Published online 7 May 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W323–W328***
doi:10.1093/nar/gks376

KoBaMIN: a knowledge-based minimization web server for protein structure refinement

João P. G. L. M. Rodrigues[1,2], Michael Levitt[1] and Gaurav Chopra[1,*]

[1]Department of Structural Biology, 299 Campus Dr W, Fairchild Bldg, Room D100, Stanford University, Stanford, CA 94305, USA and [2]Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Received February 17, 2012; Revised April 6, 2012; Accepted April 12, 2012

ABSTRACT

The KoBaMIN web server provides an online interface to a simple, consistent and computationally efficient protein structure refinement protocol based on minimization of a knowledge-based potential of mean force. The server can be used to refine either a single protein structure or an ensemble of proteins starting from their unrefined coordinates in PDB format. The refinement method is particularly fast and accurate due to the underlying knowledge-based potential derived from structures deposited in the PDB; as such, the energy function implicitly includes the effects of solvent and the crystal environment. Our server allows for an optional but recommended step that optimizes stereochemistry using the MESHI software. The KoBaMIN server also allows comparison of the refined structures with a provided reference structure to assess the changes brought about by the refinement protocol. The performance of KoBaMIN has been benchmarked widely on a large set of decoys, all models generated at the seventh worldwide experiments on critical assessment of techniques for protein structure prediction (CASP7) and it was also shown to produce top-ranking predictions in the refinement category at both CASP8 and CASP9, yielding consistently good results across a broad range of model quality values. The web server is fully functional and freely available at http://csb.stanford.edu/kobamin.

INTRODUCTION

Protein structure determines protein function and is thus of great importance for biological sciences (1). However, the difficulty and cost associated with the experimental determination of structures result in a lag between the number of known protein sequences and their corresponding structures (2). To bridge this gap, computational protein structure prediction methods have been developed that aim to generate approximate models of the native state (3), many of which incorporate steps that minimize an energy function. During this energy minimization, the protein structure moves downhill across the energy landscape defined by the force field, bringing it to an end point whose potential energy is lower than before. The chance that such minimization will lead to a structure that is more like the real native structure depends on the accuracy of the underlying force field (4). Furthermore, despite great advances since the early attempts at energy minimization of protein structures (5), crossing the apparently small barrier between near-native models and the native structure, a difference of 1–3 Å C_α root mean square deviation (C_α RMSD), is still a challenging problem (6,7).

Here we describe the KoBaMIN web server, a protein structure refinement server that employs a simple, accurate, consistent and computationally efficient protocol based on a knowledge-based potential energy function (8). The refinement protocol involves a two-step process (8). First, the server uses ENCAD (9) to refine the protein by a highly convergent energy minimization algorithm with an all-atom knowledge-based potential of mean force that implicitly includes the effect of solvent, KB01 (6,7). ENCAD's implementation of the KB01 potential enables rapid refinement of structures (<5 min for a protein of chain length 200 residues), while bringing them closer to the true native conformation. Second, a restrained energy minimization is performed using MESHI (10) to correct side chain rotamer positions and other fine details of the stereochemistry. Finally, the server calculates several comparative values (C_α RMSD, both the GDT-TS and GDT-HA values and the energy of the knowledge-based term of KB01) of the refined structures to either the starting model or to a reference structure if the latter is provided.

The performance of this protocol has been tested for large-scale benchmarks (7,8) and prospectively verified at

*To whom correspondence should be addressed. Tel: +1 650 725 0754; Fax: +1 650 723 8464; Email: gaurav.chopra@stanford.edu critical assessment of techniques for protein structure prediction (CASP) blind experiments (11,12) using the group name KnowMIN, where it showed a consistent improvement of the quality of the models. The server does not require registration and thus provides open access to a powerful refinement method, otherwise complex to install and maintain. This allows the KoBaMIN server to be used for models coming from state-of-the-art protein structure prediction methods, or else to act as the end step in a homology modelling pipeline.

