111111 -14.55   0.56   4.37 0.051   76.71  -58.90 -766.90  312.35
 10.50  111111111001101110111111 -14.00   0.56   4.91 0.058   76.71  -58.90 -766.90  312.35
 10.60  111111111001111110111111 -14.05   8.12   4.09 0.145  146.90  -58.82 -886.98  363.04
   .
   .
   .
```

The 1st column specifies the pH.

For each pH of the titration, the 2nd column specifies the ionization state having the lowest free energy. The header of this column is a string of 1-letter code names for the set of ionizable groups listed in the previous section. The binary choice of { 0, 1} is used to indicate { protonated, unprotonated}.

The remaining columns characterize the most-probable ionization state (specified in column 2). Column headings use the following notation.

dG is the calculated free energy.

dF, calculated as f(ionstate) $-$f(fully protonated) $-$df(peptide), is the model energy of the ionization state minus the model energy of the fully protonated reference state minus the sum, over deprotonated ionizable groups, of the change in model energy for the process of deprotonation for the ionizable group in a model peptide.

g is a sum (over the subset of ionizable groups such that the ionization state differs form the ionization state observed for a model peptide at the current pH) of the experimental free energy required to change the state of the ionizable group in a model peptide.

More conceptually, g is the free energy required to create the whole protein ionization state in the absence of the protein environment, and dF is the energy contributed by the protein environment to stabilization of this ionization state. dG is obtained by a merging of dF and g.

p is the probability.

pp=(Fr+Fe+Fs+Ft+Fb+Fg+Fp), the interaction energy of the protein with itself.

h= Fh, the hydrophobic component of the hydration free energy.

ps= Fps, the electrostatic interaction energy of the protein with the boundary surface.

ss= Fss, the electrostatic interaction energy of the boundary surface with itself.

As shown in the above section, 3 of the 12 ASP and GLU side chains change protonation state in the pH range (3.4, 3.8).

In the pH range (6.4, 6.8), HIS 53 and ASP 79 change protonation state.

In the pH range (10.2,10.6), 4 of the 8 TYR side chains change protonation state.

In Table II, calculated values of $\mathrm{pK_a}$ are compared to experimental values for all ionizable groups of 1lni. Also included in Table II are experimental values of $\mathrm{pK_a}$ obtained for the ionizable group in a model peptide.

Table II. Comparison of Calculation to Experiment for $\mathrm{pK_a}$Values of Ionizable Groups of 1lni (a crystal structure of Ribonuclease SA).

	$\mathrm{pK_a}$		
residue	pept[a]	expt[b]	calc[c]
1 eASP	[7.50]	9.14	8.70
1 eASP	[4.00]	3.44	3.10
17 ASP	[4.00]	3.72	4.25
25 ASP	[4.00]	4.87	5.20
33 ASP	[4.00]	2.39	1.90
79 ASZ	[4.00]	7.37	6.65
84 ASP	[4.00]	3.01	4.00
93 ASP	[4.00]	3.09	6.05
14 GLU	[4.40]	5.02	4.15
41 GLU	[4.40]	4.14	3.55
54 GLU	[4.40]	3.42	2.90
74 GLU	[4.40]	3.47	4.35
78 GLU	[4.40]	3.13	3.65
53 HIS	[6.30]	8.27	6.65
85 HIS	[6.30]	6.35	4.40
30 TYR	[9.60]	11.30	10.45
49 TYR	[9.60]	10.60	9.60
51 TYR	[9.60]	11.50	10.95
52 TYR	[9.60]	11.50	12.00
55 TYR	[9.60]	11.50	10.60
80 TYR	[9.60]	11.50	12.35
81 TYR	[9.60]	11.50	10.45
86 TYR	[9.60]	11.50	10.40
96 CYSe	[3.80]	2.42	2.70

[a] Experimental value for the ionizable group in a model peptide.
[b] Experimental value in the protein.
[c] Calculated value in the protein.

For 18 of the 24 ionizable groups, calculation reproduces the experimental direction of change in the $\mathrm{pK_a}$ value from a model peptide environment to the protein environment.

An unusual property of this protein is the observed protonation of an ASP residue (ASP 79) at pH 7.0. The calculated $\mathrm{pK_a}$ value for Asp 79 is 6.65.

6.10 edoc

FUNCTIONALITY
docking prediction

SYNTAX
edoc FAM MOL1, MOL2, ... MOLn

INPUT FILES	OUTPUT FILES
/FAM/exp/MOL1.pdb	/FAM/dgn/edoc.MOL1_MOL2
/FAM/exp/MOL2.pdb	/FAM/car/MOL1_MOL2.e????.pdb

SUMMARY

1) in the most common usage, inputs 2 pdb-format files, "FAM/exp/MOL1.pdb" and "FAM/exp/MOL2.pdb", packing of 3 or more bodies can be directed by adding structure names to the list of command line arguments

2) regularizes geometry of all input structures

3) generates a collection of 64 docked conformations

4) outputs the collection of docked conformations as "FAM/car/MOL1_MOL2.e????.pdb", where ???? is a 4-digit number ordering the collection based on packing score

5) outputs, in file "FAM/dgn/edoc.MOL1_MOL2", a diagnostic summary of the execution of the command

NOTES

The input structures are packed as rigid bodies. For docking of 2 rigid bodies, the space of conformations is defined by 6 degrees of freedom, translation and rotation of the 2nd body with respect to the 1st. This continuous space is replaced with a grid consisting of 104857600 discrete conformations. A packing score combines a measure of surface complementarity, a measure of the depth of interpenetration, and a fast approximation of the Fe and Fp components of the full energy. The search algorithm optimizes the packing score over the discretized space, generating a collection of lowest scoring docked conformations.

EXAMPLE TEST CASE

PDB database entry 3bzd contains a protein-protein complex between a fragment of a mouse T cell receptor and a bacterial toxin. To prepare for execution of the test case, file "3bzd.pdb" has been entered into directory "test/exp". Files "3bzda.pdb" and "3bzdb.pdb" have been created from file "3bzd.pdb" by copying and hand editing to isolate the 2 proteins of the complex.

To execute the command, open a window to directory "src/str", and type the following line to the macOS (or linux) prompt.

