Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W549–W555***
doi:10.1093/nar/gkm266

MODPROPEP: a program for knowledge-based modeling of protein–peptide complexes

Narendra Kumar and Debasisa Mohanty*

National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi 110067, India

Received January 31, 2007; Revised March 28, 2007; Accepted April 8, 2007

ABSTRACT

MODPROPEP is a web server for knowledge-based modeling of protein–peptide complexes, specifically peptides in complex with major histocompatibility complex (MHC) proteins and kinases. The available crystal structures of protein–peptide complexes in PDB are used as templates for modeling peptides of desired sequence in the substrate-binding pocket of MHCs or protein kinases. The substrate peptides are modeled using the same backbone conformation as in the template and the side-chain conformations are obtained by the program SCWRL. MODPROPEP provides a number of user-friendly interfaces for visualizing the structure of the modeled protein–peptide complexes and analyzing the contacts made by the modeled peptide ligand in the substrate-binding pocket of the MHC or protein kinase. Analysis of these specific inter-molecular contacts is crucial for understanding structural basis of the substrate specificity of these two protein families. This software also provides appropriate interfaces for identifying, putative MHC-binding peptides in the sequence of an antigen or phosphorylation sites on the substrate protein of a kinase, by scoring these inter-molecular contacts using residue-based statistical pair potentials. MODPROPEP would complement various available sequence-based programs (SYFPEITHI, SCANSITE, etc.) for predicting substrates of MHCs and protein kinases. The program is available at http://www.nii.res.in/modpropep.html

INTRODUCTION

Proteins involved in a majority of cellular processes usually perform their function by binding to some target proteins and forming protein–protein complexes. Interactions between two or more proteins often occur over short contiguous stretches of amino acids within one protein. For example, recognition of substrate proteins by various protein kinases during cell signaling events is governed primarily by specific interactions between the kinase and a contiguous peptide stretch containing the phosphorylation site. Several receptors have peptide fragments as ligands e.g. the major histocompatibility complex (MHC) (1). Thus understanding molecular details of interactions between proteins and short peptide motifs is essential for dissecting underlying mechanism of several major cellular processes. Among the various proteins which interact specifically with short peptide motifs, protein kinases and MHCs represent two major protein families whose substrate specificities have been extensively studied by various experimental approaches (2–4).

Although a number of computational tools such as NetPhosK (5), KinasePhos (6), GPS (7), Scansite (8), SYFPEITHI (9), ProPred (10), etc. are available for predicting the putative substrate peptides for protein kinases and MHC proteins, these methods are mostly based on available experimental binding data for a given class of protein kinase or MHC. These tools predict substrate peptides based on identification of the conserved motifs in a set of known peptide substrates and do not use information from the three dimensional structure of the protein–peptide complex. Hence, these sequence-based prediction tools do not give information about key residues in kinases and MHCs which control substrate specificity. Information about specificity determining residues (SDR) can help in design of novel peptide ligands. Correct identification of SDRs of a given protein kinase or MHC can help in prediction of substrates for those protein kinases or MHCs for which no peptide-binding data is available, as demonstrated successfully in structure-based substrate prediction methods like PREDIKIN (11) and PREDEP (12). These studies have demonstrated that structural analysis of interactions in protein–peptide complexes can lead to novel insight into the mode of substrate recognition. Therefore, molecular modeling of peptide–MHC and peptide–kinase interactions have been carried out by several groups using *ab initio* docking (13) or MD simulation approach (14). However, the compute

*To whom correspondence should be addressed. Tel: +91 11 26703749; Fax: +91 11 26162125; Email: deb@nii.res.in

intensive nature of these calculations has limited such studies to few protein–peptide complexes. Since knowledge-based methods are less compute intensive, and have better prediction accuracy, development of suitable knowledge-based tools for modeling protein–peptide complexes would permit quick structural analysis of MHCs and protein kinases with their substrate peptides. A knowledge-based approach has been used recently for developing kinDOCK (15), a powerful tool for modeling of ATP analogs into the active site pocket of protein kinases. However, no such user-friendly tool is presently available for knowledge-based modeling of peptides in the binding pockets of MHCs or protein kinases.

