Published online 7 June 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W265–W269***
doi:10.1093/nar/gkn346

KFC Server: interactive forecasting of protein interaction hot spots

Steven J. Darnell[1], Laura LeGault[2] and Julie C. Mitchell[1,3,*]

[1]Department of Biochemistry, [2]Department of Computer Sciences and [3]Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706, USA

Received January 31, 2008; Revised April 17, 2008; Accepted May 8, 2008

ABSTRACT

The KFC Server is a web-based implementation of the KFC (Knowledge-based FADE and Contacts) model—a machine learning approach for the prediction of binding hot spots, or the subset of residues that account for most of a protein interface's binding free energy. The server facilitates the automated analysis of a user submitted protein–protein or protein–DNA interface and the visualization of its hot spot predictions. For each residue in the interface, the KFC Server characterizes its local structural environment, compares that environment to the environments of experimentally determined hot spots and predicts if the interface residue is a hot spot. After the computational analysis, the user can visualize the results using an interactive job viewer able to quickly highlight predicted hot spots and surrounding structural features within the protein structure. The KFC Server is accessible at http://kfc.mitchell-lab.org.

INTRODUCTION

Protein–protein interactions underlie most biological processes, hence there is a great interest in modifying protein interfaces to elicit cellular responses. A major focus of this effort involves binding 'hot spots', a small subset of residues that account for a significant part of a protein interface's free energy of binding (1–4). Many studies have successfully disrupted protein interactions by mutating individual or small numbers of interface residues, to a point that databases cataloging hundreds of experimental interface mutations exist (5,6). In addition to mutation studies, hot spots are also receiving attention as potential binding motifs for small molecule inhibitors of protein interactions (7,8).

Hot spot identification requires the experimental characterization of a mutation's effect on binding affinity. Since the stability of protein complexes is mediated by a collection of biophysical properties (including hydrophobicity, van der Waals forces, shape specificity, hydrogen bonds, salt bridges and solvent accessibility (2,9–12) among others), hot spot searches typically focus on mutations that disrupt hydrogen bonds, van der Waals contacts and chemical complementarity. Predictive models can improve the efficiency of this process. Even though the forces that mediate binding are not fully understood (12), computational models that use physics and knowledge-based methodologies (13–20) can successfully predict subsets of hot spot residues using different aspects of these forces.

Here, we present the KFC Server—a web-based implementation of the KFC (Knowledge-based FADE and Contacts) model, which uses a physical and knowledge-based approach to predict binding hot spots. Normally, predictive web servers generate textual output, which a user must then manually incorporate into their existing structural model. To streamline this process, we designed a customized interface for the KFC Server that allows users to visualize hot spot predictions along with the protein structure. In addition, users may upload scores for Robetta's alanine scanning (16), ConSurf sequence conservation (21,22) or known experimental data, such as that available through the Alanine Scanning Energetics Database (5) or the Binding Interface Database (6). Individual controls are provided to display a space-filling, stick, or surface representation of each interface residue. The controls are color-coded to indicate the chemical property of each residue, and whether the residue is a predicted hot spot. The control panel also simplifies the execution of several standard molecular viewer features, such as altering visual representations, changing color schemes and rendering molecular surfaces. Additionally, the server includes two important and unique features: the ability to highlight local regions of well-matched or mismatched shape specificity between the binding partners, and the ability to save and recall snapshots from the user's session. The result is an interactive tool that enables users to intuitively understand the role of shape specificity and biochemical contacts associated with binding hot spots.

*To whom correspondence should be addressed. Tel: 608 890 0365; Fax: 608 262 3453; Email: jcmitchell@wisc.edu

HOT SPOT PREDICTIONS

Recently, we introduced the KFC model, a machine learning approach for predicting binding hot spots within protein–protein interactions. Since this model is the basis of predictions generated by the KFC Server, the following section summarizes its construction and performance. Further details are available in the original manuscript (23).