```
%  edoc test 3bzda 3bzdb
```

The collection of 64 lowest scoring docked conformations is enriched in conformations close to the experimental structure. The match to experiment is almost perfect for CNF=0017 and CNF=0059. For CNF=0001, the toxin is placed in the correct binding site of the T cell receptor, but slightly rotated relative to the experimental conformation.

file="test/dgn/edoc.3bzda_3bzdb"

Diagnostic file created by the command:
% edoc test 3bzda 3bzdb

```
2-BODY ENERGIES
PAIR OF BODS=( 0, 1)
SEARCH ON SPARSE GRID
 oD9 oZ9
```

```
4096   2
  D9  tot     e      m      s      h      p      i      d     dexp  pF pT lF lT iJ
   0-19.503  0.672  1.026 -6.275 -0.701 -6.073 -0.099 -8.053   0.25  2 21  8 21 20
   1-19.434  1.358  1.400 -7.779 -1.313 -7.383 -0.074 -5.643   2.44 15  5  1 53 60
   2-19.347  1.720  0.833 -6.113 -1.362 -6.949  0.291 -7.768   3.86 13 37  8 21  4
   3-19.268  0.381  0.601 -5.589 -0.776 -5.981  0.008 -7.913  -0.91  2 37  8 21 32
  .
  .
  .
4095 -7.770  0.266  3.910 -2.481 -0.428 -0.946  0.070 -8.161  -1.46  9 21  9  5 28
```

A listing of lowest scoring docked conformations obtained by sparse search over 102400 conformations of the full grid. The full grid consists of 104857600 conformations.

Score components s, h, p, and i measure complementarity of docked molecular surfaces for, respectively, shape, hydrophobicity, polarity, and charge induced on the boundary of a dielectric continuum. Score components e and m are fast estimates of, respectively, the Fe and Fp components of the full energy. Score component d is a measure of the depth of interpenetration for the packed structures.

Column heading dexp is a geometric measure of the match between the predicted and observed conformations. This measure, is meaningful only if the coordinates of the input structures being packed are the coordinates of the observed complex. The range of dexp is $[-4,\ 4]$. When meaningful, higher values are better matches.

In this example, input structures were created by copying the crystal structure of the complex, and editing to isolate the receptor and the toxin. The calculation and listing of dexp facilitates testing of the algorithm.

Column headings pF, pT, lF, lT, and iJ are indices over the 5-dimensional full grid. Indexes pF and lF range over 20 faces of an icosahedran. Indexes pT and lT range over 64 directions within a face. Index iJ ranges over 64 rotations about an axis separating the 2 bodies. For each grid point, a sixth degree of freedom, the separation distance between the 2 bodies is optimized as part of the calculation of the surface complementarity components.

Based on values of dexp, conformations 1 and 2 are good matches.