MATERIALS AND METHODS

User input

The input to the KoBaMIN server is single or multiple protein structures in PDB format using the web interface shown in Figure 1. A single protein structure can also be conveniently pasted in the text area or the file can be uploaded to the server. For refinement of multiple protein structures, a compressed archive is uploaded to the server, which supports several common compression formats (zip, tar.gz and tar.bz2). The KoBaMIN server also supports insertion codes in the PDB files. When the user submits a PDB file with insertion codes, refinement is done on the model with the first insertion code for each residue. The user can also upload a reference structure (optional) to compare with the results from refinement. The user is not required to provide an email address or a job name but this is recommended for enhanced user experience.

The KoBaMIN server workflow

After several initial data validation checks, our server runs a two-step refinement protocol (8) before final quality assessment is done for the refined structure (Figure 1). This is explained in more detail in what follows.

Structures submitted by the user undergo a two-step validation process. The first step checks for file formats and conflicting input options necessary as the server allows more than one structure submission method. If the input data is valid, it is stored on the server and passed over to the processing machine to be added to the queuing system. If the user provided an e-mail address during submission, a notification email is also sent. Alternatively, the user can bookmark the link presented in the status page to facilitate the retrieval of results at a later time. A second more stringent validation step checks for structural inaccuracies that might hinder the refinement protocol, namely, missing atoms, missing residues, handling of multiple chains and errors in the PDB format and using the Bio.PDB module of the Biopython tools (13,14). Only after successfully passing both steps of validation is a structure considered for refinement.

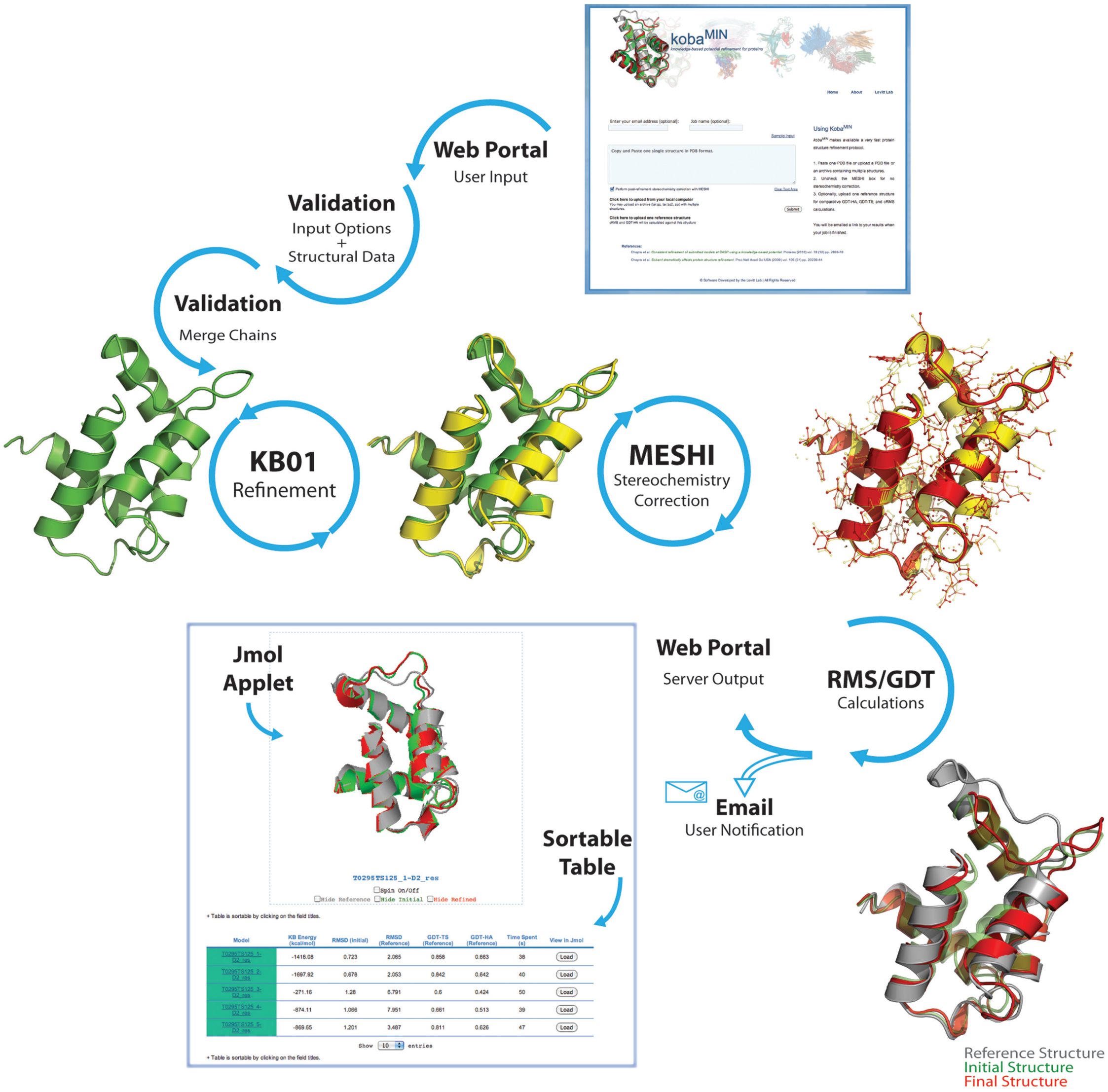
The refinement protocol works using a custom queuing system designed to maximize speed for all users using the server at any given time. One queue slot is allotted per processing core of the machine (currently on eight cores), while one core is permanently reserved for MESHI stereochemistry correction calculations due to its high memory requirements. This is a two-step process (Figure 2) (8). Each structure coming from the validation step above is first submitted to a highly convergent limited memory Broyden–Fletcher–Goldfarb–Shannon (L-BFGS) (15) energy minimization procedure implemented in the ENCAD software (9) using the KB01 potential (8). The minimization runs for 200 000 steps or until convergence to machine precision (mean gradient <1e-7 kcal/(mol Å))