Therefore, we have developed MODPROPEP, a web server for structural modeling of peptides of any desired sequences in the active site pockets of kinases/MHCs having known crystal structures or homology models of kinases/MHCs. In this manuscript, we give a brief description of the development of MODPROPEP, various assumptions made in the knowledge-based modeling protocol, various features of MODPROPREP and few examples of its use.

METHODS

Compilation of crystal structures

The available crystal structures of MHC and protein kinases were downloaded from PDB website at http://www.rcsb.org (16). The structures were divided into two groups, i.e. structures in complex with substrate peptide ligand and structures without the bound peptide ligand. These crystal structures were manually examined and chain/residue numbering was appropriately edited if necessary. All the crystal structures were categorized into three major classes, i.e. class I MHC, class II MHC and protein kinases. Each of these three classes was further grouped into various functional families of protein kinases or MHC alleles.

Detailed analysis of these crystal structures indicated, that all the protein kinases shared a conserved structural fold despite their sequence divergence. For example, crystal structures of IR and PHK, which share a sequence identity of only 40% can be superposed with a C^{α} RMSD of 1.6 Å. Similar conservation of structures was also observed both for class I and class II MHC structures which share a higher degree of sequence identity within themselves. BLAST alignment of large number of protein kinases and MHC proteins available in sequence databases with these crystal structures indicated that, homology models can be obtained for most of these sequences with reasonable accuracy. Comparison of the bound peptide structures indicated that in all these three classes of proteins, the substrate peptides bind at a structurally homologous site on the conserved fold and the bound peptides maintain a more or less similar extended conformation. This suggested the possibility that bound peptides from peptide–protein complexes can be transformed to the protein structures lacking the bound peptide based on optimum superposition of the protein structures. It may be noted that similar assumption has been used successfully in structural modeling studies of protein–ligand complexes involving protein kinases (11), MHCs (12,17) and other enzyme families (18,19). There are several examples where more than one crystal structure of an allele is found with bound peptides of different length. It is generally assumed that, three residues on each side of the phosphorylation site make significant contact with the protein kinase and are responsible for the specificity of a kinase (11). Therefore, bound peptides having more than seven amino acids were truncated to three amino acids on either side of the phosphorylation site. All these structures were stored in the template library of MODPROPEP.

Modeling of protein–peptide complexes

The current template library of MODPROPEP has protein–peptide complex crystal structures for 16 alleles of class I, 12 alleles of class II MHC proteins and six different protein kinase families. Figure 1 shows a flowchart depicting various tasks which can be performed using MODPROPEP. For these MHC alleles and protein kinase families, substrate peptide of any desired sequence can be modeled. Modeling of peptide in the binding pocket of MHC or protein kinase is carried out by using the same backbone conformation as in the template complex and the side-chain conformations are generated by the program SCWRL (20), which uses a backbone-dependent rotamer library approach. The template library of MODPROPEP has structures for many MHC alleles or kinases families without the bound peptide substrate. For modeling of peptide substrates in complex with any of these MHC alleles or kinase families, peptide conformations are transformed from the available crystal structures of the protein–peptide complexes after optimum superposition of the proteins. If no crystal structures are available for a given protein kinase or MHC protein, the program can model its structure in complex with peptides of desired sequence using the crystal structure of the closest homologous protein–peptide complexes. Sequences of various MHC alleles have been obtained from the IMGT/HLA database (21) and stored locally so that the user can select from the list of alleles the protein to be modeled. The crystal structure having maximum sequence similarity is used as a template for modeling the structure of query allele. All sequence alignments are carried out using a local version of the program BLAST. The SCWRL program is used for mutating the residues as per the BLAST alignment and generate the desired homology model. Since only protein–peptide complexes are used to generate the homology models, the backbone of the bound peptide is appropriately mutated by SCWRL to model the substrate of desired sequence. Thus, MODPROPREP provides options for modeling peptide of any desired sequence in complex with any MHC protein or protein kinase.