```
SEARCH ON INTERMEDIATE GRID
 oD9 oZ9
8192   2
  D9  tot     e      m      s      h      p      i      d     dexp  pF pT lF lT iJ
   0-21.688  0.203  0.439 -6.424 -1.482 -6.649 -0.277 -7.499  -1.08 17 25  9  9 24
   1-21.324  0.605  0.477 -6.413 -1.413 -6.639 -0.422 -7.519  -1.12 17 25  9  9 22
   2-20.947  0.081  0.389 -6.022 -1.449 -6.238 -0.138 -7.571  -1.01 17 25  9  9 26
  .
  .
  .
  18-19.839  0.817  1.141 -6.206 -1.316 -6.840  0.110 -7.545   3.20 13 37  8 41 60
  .
  .
  .
  26-19.569  1.749  0.667 -5.719 -1.295 -7.171  0.086 -7.887   3.93 13 45  8 21  2
  .
  .
  .
8191-11.989  1.995  2.289 -3.638 -0.911 -4.473 -0.275 -6.976  -0.38  2  5  9  9 58
```

A listing of lowest scoring docked conformations obtained by intermediate search over the grid in the neighborhoods of the previously identified lowest scoring conformations.

Conformations 18 and 26 are strong matches.

```
SEARCH ON FULL GRID
 oD9 oZ9
8192   2
  D9  tot     e      m      s      h      p      i      d     dexp  pF pT lF lT iJ
   0-22.429 -0.550  1.411 -7.856 -1.513 -7.505 -0.268 -6.149   1.03  7 26  1 61  9
   1-22.248 -0.122  1.564 -8.123 -1.492 -7.416 -0.162 -6.497   0.85  7 26  1 62 11
   2-21.892  0.639  1.430 -8.070 -1.454 -7.536 -0.462 -6.437   0.99 14 63  1 62 10
   3-21.852  1.438  0.952 -7.788 -1.314 -7.093  0.022 -8.069   2.59 11  3  8 22 16
  .
  .
  .
  58-20.777 -0.055  1.053 -5.803 -1.250 -7.144  0.220 -7.798   3.43 13 40  8 42 63
  .
  .
  .
  98-20.450  0.063  1.667 -6.271 -1.383 -7.026  0.351 -7.852   3.85 13 11  8 21 60
  .
  .
  .
8191-11.991  1.339  5.253 -6.272 -1.267 -5.739 -0.093 -5.212   0.17  6  9  1 57  6
```

A listing of lowest scoring docked conformations obtained by search over the full grid in the neighborhoods of the previously identified lowest scoring conformations.

Conformations 58 and 98 are strong matches.

```
CLUSTER
COLLECTION OF DOCKED CONFORMATIONS
 oD9 oZ9
 273   2
  D9  tot     e      m      s      h      p      i      d     dexp  pF pT lF lT iJ
   0-22.429 -0.550  1.411 -7.856 -1.513 -7.505 -0.268 -6.149   1.03  7 26  1 61  9
   1-21.852  1.438  0.952 -7.788 -1.314 -7.093  0.022 -8.069   2.59 11  3  8 22 16
   2-21.819 -0.017  0.121 -6.139 -1.414 -6.418  0.162 -8.115  -0.17 17 29  8 23 54
  .
  .
  .
  17-20.777 -0.055  1.053 -5.803 -1.250 -7.144  0.220 -7.798   3.43 13 40  8 42 63
  .
  .
  .
  59-18.779  0.773  0.793 -5.196 -0.994 -6.483  0.158 -7.829   3.82 13 45  8 18  5
  .
  .
  .
 272-12.012  0.647  0.865 -3.180 -0.477 -1.753  0.160 -8.275  -0.46  1 29  8 29 38
```

Clustering partitions the previous collection of 8192 lowest scoring conformations into 272 clusters. For each cluster, the lowest scoring conformation is retained as representative.

Following clustering, conformation 1 is a good match to experiment. Conformations 17 and 59 are almost perfect.

The "edoc" command does not access the energy surface but can be computationally intensive. Using an Apple M1 Pro processor, this test case requires computation time of about 5.6 hours.

6.11 Package Unique Functionalities

Because the "ereg", "estp", "rcyc", "ptra", and "ionstate" commands access the energy surface and the conformational search algorithms, the functionality provided by these commands is unique to this package. Also, the functionality of the "greg" and "igor" commands is, we believe, unique. In contrast, for the "prof", "hlog", and "edoc" commands, there exist alternative, possibly better, almost certainly faster, software packages capable of accomplishing structure quality assessment, homology model building, and docking. However, as opposed to interfacing to external software, the above commands have the benefit of maintaining consistent rigid geometry along with the file and directory organization of the ereg package.

7 UTILITY COMMANDS

The following 2 commands facilitate format conversion.

7.1 seqformat

FUNCTIONALITY
   sequence format conversion

SYNTAX
   seqformat FAM MOL

INPUT FILES	OUTPUT FILES
/FAM/exp/seq.MOL	/FAM/seq/seq.MOL

SUMMARY

1) inputs the sequence of a single polypeptide chain in 1-letter code format

2) outputs the sequence in an alternative format that allows specification of disulfide crosslinks, multiple chains, alternative ionization states, and residues outside the set of 20 naturally occurring amino acids

EXAMPLE TEST CASE

In this example, MOL="grb2" is the sh3 domain of growth factor bound protein 2.

To execute the command, open a window to directory "src/str", and type the following line to the macOS (or linux) prompt.