Since KB01 sometimes spoils the local bonded stereochemistry of the refined structures, an additional stereochemistry correction step is applied. This step is optional, but recommended; it is set as default at the time of submission unless unchecked by the user (Figure 1). It uses the MESHI software library (10) to first identify and correct fragments with bad torsion angles. Such fragments are replaced by fragments of corresponding lengths from native structures that best fit the C_α and C_β atoms of the starting fragments. Then 20 000 steps of restrained energy minimization are performed with the MESHI potential (8); this causes minimal changes to the protein coordinates to correct local stereochemistry while retaining the improvements caused by the initial pass of KB01 minimization (Figure 2).

After structure refinement, the server runs calculation scripts to provide the user with a quantitative analysis to compare refined structure with the corresponding reference and initial structures. If the optional reference structure is not provided, the initial structure is used for comparative analysis. A comprehensive measure of the extent of structural changes caused by the refinement process involves comparison using C_α RMSD and two Global Distance Test scores (16), GDT-TS and GDT-HA. Analysis of intermediate stages of the refinement process is not performed due to the previous observation that the stereochemistry correction step does not significantly change the improvements in the refined structure by KB01 (8). The log files of the ENCAD/KB01 refinement job are also parsed to provide the user with the energy difference before and after that step, measured only by the non-bonded term of the hybrid knowledge-based force field (Figure 2). If an email is provided, the user is notified about job completion and web links are provided to analyse the results. Otherwise, the user can visit the bookmarked link provided on the status page when the job was queued.

Server output

The results of the refinement protocol are presented to the user via a web page that contains a Jmol interactive molecular viewer applet (17) and a summary table showing the KB01 energy, C_α RMSD, GDT-TS and GDT-HA and comparison of the refined structure to the initial and reference (if provided) structures (Figure 3). The initial (green cartoon), refined (red cartoon) and reference (grey cartoon) structures can be viewed interactively in the Jmol applet by loading different structures for a multiple structures job. Both refined and initial structures are aligned with the reference structure; if the reference is not provided, the refined structure is aligned with the initial structure for refinement analysis. The links for downloading the coordinate files of the refined structures are provided in the same table.



koba^{MIN}
knowledge-based potential refinement for proteins

Home About Levitt Lab

Enter your email address [optional]: Job name [optional]:

Sample Input

Copy and Paste one single structure in PDB format.