In order to analyze the interactions between the peptide and the protein, the residues of the MHC or the kinase, which are in contact with different side chains of the modeled peptide, are identified using a distance-based cut off. Based on these contact residues, putative binding

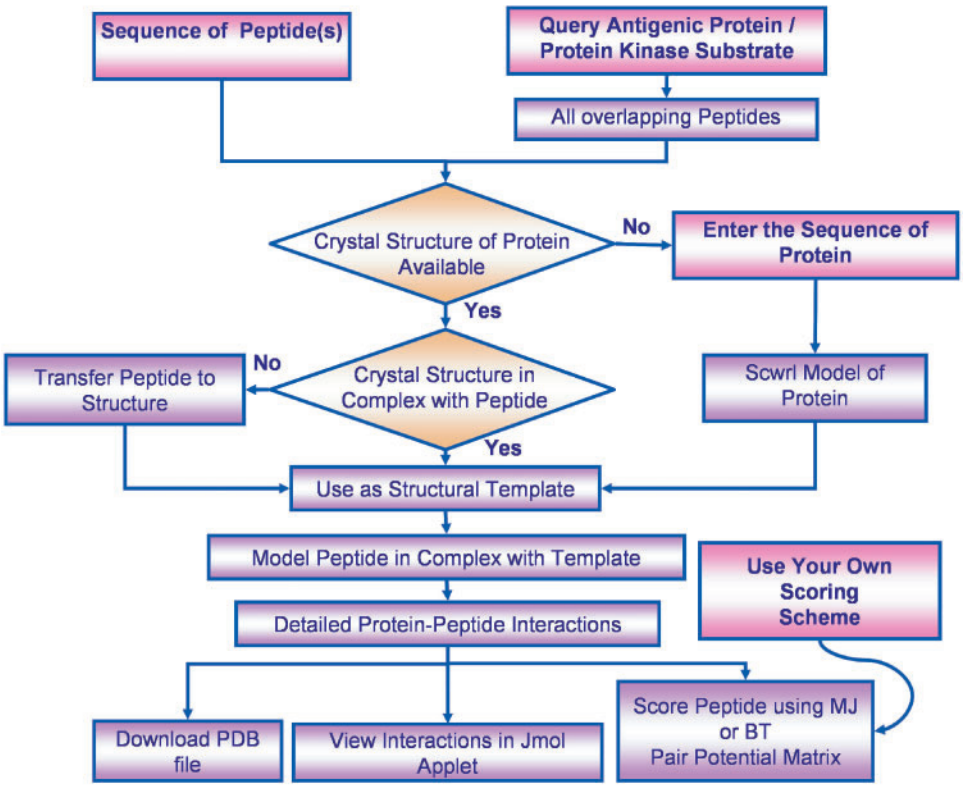

Figure 1. A flowchart depicting the organization and features of MODPROPEP. Pink boxes represent the information provided by the user as input.

pockets are defined for each of the residues in the peptide. MODPROPEP provides a user-friendly Jmol java applet interface (http://jmol.sourceforge.net) for visualizing the modeled complexes and analyzing the binding pockets in detail.

Apart from structural modeling of the peptide of a given sequence in the substrate-binding pocket of MHC protein or protein kinase, MODPROPEP is also capable of scanning an antigenic protein for potential MHC-binding peptides. Similarly putative substrate proteins for various protein kinases can be scanned for potential phosphorylation sites. Scanning of input sequence is done by breaking the protein sequence into all possible overlapping peptides of a given length. This length is usually the length of the bound peptide present in the template protein–peptide complex, i.e. 9 or 10 mer for class I MHC and longer peptides for class II MHC. However, for protein kinases only heptameric peptides containing Ser/Thr/Tyr as central residue are chosen. For each of these peptides, instead of building all atom side-chain conformations, as a first step, contacting residue pairs between peptide and the protein are identified based on C^{β}–C^{β} distances. The binding score of these peptides with the MHC or kinase is evaluated using residue-based statistical energy function by Miyazawa and Jernigan (MJ) (22). It may be noted that a similar scoring scheme has been used earlier for identifying MHC-binding peptides using a threading approach (12). Apart from MJ statistical potential, the program also has options for ranking peptide-binding affinities using residue-based statistical energy function by Betancourt and Thirumalai (BT) (23) or other user-defined residue-based schemes. The peptides are sorted according to their binding score and the user can select some or all of these peptides for detailed side-chain modeling by SCWRL depending on their preliminary scores.