```
%  seqformat test grb2
```

7.2 initcar

FUNCTIONALITY
initial structure generation

SYNTAX
initcar FAM MOL CNF

INPUT FILES	OUTPUT FILES
/FAM/seq/seq.MOL	/FAM/dgn/initcar.MOL
/FAM/tor/tor.MOL.CNF (optional)	/FAM/car/MOL.CNF.pdb
	/FAM/tor/tor.MOL.CNF (if not present)

SUMMARY

1) inputs a sequence, "FAM/seq/seq.MOL"

2) inputs (if present) an initial conformation in torsion angle format, "FAM/tor/tor.MOL.CNF"

3) if torsion angle input is not present, assigns, for each residue, backbone torsion angles consistent with the most probable residue state calculated using a residue Ising model, and an extended conformation of the side chain

4) outputs, in file "FAM/dgn/initcar.MOL", a diagnostic summary of the execution of the command

5) outputs, in pdb-format file "FAM/car/MOL.CNF.pdb", cartesian coordinates of the initial conformation

6) if not present initially, outputs in file "FAM/tor/tor.MOL.CNF", torsion angle coordinates of the assigned conformation

NOTES

To initiate a new conformation by hand editing, or to alter an existing program generated conformation, the easiest way is by creating or modifying a torsion angle format file "FAM/tor/tor.MOL.CNF".

The most common use of the "initcar" command is to generate a corresponding pdb-format file for a hand modified torsion angle format file. This functionality is a simple format conversion for the specification of a conformation.

EXAMPLE TEST CASE

A conformation CNF="unfolded" is created for protein MOL="1pga". The most probable secondary structure state is shown in diagnostic file "test/dgn/initcar.1pga".

To execute the command, open a window to directory "src/str", and type the following line to the macOS (or linux) prompt.

```
%  initcar test 1pga unfolded
```

8 USER CONTROL OF EXECUTION

For some commands, the user can modify execution by changing the default values of variables listed in file "src/str/thread_config".

For most commands, changing the value of the VERBOSE variable from 1 to 0 silences some of the output to the diagnostic summary of the execution of the command.

The "ereg" command directs a sequence of local minimization trajectories on a sequence of approximate RESTBCHPW energy surfaces, where the approximate energy surfaces differ only by the coefficient for the Fc component, which decreases to zero. The Fc component is a sum of harmonic distance constraints to values taken from the starting conformation. The full RESTBCHMGP energy is evaluated at the endpoint of each local minimization trajectory. By default, this sequence is terminated, and the current local minimization trajectory is discarded, if the full energy increases relative to the previous single point evaluation. Changing the value of the ALTEREG variable from 0 to 1 continues the sequence of local minimization trajectories, even after full energy increases.

The NCYCLES variable controls the number of cycles directed by the "ptra" command.

9 FILE FORMATS

In this section, we describe, using examples, the input file types of Table III.

Table III. Types of Input Files.

General Name	Content
/FAM/seq/seq.MOL	residue sequence
/FAM/tor/tor.MOL.CNF	torsion angles
/FAM/stp/stp.SUB.MOL	subset of degrees of freedom
/FAM/car/dprof.MOL.CNF	defect energy density profile
/FAM/arg/exptstructs.GRP	group of templates
/FAM/exp/seq.MOL	1-letter code residue sequence

9.1 residue sequence

Frame boxes enclose descriptions of the file format. Content of these boxes should not be included in actual input files. Formats for input lines are specified using the compact Fortran notation for fields and repeat counts.

file="test/seq/seq.1crn"

```
    1
```

Number of polymer chains. FORMAT=(1x,i4).

```
   46
```

Number of residues in first chain. FORMAT=(1x,i4).