☑ Perform post-refinement stereochemistry correction with MESHI

Clear Text Area

Click here to upload from your local computer
You may upload an archive (tar.gz, tar.bz2, zip) with multiple structures.

Submit

Click here to upload one reference structure
cRMS and GDT-HA will be calculated against this structure

Using Koba^{MIN}

koba^{MIN} makes available a very fast protein structure refinement protocol.

1. Paste one PDB file or upload a PDB file or an archive containing multiple structures.
2. Uncheck the MESHI box for no stereochemistry correction.
3. Optionally, upload one reference structure for comparative GDT-HA, GDT-TS, and cRMS calculations.

You will be emailed a link to your results when your job is finished.

References:
Chopra et al. *Consistent refinement of submitted models at CASP using a knowledge-based potential.* Proteins (2010) vol. 78 (12) pp. 2668-78
Chopra et al. *Solvent dramatically affects protein structure refinement.* Proc Natl Acad Sci USA (2008) vol. 105 (51) pp. 20239-44

© Software Developed by the Levitt Lab | All Rights Reserved

T0295TS125_1-D2_res
□Spin On/Off
□Hide Reference □Hide Initial □Hide Refined

+ Table is sortable by clicking on the field titles.

Model	KB Energy (kcal/mol)	RMSD (Initial)	RMSD (Reference)	GDT-TS (Reference)	GDT-HA (Reference)	Time Spent (s)	View in Jmol
T0295TS125_1-D2_res	-1418.08	0.723	2.065	0.858	0.663	38	Load
T0295TS125_2-D2_res	-1697.92	0.678	2.053	0.842	0.642	40	Load
T0295TS125_3-D2_res	-271.16	1.28	6.791	0.6	0.424	50	Load
T0295TS125_4-D2_res	-874.11	1.066	7.951	0.661	0.513	39	Load
T0295TS125_5-D2_res	-869.65	1.201	3.487	0.811	0.626	47	Load

Show 10 entries

+ Table is sortable by clicking on the field titles.

Figure 1. KoBaMIN server workflow. The only required field for job submission to the web page of KoBaMIN is an initial structure (cartoon in green) in PDB format, which can be pasted in the designated text area, or uploaded through the web form. Multiple structures can only be submitted for refinement using the upload option. The structures undergo a two-step validation process and various exceptions are built into the web server for proper error handling. If the user specifies an email address during submission (not required but recommended for enhanced user experience), a notification is sent when the job is queued; a similar notification is also sent after job completion. Once the initial structure is validated, it undergoes a two-step structural refinement step (also described in Figure 2) resulting in a refined structure (cartoon in red). It is optional to provide a reference structure (cartoon in grey) for structural comparison after GDT/RMSD calculations. If the reference is not provided, the initial structure is used for comparative calculations with the refined structure. A table of knowledge-based energy (kcal/mol), C_α RMSD (Å), GDT-TS, GDT-HA and time spent in computation (seconds) is generated for the refined structure with respect to the reference (or initial) structure in a tabular format. A Jmol applet is used for structural visualization. All these results are shown in a web page that the user can bookmark for future viewing.

Several downloadable links are also provided on the results web page so that the user can download an archive with only the refined structures as well as an archive containing the complete run data. All these links, including the results web page, are also sent in an e-mail after job completion. The complete run archive contains all structures, including initial structures submitted by the user and the intermediates of the two-step refinement protocol. It also contains ENCAD and MESHI log files for all structures for a more detailed inspection of the refinement run, as well as a general KoBaMIN server log file. Any structures that failed structural validation or the refinement protocol are placed in a specific folder for easy identification. The details of the directory structure for the output as well as the explanation of the entire server workflow is given in the 'About' page of the web server (http://csb.stanford.edu/kobamin/about.html). The results are stored on the server for 1 month, after which they will be deleted.