Query interface

Currently, the structural library of program contains crystal structures of class I MHC, class II MHC and protein kinases. The modeling of protein–peptide complexes involving these three classes of protein is possible. User can access the features involving each class by clicking the links on the horizontal bar just below the header graphics.

The program requires user to select a MHC allele or protein kinase from the pull-down menu. The program automatically shows the peptide length options available for modeling for that MHC allele or protein kinase. Program takes the user to available crystal structure templates for the selected protein and peptide length. From here the user can decide a task, which is either modeling of peptides or scanning a protein sequence for favorable binders. The user is prompted to enter the sequence of peptides as one letter code of amino acids.

The program models the peptides in complex with the selected protein that are available for download as files in PDB format. If no ligand bound structure is available for the selected protein, the peptide is modeled by transferring the ligand peptide coordinates from a homologous protein–peptide complex. Figure 2 shows an example where a peptide has been modeled in complex with the kinase GSK3-beta by transferring the coordinates from CDK2. In order to test the accuracy of this ligand transformation approach, we modeled a peptide in complex with PKB by transforming the bound peptide from PKA. The tutorial section of MODPROPEP shows the superposition of the modeled and the experimentally determined bound peptide in the active site of PKB. As can be seen, backbone of both the peptides superpose quite well with an RMSD of 1.3 Å.

MODPROPEP provides a user-friendly interface to analyze each modeled peptide in detail for contact with the protein. Inter-residue contacts can be calculated either based on the distance between C^{β} atoms or based on the distance between any two atoms in a pair of residues. A list of neighboring residues in the protein is displayed for each residue in the peptide. These amino acids on the protein define the binding subsite for each of the peptide residues. Residue pairs having steric clashes are highlighted in yellow. The program also provides interface for analyzing detailed atomic contacts between each pair of residue. Additionally, MODPROPEP uses Jmol applet for the rapid visualization of these subsites in the proteins. Mouse click on a peptide residue shows that residue and the neighboring residues in the protein in Jmol applet on right-hand side. Clicked peptide residue is depicted in ball and stick, while the neighboring residues are shown in CPK. The protein backbone is shown in ribbon while the peptide backbone is shown in the sticks.

As mentioned earlier, the current version of MODPROPEP permits scoring various bound peptides using residue-based statistical scoring matrices given by MJ and BT. Both these scoring matrices have been used in the literature for evaluating binding energy of protein–peptide complexes. It has been reported that, while MJ potential gives better results for binding of peptides involving hydrophobic interfaces, BT potential is more appropriate for binding of peptides involving