```
|eTHR|THR |CYS |CYS |PRO |SER |ILE |VAL |ALA |ARG |SER |ASN |PHE |ASN |VAL |CYS
|ARG |LEU |PRO |GLY |THR |PRO |GLU |ALA |ILE |CYS |ALA |THR |TYR |THR |GLY |CYS
|ILE |ILE |ILE |PRO |GLY |ALA |THR |CYS |PRO |GLY |ASP |TYR |ALA |ASNe
```

Sequence of residues. FORMAT=(16(1x,a4)). N- and C-terminal residues, because these differ in composition and connectivity, are distinct from interior residues having the same amino acid side chain.

```
    3
```

Number of disulfide bonds. FORMAT=(1x,i4).

```
    1    3    1   40
    1    4    1   32
    1   16    1   26
```

(chain index, residue index) of 1st then 2nd half cystine. One disulfide bond per line. FORMAT=(4(1x,i4)).

If a system of molecules contains n chains, the residue sequences of chains 2–n are specified following the specification for chain 1, and prior to specification of disulfide bonds.

For nucleic acid chains, the most natural choice for the unit of composition, the nucleotide, was not used. Instead, each nucleotide is partitioned into an ordered triple of residues (a pentose ring, a nucleic acid base, a backbone phosphate). This choice, although it imposes a burden of required preprocessing for structures taken from the PDB database, was considered preferable to a combinatorial explosion of residue types.

Table IV. Currently Available Residues.

Num	Res	Class[a]	Description[b]
1	ALA	a	alanine
2	ASP	a	aspartic acid (ionized)
3	CYS	a	half cystine
4	GLU	a	glutamic acid (ionized)
5	PHE	a	phenylalanine
6	GLY	a	glycine
7	HIS	a	histidine (N-delta protonated)
8	ILE	a	isoleucine
9	LYS	a	lysine (ionized)
10	LEU	a	leucine
11	MET	a	methionine
12	ASN	a	asparagine
13	PRO	a	proline
14	GLN	a	glutamine
15	ARG	a	arginine (ionized)
16	SER	a	serine
17	THR	a	threonine
18	VAL	a	valine
19	TRP	a	tryptophan

continued on next page

continued from previous page

Num	Res	Class	Description
20	TYR	a	tyrosine
21	AIB	a	aminoisobutyric acid
22	ABU	a	aminobutyric acid
23	NLE	a	norleucine
24	ORN	a	ornithine (ionized)
25	CYH	a	cysteine
26	HIE	a	histidine (N-epsilon protonated)
27	HIP	a	histidine (ionized)
28	ASZ	a	aspartic acid (neutral, protonated)
29	CYZ	a	cysteine (ionized, deprotonated)
30	GLZ	a	glutamic acid (neutral, protonated)
31	LYZ	a	lysine (neutral, deprotonated)
32	TYZ	a	tyrosine (ionized, deprotonated)
33	ACE	e	N-terminal acetyl
34	NME	e	C-terminal N-methyl
35	eALA	a	N-terminal ALA
36	eASP	a	N-terminal ASP
37	eCYS	a	N-terminal CYS
38	eGLU	a	N-terminal GLU
39	ePHE	a	N-terminal PHE
40	eGLY	a	N-terminal GLY
41	eHIS	a	N-terminal HIS
42	eILE	a	N-terminal ILE
43	eLYS	a	N-terminal LYS
44	eLEU	a	N-terminal LEU
45	eMET	a	N-terminal MET
46	eASN	a	N-terminal ASN
47	ePRO	a	N-terminal PRO
48	eGLN	a	N-terminal GLN
49	eARG	a	N-terminal ARG
50	eSER	a	N-terminal SER
51	eTHR	a	N-terminal THR
52	eVAL	a	N-terminal VAL