Web server configuration

The KoBaMIN protein refinement server does not require registration and is freely available at

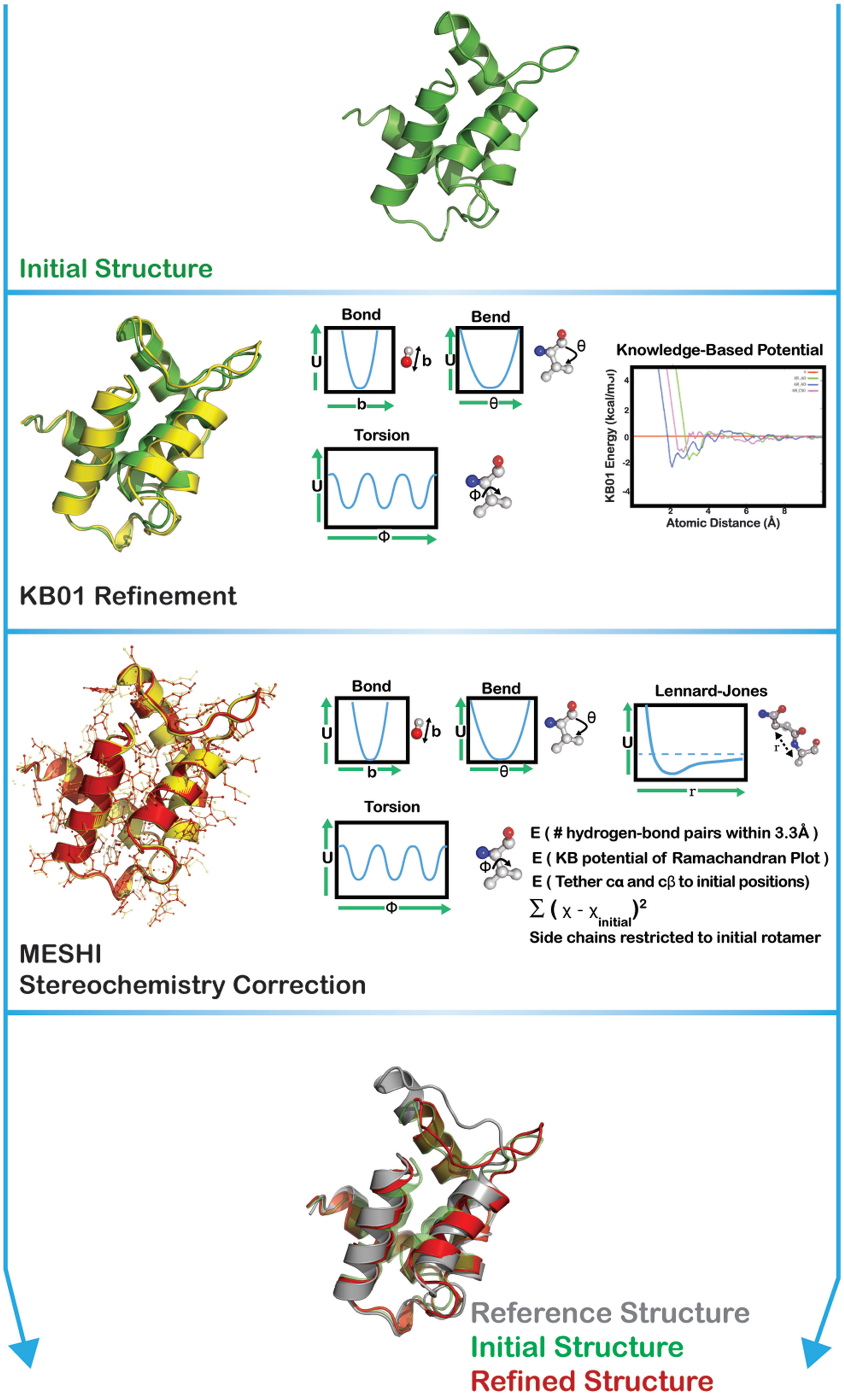
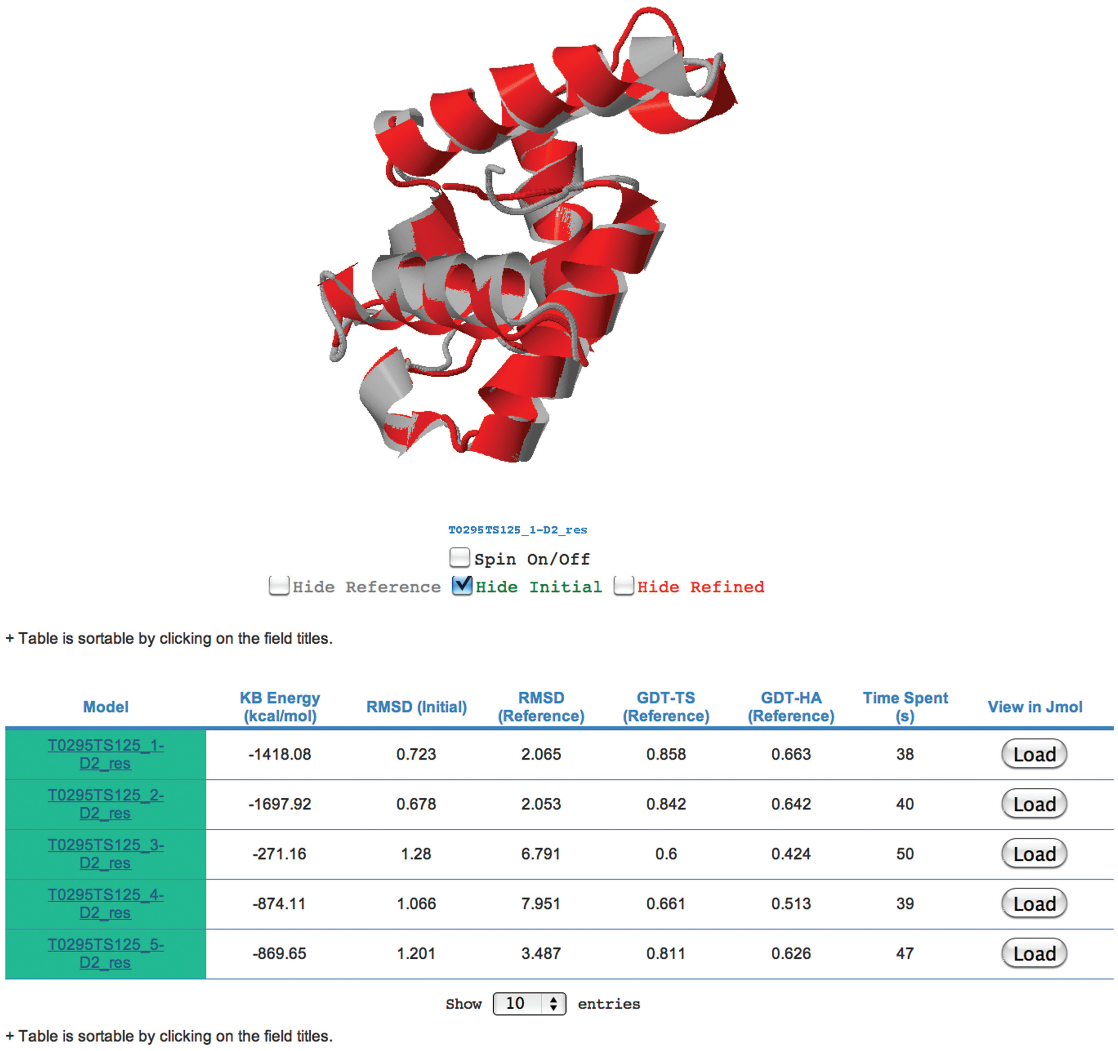