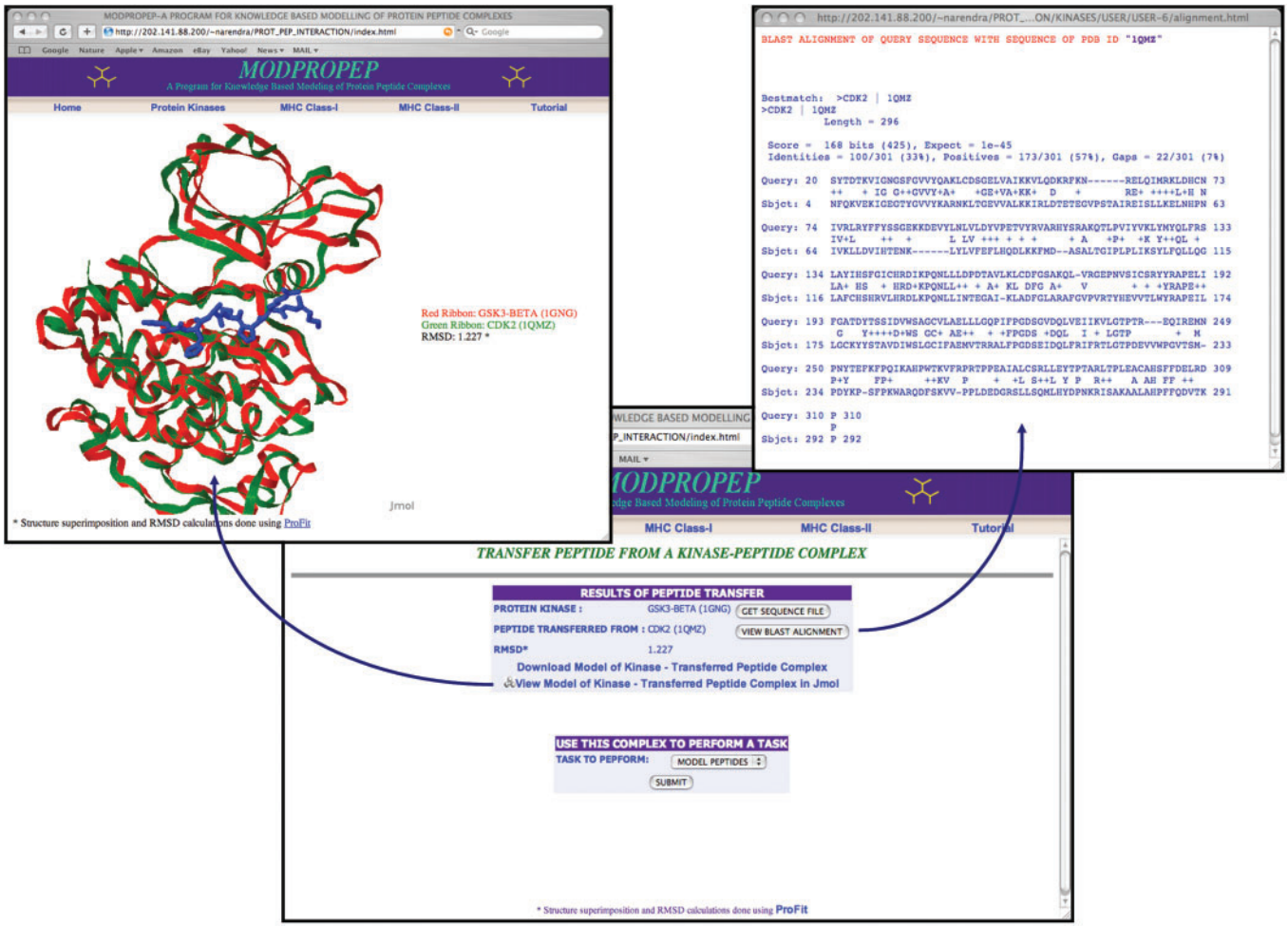

Figure 2. A snapshot from MODPROPEP showing the result of transfer of bound peptide from the kinase CDK2 (template:1QMZ) to GSK3-beta (template:1GNG). Links are provided for downloading pdb coordinates of the modeled complex and viewing the superposition of the two protein structures along with the peptide in the Jmol applet. A pop-up window shows the BLAST alignment between CDK2 and GSK3-beta.

polar contacts. Here, we discuss a typical example of ranking the site of phosphorylation on the beta-adducin protein (accession no: P35612) by protein kinase A (24). Out of a total of 118 S/T containing heptamers, RTPSFLK containing the experimentally identified phosphorylation site S713, is ranked 8 by MJ potential, while scoring by BT matrix gives it a rank of 3. Modeling of this peptide in complex with PKA shows R710 is stabilized by contacts with E127 and E170. The screenshots for this example are available at the tutorial page of MODPROPEP. Prediction of phosphorylation site in *Limulus* myosin III by PKA (25), and PS1 by GSK3-beta (26) indicates that, the true phosphorylation sites identified in recent experiments are ranked as high-scoring peptides by MODPROPEP using BT matrix. Figure 3 shows the ranking of a recently identified class I MHC allele HLA-A*0201 ligand by MODPROPEP (27). As can be seen, out of a total of 625 nonameric peptides present in the antigen CABL1_HUMAN (accession no: Q8TDN4), VALEFALHL has a rank of 13 and 26 by MJ and BT potentials, respectively. Analysis of inter-molecular contacts indicates that, this peptide is stabilized by interactions involving K66, A150, V152, Y159 and W167. We have also tested the predictive ability of MODPROPEP using all the known substrates of PKA cataloged in phospho.ELM (28). Our results indicate that, in 76% of cases the true phosphorylation site can be ranked within top 30% using BT matrix. Similar benchmarking on 90 class I MHC–peptide complexes shows that, MODPROPEP can rank the true binder within top 30% in 61% of cases.

