MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology

Michael Feig, John Karanicolas, Charles L. Brooks III

Presented by Kate Stafford
In consultation with Alex Ismail and Jay Shukla
MMMTSB?

• **What does that mean?**
  Feig et al. describe the **Multiscale Modeling Tools for Structural Biology Tool Set**, available at [http://mmtsb.scripps.edu/software/mmtsbToolSet.html](http://mmtsb.scripps.edu/software/mmtsbToolSet.html).

• **What is it for?**
  The MMTSB Tool Set is a collection of Perl scripts and packages used to interface with a variety of existing structural modeling programs. Many such programs (e.g., CHARMM, MONSSTER, etc.) evolved independently and have their own proprietary scripting languages and input formats; however, they are frequently used together. Enabling ‘communication’ between them is useful in protein modeling.
The Protein Folding Problem

- Knowledge of the ‘native’ three-dimensional conformation assumed by proteins in solution provides clues to protein function.
- Currently, structural information comes from experimental sources: X-ray crystallography (which is difficult and time-consuming to perform) and NMR (which is limited to small soluble proteins).
- Ideally, we could use computational methods to predict protein structure with high accuracy given only the primary sequence.
- This information can be used to understand disease processes and etiologies – for example, protein misfolding diseases such as Alzheimer’s.
Aβ peptide is produced by abnormal proteolysis of amyloid precursor protein (APP).

In its soluble form, Aβ is largely α-helical or disordered. In a nucleation-dependent polymerization process, Aβ monomers self-assemble into large, highly organized fibrillar aggregates with high β-sheet content. These fibers comprise the majority of the extracellular ‘plaques’ in the classical Alzheimer’s brain.
Structural Modeling of Proteins

There are two main approaches to computational prediction of protein structure (Fiser et al. 2002):

• *Ab initio* protein folding
  – Limited to relatively small peptides
  – Computationally expensive and of relatively low accuracy
  – Permit the exploration of structural space not covered by known, solved structures

• Comparative modeling based on a known structural template
  – Requires sequence alignment with a suitable template whose structure has already been solved
  – With sufficiently high sequence identity and good alignment, can be performed with high accuracy
  – Significant decline in accuracy occurs when the template and target have less than 40% sequence identity (Jaroszewski et al. 2002)
Structural Genomics

• The Protein Structure Initiative (Norvell and Machalek 2000) aims to sort known sequences into families and sponsor efforts to solve the structure of a representative member of each family.

• In theory, most of the remaining sequences in each family can then be modeled on the basis of the known structure.

• This effort, like all structural modeling efforts, depends on the availability of good tools for the production and evaluation of models – such as the MMTSB tool set.
Applications of the Tool Set

• Evaluating ensembles of predicted protein conformations to determine the structure with the lowest energy

• Determining peptide conformations ab initio with replica exchange molecular dynamics simulations

• Identifying and predicting the conformation adopted by missing portions of a larger protein structure
Applications: Scoring Models

• A set of structural models is produced from a source outside of the tool set
  – Feig et al. performed an example analysis on the set of structural predictions submitted as entries for a particular target sequence in CASP4
  – In my case, a set of 100-500 suboptimal sequence alignments is created between a target sequence and a template of known structure, and a model is produced from each alignment using the program MODELLER (Šali and Blundell 1993)

• The tool set allows the researcher to evaluate and rank the energies of the models in the ensemble
Procedure: Scoring Models

1. The set of models is converted to an MMTSB ensemble
2. The MMTSB ‘completion’ utility is run on the models to restore missing atoms
   • CHARMM is used to add hydrogens
   • The MMTSB rebuild utility is used to restore side chains if both α- and β-carbons are available
   • SCRWL is used to restore side chains if the model consists only of main-chain carbons
3. The structures are minimized using CHARMM
4. The minimized structures are scored in CHARMM with generalized Born solvation
5. If a known structure exists for comparison, properties such as RMSD can be calculated for each member of the model set
6. Finally, the models’ RMSD’s, energies, and other properties can be extracted for statistical analysis or ranked to identify the best model in the set
Plot of the all-atom energy score computed with a CHARMM22 force field with generalized Born solvation vs alpha-carbon RMSD, provided as example data in Feig et al. 2004. Structures whose energies are greater than -5000 kcal/mol or whose RMSD’s are greater than 20 angstroms have not been plotted.