Initial Structure
Bond
U
b
Bend
θ
Torsion
Φ
Knowledge-Based Potential
KB01 Energy (kcal/mol)
4
2
0
-2
-4
2
4
6
8
Atomic Distance (Å)
KB01 Refinement
Bond
Bend
Lennard-Jones
r
Torsion
E (# hydrogen-bond pairs within 3.3Å)
E (KB potential of Ramachandran Plot)
E (Tether cα and cβ to initial positions)
$\sum (x - x_{initial})^2$
Side chains restricted to initial rotamer
MESHI
Stereochemistry Correction
Reference Structure
Initial Structure
Refined Structure

Figure 2. KoBaMIN structure refinement protocol. After successful structure validation, the initial structures are first submitted to a powerfully convergent energy minimization using the hybrid knowledge-based potential (KB01), which includes ENCAD bonded terms and a statistically derived continuously differentiable potential of mean force for non-bonded terms. Side chain optimization is done by fragment replacement and the final stereochemistry correction step is performed using MESHI minimization, which included bonded terms and a combination of van der Waals, hydrogen bond, knowledge-based potential for Ramachandran plot, a tethering term for C_α and C_β atoms to remain close to the initial position as well as side chain restraints to initial rotamers. Validation checks occur before and after each refinement step, and calculations to inform the user of the extent of the refinement process are performed at the end of the stereochemistry correction step.

http://csb.stanford.edu/kobamin as a community service. It runs on a dedicated Linux machine, a 2.8 GHz Intel i7 processor with eight cores and 12 GB of RAM. The web application uses the Python programming language (http://www.python.org) to serve web pages, parse user data and to bind all separate steps together into an integrated workflow. The web server runs on the Apache HTTP server version 2.0.51. The user interface is designed



T0295TS125_1-D2_res

□ Spin On/Off

□ Hide Reference ☑ Hide Initial □ Hide Refined

+ Table is sortable by clicking on the field titles.

Model	KB Energy (kcal/mol)	RMSD (Initial)	RMSD (Reference)	GDT-TS (Reference)	GDT-HA (Reference)	Time Spent (s)	View in Jmol
T0295TS125_1-D2_res	-1418.08	0.723	2.065	0.858	0.663	38	Load
T0295TS125_2-D2_res	-1697.92	0.678	2.053	0.842	0.642	40	Load
T0295TS125_3-D2_res	-271.16	1.28	6.791	0.6	0.424	50	Load
T0295TS125_4-D2_res	-874.11	1.066	7.951	0.661	0.513	39	Load
T0295TS125_5-D2_res	-869.65	1.201	3.487	0.811	0.626	47	Load

Show 10 entries

+ Table is sortable by clicking on the field titles.

Figure 3. KoBaMIN web server output. The user is presented with a table reflecting the extent of the refinement process (C_α RMSD, GDT-TS and GDT-HA of refined structure with respect to the reference structure) and the change in energy (reported only by the non-bonded knowledge-based term of ENCAD force field), as well as the time (in seconds) spent to refine each individual model. Click on model names in the table to download each structure. The table fields can be sorted in ascending or descending order by clicking on them. Each model is viewed in Jmol applet by loading it. The Jmol applet shows all three structures: refined (red cartoon), initial (green cartoon) and reference (grey cartoon). The user can hide or show different structures; here the initial structure is hidden and only reference and refined models are shown in the Jmol applet. The applet is interactive, in that, the user can change the view, style or check 'Spin On/Off' to rotate the molecules continuously. The results web page also contains links to download an archive of the final refined structures or a complete run data archive with log files (links not shown in this picture). An example results page can be viewed at http://csb.stanford.edu/kobamin/results/example/.

using JavaScript libraries jQuery and DataTables, and the Jmol interactive molecular viewer applet is used for structural visualization.

Performance of the web server

The KoBaMIN refinement protocol has been extensively tested on large-scale benchmarks. The protocol is based on direct energy minimization with that potential and thus requires a few minutes (typically <5 min) of CPU time for proteins of varying chain length (see Figure S4 in publication) (8). It was first tested on a set of 75 native non-redundant proteins and 729 near-native decoys for each native protein. The KB01 refinement was more consistent than all other potentials (7). Our second benchmark was done on more realistic set of models by applying it to 36 802 models from 178 groups in CASP7 as well as 21 refinement targets from the refinement categories of both CASP7 (nine models) and CASP8 (12 models). The CASP benchmark is highly diverse in the difficulty of the prediction targets, the variety the prediction methods used and the success of participants. Our protocol shows consistent average improvement of all models across prediction groups (including the top ranking groups) and target difficulty (8).