Implementation of the web server

MODPROPEP has been implemented using Perl, CGI scripts, java scripts, Jmol applet and apache web server. BLAST program downloaded from NCBI website is used for local alignments. SCWRL3 is used for the side-chain modeling. Various structural superpositions have been carried out using the program ProFit (http://www.bioinf.org.uk/software/profit).

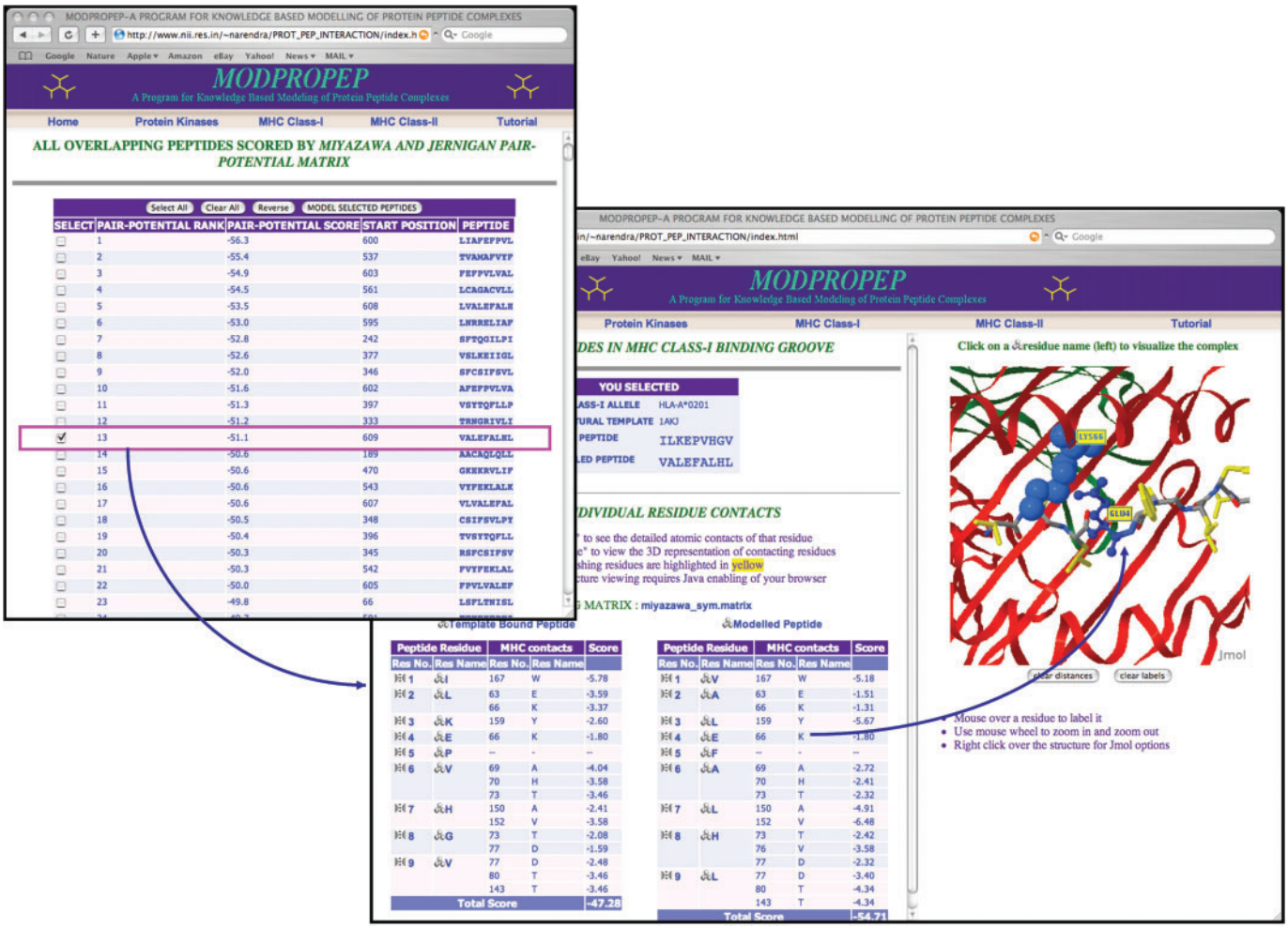

Figure 3. A snapshot from MODPROPEP showing the result of scanning of CABL1_HUMAN protein for HLA-A*0201 restricted antigenic peptides. The experimentally identified substrate peptide VALEFALHL, is chosen for modeling in complex with HLA-A*0201 using 1AKJ as template. The residues of HLA-A*0201 in contact with the peptide residues are depicted in tabular format. The right-hand side frame shows the 3D structure of a selected peptide residue and its contacts with HLA-A*0201.

DISCUSSION

MODPROPEP is a web server for knowledge-based modeling of peptide ligands in the active site of various MHCs and protein kinases. The software uses available crystal structures as templates and uses the program SCWRL to mutate the sequence of the protein as well as the peptide to model any peptide–MHC or peptide–kinase complex. It provides a number of user-friendly interfaces for visualization and analysis of binding pockets in these protein–peptide complexes. This software has been developed based on the assumption that MHCs and protein kinases have conserved structural fold and the ligand peptides bind essentially at the same site. A major advantage of MODPROPEP over other structural modeling programs is that, it can be used to quickly model a large number of peptides