53	eTRP	a	N-terminal TRP
54	eTYR	a	N-terminal TYR
55	eAIB	a	N-terminal AIB
56	eABU	a	N-terminal ABU
57	eNLE	a	N-terminal NLE
58	eORN	a	N-terminal ORN
59	eCYH	a	N-terminal CYH
60	eHIE	a	N-terminal HIE
61	eHIP	a	N-terminal HIP
62	eASZ	a	N-terminal ASZ
63	eCYZ	a	N-terminal CYZ
64	eGLZ	a	N-terminal GLZ
65	eLYZ	a	N-terminal LYZ
66	eTYZ	a	N-terminal TYZ
67	ALAe	a	C-terminal ALA

continued on next page

continued from previous page

Num	Res	Class	Description
68	ASPe	a	C-terminal ASP
69	CYSe	a	C-terminal CYS
70	GLUe	a	C-terminal GLU
71	PHEe	a	C-terminal PHE
72	GLYe	a	C-terminal GLY
73	HISe	a	C-terminal HIS
74	ILEe	a	C-terminal ILE
75	LYSe	a	C-terminal LYS
76	LEUe	a	C-terminal LEU
77	METe	a	C-terminal MET
78	ASNe	a	C-terminal ASN
79	PROe	a	C-terminal PRO
80	GLNe	a	C-terminal GLN
81	ARGe	a	C-terminal ARG
82	SERe	a	C-terminal SER
83	THRe	a	C-terminal THR
84	VALe	a	C-terminal VAL
85	TRPe	a	C-terminal TRP
86	TYRe	a	C-terminal TYR
87	AIBe	a	C-terminal AIB
88	ABUe	a	C-terminal ABU
89	NLEe	a	C-terminal NLE
90	ORNe	a	C-terminal ORN
91	CYHe	a	C-terminal CYH
92	HIEe	a	C-terminal HIE
93	HIPe	a	C-terminal HIP
94	ASZe	a	C-terminal ASZ
95	CYZe	a	C-terminal CYZ
96	GLZe	a	C-terminal GLZ
97	LYZe	a	C-terminal LYZ
98	TYZe	a	C-terminal TYZ
99	H2O	s	water
100	NH2	e	C-terminal NH2
101	UNK	a	unknown
102	D	r	deoxyribose
103	R	r	ribose
104	RME	r	2’O-methyl
105	RF	r	2’flouro
106	MOE	r	2’O-methoxyethyl
107	LNA	r	locked nucleic acid
108	CET	r	2’O-constrained ethyl
109	PO	p	phosphodiester
110	PSR	p	phosphorothioate R isomer
111	PSS	p	phosphorothioate S isomer
112	5OH	p	5’OH
113	3OH	p	3’OH
114	5PO	p	5’phosphate
115	3PO	p	3’phosphate

continued on next page

continued from previous page

Num	Res	Class	Description
116	N	b	NH2 (a minimal replacement of base)
117	A	b	adenine
118	AP	b	ionized adenine (N1 protonated)
119	G	b	guanine
120	GP	b	ionized guanine (N7 protonated)
121	GM	b	ionized guanine (N1 deprotonated)
122	T	b	thymine
123	TM	b	ionized thymine (N3 deprotonated)
124	C	b	cytosine
125	CP	b	ionized cytosine (N3 protonated)
126	U	b	uracil
127	UM	b	ionized uracil (N3 deprotonated)
128	SEP	a	phosphorylated SER
129	THP	a	phosphorylated THR
130	TYP	a	phosphorylated TYR
131	eSEP	a	N-terminal SEP
132	eTHP	a	N-terminal THP
133	eTYP	a	N-terminal TYP
134	SEPe	a	C-terminal SEP
135	THPe	a	C-terminal THP
136	TYPe	a	C-terminal TYP

[a] classes: a= amino acid, r= nucleic acid pentose ring, b= nucleic acid base, p= nucleic acid backbone phosphate, c= carbohydrate, l= lipid, e= end group, s= small molecule

[b] All N-terminal amino and C-terminal carboxyl groups are ionized. For each N- and C-terminal amino acid residue, a corresponding residue, with e replaced by z in the name, having a neutralized amino or carboxyl group is created by the program.

9.2 torsion angles

file="test/tor/tor.1pga.05"