The KoBaMIN refinement was also prospectively verified to produce top-ranking predictions in the refinement category at both CASP8 (11) and CASP9 (12), yielding consistently good results. In the recent community-wide CASP9 experiment, KoBaMIN (the KnowMIN group) produced the best average GDT-HA over all refinement targets, improving 13 out of 14 refinement targets (12). The KoBaMIN server (http://csb.stanford.edu/kobamin/) has been up and running since May 2010, and has successfully generated thousands of refined models for the biologists worldwide.

CONCLUSIONS

The KoBaMIN protein refinement server makes available a fast and consistent protocol for near-native protein structure refinement across a broad range of model quality values. The improvement in accuracy of the refined structure is chiefly caused by KB01 minimization and not by the MESHI stereochemistry correction step (8). The KB01 pairwise potential encodes information about favoured pairwise interatomic distances (see local minima for knowledge-based potential in Figure 2), which seems to be essential for structural improvement. The KoBaMIN server can be most useful to models coming from state-of-the-art protein structure prediction methods, or as an end step in homology and comparative modelling pipelines that contain a high percentage of helical and coil residues (8). The server does not require registration and thus provides open access to a powerful refinement method, otherwise complex to install and maintain. The low computational cost and high accuracy of the KoBaMIN protocol will allow this consistent refinement method to be run on a genome scale.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Nir Kalisman for useful insights and help with the MESHI software.

FUNDING

Funding for open access charge: National Institutes of Health award [GM063817 to M.L.]; who is the Robert W. and Vivian K. Cahill Professor of Cancer Research.

Conflict of interest statement. None declared.

REFERENCES

1. Anfinsen,C. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
2. Levitt,M. (2009) Nature of the protein universe. *Proc. Natl Acad. Sci. USA*, **106**, 11079.
3. Zhang,Y. (2008) Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.*, **18**, 342–348.
4. Misura,K. and Baker,D. (2005) Progress and challenges in high-resolution refinement of protein structure models. *Proteins*, **59**, 15–29.
5. Levitt,M. and Lifson,S. (1969) Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.*, **46**, 269–279.
6. Summa,C. and Levitt,M. (2007) Near-native structure refinement using *in vacuo* energy minimization. *Proc. Natl Acad. Sci. USA*, **104**, 3177.
7. Chopra,G., Summa,C.M. and Levitt,M. (2008) Solvent dramatically affects protein structure refinement. *Proc. Natl Acad. Sci. USA*, **105**, 20239–20244.
8. Chopra,G., Kalisman,N. and Levitt,M. (2010) Consistent refinement of submitted models at CASP using a knowledge-based potential. *Proteins*, **78**, 2668–2678.
9. Levitt,M., Hirshberg,M., Sharon,R. and Daggett,V. (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comput. Phys. Commun.*, **91**, 215–231.
10. Kalisman,N., Levi,A., Maximova,T., Reshef,D., Zafriri-Lynn,S., Gleyzer,Y. and Keasar,C. (2005) MESHI: a new library of Java classes for molecular modeling. *Bioinformatics*, **21**, 3931–3932.
11. MacCallum,J.L., Hua,L., Schnieders,M.J., Pande,V.S., Jacobson,M.P. and Dill,K.A. (2009) Assessment of the protein-structure refinement category in CASP8. *Proteins*, **77(Suppl. 9)**, 66–80.
12. MacCallum,J.L., Pérez,A., Schnieders,M.J., Hua,L., Jacobson,M.P. and Dill,K.A. (2011) Assessment of protein structure refinement in CASP9. *Proteins*, **79**, 74–90.
13. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al*. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
14. Hamelryck,T. and Manderick,B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.
15. Dong,C. and Nocedal,J. (1989) On the limited memory BFGS method for large scale optimization. *Math. Program.*, **45**, 503.
16. Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.
17. Jmol: *An Open-Source Java Viewer for Chemical Structures in 3D*, http://www.jmol.org (April 2012, date last accessed).