in the binding pockets of MHCs and protein kinases. Thus MODPROPEP will complement various available sequence-based programs for predicting peptide ligands for MHCs and protein kinases. Using this software the user can identify amino acids on the MHC or kinases, which are crucial for selection of a peptide ligand. Such information is important for design of novel peptide ligands or assigning specificities to new alleles of MHCs or novel families of kinases. This software also has an option for searching the MHC-binding peptides in the sequence of an antigen or phosphorylation sites on the substrate protein of a protein kinase using structure-based approach. Presently, the binding energy is being accessed using residue-based statistical potential. This scoring function is appropriate for quick preliminary ranking of putative peptide ligands. High-ranking peptides need to be modeled and detailed interactions with the proteins should be analyzed for prediction of actual binders.

ACKNOWLEDGEMENTS

Authors thank Director, NII for his encouragement and support. N.K. is a recipient of Senior Research Fellowship from CSIR, India. The work has been supported by core grants to National Institute of Immunology and project grants to D.M. from Department of Biotechnology, Government of India. Computational resources provided under BTIS project of DBT, India are gratefully acknowledged. The authors are thankful to Prof. R.L. Dunbrack Jr for allowing the use of SCWRL for modeling of side chains. Funding to pay the Open Access publication charges for this article was provided by Department of Biotechnology, Government of India.

Conflict of interest statement. None declared.