```
 1   26.7353151  34.1570577  35.8877045  72.4057615 128.3637055  -7.4901440
```

```
Chain translation and rotation, one chain per line.
           chain index
           |      x translation (angstrom)
           |      |      y translation
           |      |      |      z translation
           |      |      |      |      euler angle α (degree)
           |      |      |      |      |      euler angle β
           |      |      |      |      |      |      euler angle γ
FORMAT(i2,1x,f12.7,f12.7,f12.7,f12.7,f12.7,f12.7).
```

```
------------------------------------------------------------------------------
```

Horizontal lines separate chains.

```
eMET -171.0208 136.7638 173.4534 -72.0838-178.8269-108.2172  56.1787
THR   -99.1724 130.3982-179.6647 -59.9758 -55.6014  55.6825
TYR  -121.5744 143.7055-175.4898 -79.4262  84.6499 -20.0772
```

```
   .
   .
   .
VAL  -120.5247 120.6641-176.3758  53.0370  60.0239  61.0509
THR  -121.8101 132.1697 167.8736 -63.1352  88.4383  54.9119
GLUe  -99.8123 114.7693          -68.3970-147.0508 -97.3108
```

Torsion angles, IUPAC definitions, one residue per line.

```
        residue name
        |    φ (degree)
        |    |   ψ
        |    |   |   ω
        |    |   |   |   χ1
        |    |   |   |   |   χ2
        |    |   |   |   |   |   χ3
        |    |   |   |   |   |   |   χ4
        |    |   |   |   |   |   |   |   χ5
        |    |   |   |   |   |   |   |   |   χ6
        |    |   |   |   |   |   |   |   |   |   χ7
FORMAT(a4,1x,f9.4,f9.4,f9.4,f9.4,f9.4,f9.4,f9.4,f9.4,f9.4,f9.4).
```

Torsion angles that do not exist for a residue, for example ω for a C-terminal residue, can be left blank.

For residues of the classes r, b, and p that compose nucleic acid chains, the order used for torsion angles can be found in data file "src/dat/residue_interface", input to variable "T0tor".

Because the pdb file format, used in this package to store conformations, is a widely used standard, it is, with the following exception, defined elsewhere. For nucleic acid chains, the order used for atoms within residues of classes r, b, and p can be found in data file "src/dat/residue_interface", input to variable "P0atm".

The most common failure in reading a pdb format file occurs when a structure file contains multiple Hydrogen atoms bonded to the same heavy atom. Standard PDB database format specifies 4 character atom names placed in columns 13–16, with the element symbol in column 14. In cases of multiple H atoms bonded to C or N, for example the 3 H atoms of a methyl group, the character 1, 2, or 3 that distinguishes these H atoms should be in column 13, before the element symbol. For each residue in the dataset, the PDB database format atom names can be found in data file "src/dat/residue_interface", input to variable "P0atm".

9.3 subset of degrees of freedom

file="test/stp/stp.g006.1pga"

```
  11
```

Number of residues containing torsion angles that will be allowed to vary during energy minimization. FORMAT=(i4).

```
[ 1:   3]  0 2 0  TYR
[ 1:   5]  0 2 0  LEU
[ 1:  24]  0 0 0  ALA
[ 1:  25]  2 2 7  THR
[ 1:  26]  2 1 0  ALA
[ 1:  27]  2 2 0  GLU
[ 1:  28]  2 2 0  LYS
[ 1:  29]  2 2 0  VAL
[ 1:  30]  2 2 0  PHE
[ 1:  31]  2 2 0  LYS
[ 1:  52]  0 2 0  PHE
```

```
One line per residue.
          '[' character
          |  chain index
          |  |  ':' character
          |  |  |  residue index
          |  |  |  |  ']' character
          |  |  |  |  |
          |  |  |  |  |      backbone search parameter, controls
          |  |  |  |  |      conformational search characteristics of
          |  |  |  |  |      torsion angles ($\omega$, $\phi$, $\psi$), where $\omega$
          |  |  |  |  |      is the final backbone torsion angle of the
          |  |  |  |  |      preceding residue, [0=held fixed, 1=allowed to
          |  |  |  |  |      vary, 2=actively assigned alternative values]
          |  |  |  |  |      |
          |  |  |  |  |      |  side chain search parameter, controls
          |  |  |  |  |      |  conformational search characteristics of
          |  |  |  |  |      |  side chain torsion angles, [0=held fixed,
          |  |  |  |  |      |  1=allowed to vary, 2=actively assigned
          |  |  |  |  |      |  alternative values]
          |  |  |  |  |      |  |
          |  |  |  |  |      |  |  for the first residue of a segment to be
          |  |  |  |  |      |  |  deformed, set equal to the length of the
          |  |  |  |  |      |  |  segment, zero otherwise
          |  |  |  |  |      |  |  |
          |  |  |  |  |      |  |  |     concatenated string of names of
          |  |  |  |  |      |  |  |     residues to be substituted at this
          |  |  |  |  |      |  |  |     position
FORMAT=((1x,i2,1x,i4,1x),1x,i2,i2,i2,2x,32a4).
```

Rules:

1) The length of a deformed segment must be an odd number equal to { 5, 7, 9,11,13}.

2) For each string of residues such that the backbone search parameter is 1 or 2, if the residue preceding this string is not already included in this file, due to side chain torsion angles that will be allowed to vary, then it must be added with backbone and side chain search parameters set to zero. In addition to ϕ and ψ, the backbone search parameter controls the ω torsion angle associated with the preceding residue. As a consequence, for each residue having backbone search parameter equal 1 or 2, the set of residues containing torsion angles that will be allowed to vary during energy minimization must include the preceding residue.

3) For convenience, when used with the command "estp", if the first character of the subset name is ’l’ or ’h’, then automated expansion of the subset is bypassed. Expansion, if not bypassed, adds search of side chains that contact the original set of backbone segments and side chains marked as searchable,

4) For convenience, when used with the command "estp", if the first character of the subset name is ’h’ or ’i’, then the combinatorial search through residue sequence space is truncated to include only the initial and final sequences.

5) For segments of length 13 residues or greater, because the space of possible backbone deformations increases combinatorially with length, and because the space of productive backbone deformations, defined as those that minimize to low energy conformations, remains small, the sample of deformations generated is usually not productive. As a more highly evolved alternative, a segment of length 13 residues or greater can be partitioned into a sequence of adjacent segments of lengths 7 to 11 residues, and conformational search can be directed as simultaneous deformation of the multiple small segments.

Examples of this file type can be found in directory "test/stp/".

When used with the "ptra" command, "stp.SUB.MOL" files specify the subset of torsion angle degrees of freedom that will be allowed to vary. In this case, no distinction is made between the values 1 or 2 for the backbone and side chain search parameters. File "test/stp/stp.c2.1pga" gives an example of directing a search through sequence space. File "test/stp/stp.c777.1pga" gives an example of directing a combinatorial search of multiple, simultaneous segment deformations.

Within the program residue names consist of 4-character strings. The format for an stp file specifies either a single 4-character residue name placed in columns 19–23, or a concatenated string of multiple 4-character residue names beginning in column 19. The most common failure in reading an stp format file occurs because trailing blank characters are omitted from 4-character residue names. This causes a failure in matching the input string to any residue name in the dataset. For all residues in the dataset, the 4-character names can be found in data file "src/dat/residue_mappings", input to variable "L0aa". In this manual, residue names are listed in Table IV.

9.4 defect profile

file="test/car/dprof.grb2.cyc1"