Although this plot does not reveal strong correlation between energy scores and RMSD, the general trend is that lower energy scores indicate lower RMSD.
Applications: Replica Exchange

- The MMTSB tool set adds replica-exchange functionality to all-atom simulation software such as CHARMM or Amber, and to lattice-based simulation software such as MONSSTER
- The replica-exchange method allows the energy landscape of a given peptide to be sampled accurately and efficiently (Sugita and Okamoto 1999)
  - Simulations at ‘ordinary’ temperatures would produce an accurate energy landscape but waste time and resources ‘trapped’ in local minima
  - Simulations at elevated temperatures avoid the local minima trap but would yield a different energy landscape (extreme case: protein denaturation)
  - Temperature exchanges between parallel replicas, performed according to the standard Metropolis criterion, allow sampling of the low-temperature surface with the advantage of avoiding local minima traps

Metropolis criterion:

\[
p = \begin{cases} 
1 & \text{for } \Delta \leq 0 \\
\exp(-\Delta) & \text{for } \Delta > 0 
\end{cases}
\]

where

\[
\Delta = \left( \frac{1}{kT_i} - \frac{1}{kT_j} \right) (E_j - E_i)
\]
Multiscale Modeling and Replica Exchange

- The tool set provides two standard multiscale modeling protocols coupling all-atom energy evaluation with Monte Carlo moves performed on a low-resolution lattice.
- Both protocols can be coupled to replica exchange sampling with a built-in utility.
- Alternative protocols can be implemented by using the tool set as a programming library.
Replica Exchange Example

A replica exchange simulation of the sequence (AAQAA)₃ of known native helical structure. Black represents temperature; red represents hydrogen bonds between every fourth residue (a proxy measure of helix formation).

Simulation of a segment of the zinc endopeptidase astacin. Black represents temperature; red represents RMSD.
Applications: Refinement

- Homology modeling based on a template of known structure can only go so far. Inevitably, there will be regions of the target sequence that do not align well with the template.

- For these regions, refinement of the model will be necessary. Replica exchange simulations are effective, but it is slow and computationally wasteful to run these simulations on the entire structure when only a small segment needs to be refined.

- The MMTSB tool set offers utilities for selecting small regions of a larger protein model for simulation while forcing the remainder of the structure to remain fixed.

- Layers of residues within a certain cutoff distance of the refinement region can be restrained to varying degrees; both the distance and the restraints can be user-defined. This allows the refinement region to vary within the restraints posed by the surrounding residues in the model.
Refinement and Clustering

• After energy scores have been calculated for the set of conformations generated by the refinement, models can be sorted into clusters on the basis of their RMSD over a specific range of residues.

• MMTSB utilities allow clustering of any ensemble of models and provide an easy means of calculating average properties of each cluster and selecting their most representative members.

Energy score vs RMSD for the astacin beta-hairpin structure. Members of the lowest-scoring cluster are red; the second-best cluster is shown in blue.
• One goal of my project is to ascertain the accuracy with which scoring functions – in particular, statistical potentials – identify the model that best reproduces a protein’s native structure.

• Given a set of proteins of known structure, I am building ensembles of models based on suboptimal sequence alignments. These models are scored using the statistical potential function implemented in Prosa II (Sippl 1993); energy scores are then correlated with RMSD values, fractions of native contacts, and other measures of how closely the models follow the protein’s known structure.

• I am using the MMTSB tool set to manipulate the models as an ensemble – perform minimizations, evaluate energies, and determine structural properties. I will also use the tool set to perform *ab initio* calculations on loops within these models where sequence alignment is insufficiently accurate.
References

Acknowledgements

- Dr. Troy Wymore and Adam Marko
- Pittsburgh Supercomputing Center
- Alex Ismail and Jay Shukla
- Rajan Munshi
- BBSI program and students
- NIH/NSF