REFERENCES

1. Pamer,E. and Cresswell,P. (1998) Mechanisms of MHC class I–restricted antigen processing. *Annu. Rev. Immunol.*, **16**, 323–358.
2. Songyang,Z., Blechner,S., Hoagland,N., Hoekstra,M.F., Piwnica-Worms,H. and Cantley,L.C. (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.*, **4**, 973–982.
3. Udaka,K., Wiesmuller,K.H., Kienle,S., Jung,G., Tamamura,H., Yamagishi,H., Okumura,K., Walden,P., Suto,T. *et al.* (2000) An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics*, **51**, 816–828.
4. Pawson,T. (1994) Introduction: protein kinases. *Faseb. J.*, **8**, 1112–1113.
5. Blom,N., Sicheritz-Ponten,T., Gupta,R., Gammeltoft,S. and Brunak,S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.
6. Huang,H.D., Lee,T.Y., Tzeng,S.W. and Horng,J.T. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **33**, W226–W229.
7. Xue,Y., Zhou,F., Zhu,M., Ahmed,K., Chen,G. and Yao,X. (2005) GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res.*, **33**, W184–W187.
8. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
9. Rammensee,H., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.
10. Singh,H. and Raghava,G.P. (2001) ProPred: prediction of HLA-DR binding sites. *Bioinformatics*, **17**, 1236–1237.
11. Brinkworth,R.I., Breinl,R.A. and Kobe,B. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl Acad. Sci. USA*, **100**, 74–79.
12. Schueler-Furman,O., Altuvia,Y., Sette,A. and Margalit,H. (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci.*, **9**, 1838–1846.
13. Tong,J.C., Tan,T.W. and Ranganathan,S. (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci.*, **13**, 2523–2532.
14. Pohlmann,T., Bockmann,R.A., Grubmuller,H., Uchanska-Ziegler,B., Ziegler,A. and Alexiev,U. (2004) Differential peptide dynamics is linked to major histocompatibility complex polymorphism. *J. Biol. Chem.*, **279**, 28197–28201.
15. Martin,L., Catherinot,V. and Labesse,G. (2006) kinDOCK: a tool for comparative docking of protein kinase ligands. *Nucleic Acids Res.*, **34**, W325–W329.
16. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
17. Schueler-Furman,O., Elber,R. and Margalit,H. (1998) Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes. *Fold Des.*, **3**, 549–564.
18. Ansari,M.Z., Yadav,G., Gokhale,R.S. and Mohanty,D. (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.*, **32**, W405–W413.
19. Trivedi,O.A., Arora,P., Vats,A., Ansari,M.Z., Tickoo,R., Sridharan,V., Mohanty,D. and Gokhale,R.S. (2005) Dissecting the mechanism and assembly of a complex virulence mycobacterial lipid. *Mol. Cell.*, **17**, 631–643.
20. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L.Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001–2014.
21. Robinson,J., Waller,M.J., Parham,P., de Groot,N., Bontrop,R., Kennedy,L.J., Stoehr,P. and Marsh,S.G. (2003) IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res.*, **31**, 311–314.
22. Miyazawa,S. and Jernigan,R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, **256**, 623–644.
23. Betancourt,M.R. and Thirumalai,D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.*, **8**, 361–369.
24. Matsuoka,Y., Hughes,C.A. and Bennett,V. (1996) Adducin regulation. Definition of the calmodulin-binding domain and sites of phosphorylation by protein kinases A and C. *J. Biol. Chem.*, **271**, 25157–25166.

25. Kempler,K., Toth,J., Yamashita,R., Mapel,G., Robinson,K., Cardasis,H., Stevens,S., Sellers,J.R. and Battelle,B.A. (2007) Loop 2 of Limulus Myosin III Is Phosphorylated by Protein Kinase A and Autophosphorylation. *Biochemistry*, **46**, 4280–4293.
26. Prager,K., Wang-Eckhardt,L., Fluhrer,R., Killick,R., Barth,E., Hampel,H., Haass,C. and Walter,J. (March 14, 2007) A structural switch of presenilin 1 by GSK-3beta mediated phosphorylation regulates the interaction with beta -catenin and its nuclear signaling. *J. Biol. Chem.*, 10.1074/jbc.M608437200.
27. Kruger,T., Schoor,O., Lemmel,C., Kraemer,B., Reichle,C., Dengjel,J., Weinschenk,T., Muller,M., Hennenlotter,J. *et al.* (2005) Lessons to be learned from primary renal cell carcinomas: novel tumor antigens and HLA ligands for immunotherapy. *Cancer Immunol. Immunother.*, **54**, 826–836.
28. Diella,F., Cameron,S., Gemund,C., Linding,R., Via,A., Kuster,B., Sicheritz-Ponten,T., Blom,N. *et al.* (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.