```
Residue..................Overlap............................Unpaired Hbond Acceptor...........
|.........bb Conformation.....Disallowed (phi,psi)..........|....Exposed Hydrophobic Surface..
|.........|.sc Conformation...|....Disallowed omega.........|....|....Secondary Structure.....
|.........|.|.Total Energy....|....|....Disallowed chi......|....|....|....Statistical Contact
|.........|.|.|......isCoil...|....|....|....Buried Charge..|....|....|....|....Electrostatic.
|.........|.|.|......|isExposed....|....|....|....Cavity....|....|....|....|....|.............
|.........|.|.|......||isLoop.|....|....|....|....|....Unpaired Hbond Donor|....|.............
|.........|.|.|......|||.|....|....|....|....|....|....|....|....|....|....|....|.............
```

The above headings describe the columns below.

```
-------------------------------------------------------------------------------------
```

Horizontal lines separate chains.

```
eMET    1   +   3.00 111   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  3.0
GLU     2 E +   1.81 010   0.0  0.0  0.0  0.0  0.0  0.0  0.5  0.0  0.0  1.3  0.0  0.0
ALA     3 E     1.25 000   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.2  0.0  0.0
ILE     4 E +   1.29 010   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.2  0.0  0.0
  .
  .
  .
GLU    54 E +   2.52 111   0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  2.5  0.0  0.0
MET    55 E +   4.03 111   0.0  0.0  0.0  0.0  0.0  0.0  0.5  0.5  0.0  3.0  0.0  0.0
LYSe   56   -   4.50 111   0.0  0.0  0.0  1.0  0.0  0.0  0.0  0.5  0.0  0.0  0.0  3.0
```

Decimal numbers are free energies of chain folding in kcal/mol units. These are attributed to defects in structure given in the column headings.

For side chain conformation, a value of ’-’ indicates the conformation differs from a set of commonly observed rotamers.

```
-------------------------------------------------------------------------------------
```

9.5 group of templates

file="test/arg/exptstructs.sparse"

```
1ckaA_
1bbzA_
1cskA_
1semA_
1oebA_
1jo8A_
```

Template structures, one per line. FORMAT=(a32).

9.6 1-letter code sequence

file="test/exp/seq.grb2"

```
  56
```

Number of residues in chain. FORMAT=(i4).

```
MEAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIPKNYIEMK
```

1-letter code string of residues. FORMAT=(100a1).

```
   0
```

Number of disulfide bonds. FORMAT=(i4).

If nonzero, disulfide bonds are listed, one pair of residue indexes per line, specifying the 1st and 2nd half cystines, respectively. FORMAT=(i4,i4).

- -

For each residue of the chain, a character from the set {$-$,H,E,C}. FORMAT=(100a1).

The current file format is used for input of protein sequences to the "igor" and "seqformat" commands. For the "seqformat" command, this string of characters is not used, but must still be included. For advanced applications of the "igor" command, changing the minus sign to a secondary structure state H, E, or C will bias the igor model score to favor the selected secondary structure state at the residue position. For normal usage, the string of characters should be minus signs.

10 SUPPORT

We offer support and responsiveness to user input. Contact information is provided on the company website.