KFC MODEL

The KFC model is comprised of two decision tree-based classifiers: K-FADE [based on shape specificity features calculated by the Fast Atomic Density Evaluator, or FADE (24)] and K-CON (based on biochemical contact features). Each decision tree, which provides a set of hierarchical rules for hot spot classification, was trained by a supervised learning process to recognize the local structural environments that are indicative of hot spots. In practice, every path through the tree terminates with a prediction/classification as to whether a residue is a hot spot. The training set used for learning consisted of 249 experimentally characterized alanine mutations within the interface of 16 nonredundant protein complexes. Structures for each complex were obtained from the Protein Data Bank (PDB) (25). For this work, residues were classified as hot spots if their mutation to alanine resulted in a change of binding energy ($\Delta\Delta G$) greater than 2 kcal/mol. The data mining tool C5.0 (Rulequest, St. Ives, Australia) was used to create predictive models from many different combinations of structurally-derived chemical and physical features that describe the interface residues, and those that best described the hot spot environment were selected as features for the K-FADE and K-CON models. K-FADE predicts hot spots using the size of the residue and the radial distribution of shape specificity and interface points. K-CON predicts hot spots in terms of a residue's intermolecular atomic contacts, hydrogen bonds, interface points and chemical type.

To validate this approach, KFC's ability to predict known hot spots was compared to the Robetta Interface Alanine Scanning (Robetta-Ala) service, a leading hot spot prediction utility that predicts the $\Delta\Delta G$ of a residue's mutation to alanine (16,26). The predictive performance of each method was described in terms of *F1 score*, a statistical measure of accuracy balancing *precision* (the fraction of positive hot spot predictions that are correct) and *recall* (the fraction of known hot spots that are predicted). As described in ref. (23), we have used the F1 score as a standard measure of predictive accuracy.

A cross-validation analysis of the training data showed that KFC exceeded the predictive accuracy of Robetta-Ala, and a model combining KFC and Robetta-Ala performed significantly better than Robetta-Ala alone ($P = 0.02$). The combined model predicts a residue is a hot spot if either KFC or Robetta-Ala makes a positive prediction. In addition, this result was verified by using an independent test set of 112 mutations and the final KFC models trained on the full training set. Again, KFC slightly outperformed Robetta-Ala, and the combination of KFC and Robetta-Ala achieved a large statistical improvement in predictive accuracy over either individual model ($P = 0.0071$). In addition to its high accuracy, KFC is computationally fast. Using common computer hardware, a typical KFC analysis is complete in less than 2 min. Given its speed and accuracy, the KFC model can support hot spot predictions for multiple users in a server environment.

WEB SERVER IMPLEMENTATION

The KFC Server automates the analysis of a user submitted protein–protein or protein–DNA interface and presents an interactive visualization of its hot spot predictions. The server is organized into three main sections: the submission page, the queue and the job viewer. On the submission page, the user provides a protein structure and defines the interface to analyze. Next, the submitted job enters the server's queue for processing. Afterwards, the job viewer superimposes the results from the KFC analysis onto the protein structure. These tools help the user to quickly analyze a protein interface and to simply visualize the structural environment around putative hot spots.

Submitting a protein complex

Before the KFC analysis can begin, the user must provide the structure of a protein complex and define the interface to analyze. A protein interface is the region between two binding partners, where each partner is comprised of one or more molecules. Typically, users will analyze structures found in the PDB; however, the KFC Server will accept any PDB formatted coordinate file, such as structures generated from protein docking or other molecular modeling techniques. As a warning, model structures containing many clashes may vastly overestimate the number of hot spots in the interface. Also, the KFC model can only analyze structures containing proteins and nucleic acids. Some processing errors can be avoided if other types of molecules are removed from the PDB file prior to submission.

The user can provide a PDB file in one of two ways: enter a four character PDB code to download the file directly from the PDB, or upload a structure from their computer using the submission form. The binding partners are defined by listing the PDB chain identifiers that comprise each partner. If the chain identifiers are not found in the file, the two binding partners are defined by splitting the file at the first TER record. Files that do not contain a bound complex or specifying noncontacting partners will not yield useful results. Users may also register to have their predictions sent to their email address, but registration is not required to submit jobs or view results. The KFC Server is free and open to all users.

Computing predictions

Submitted jobs are processed in a queue, ensuring that simultaneous submissions cannot exceed the hardware resources of the server. When the job is activated, the server begins to calculate the structural features surrounding each residue in the interface. Here, an interface residue

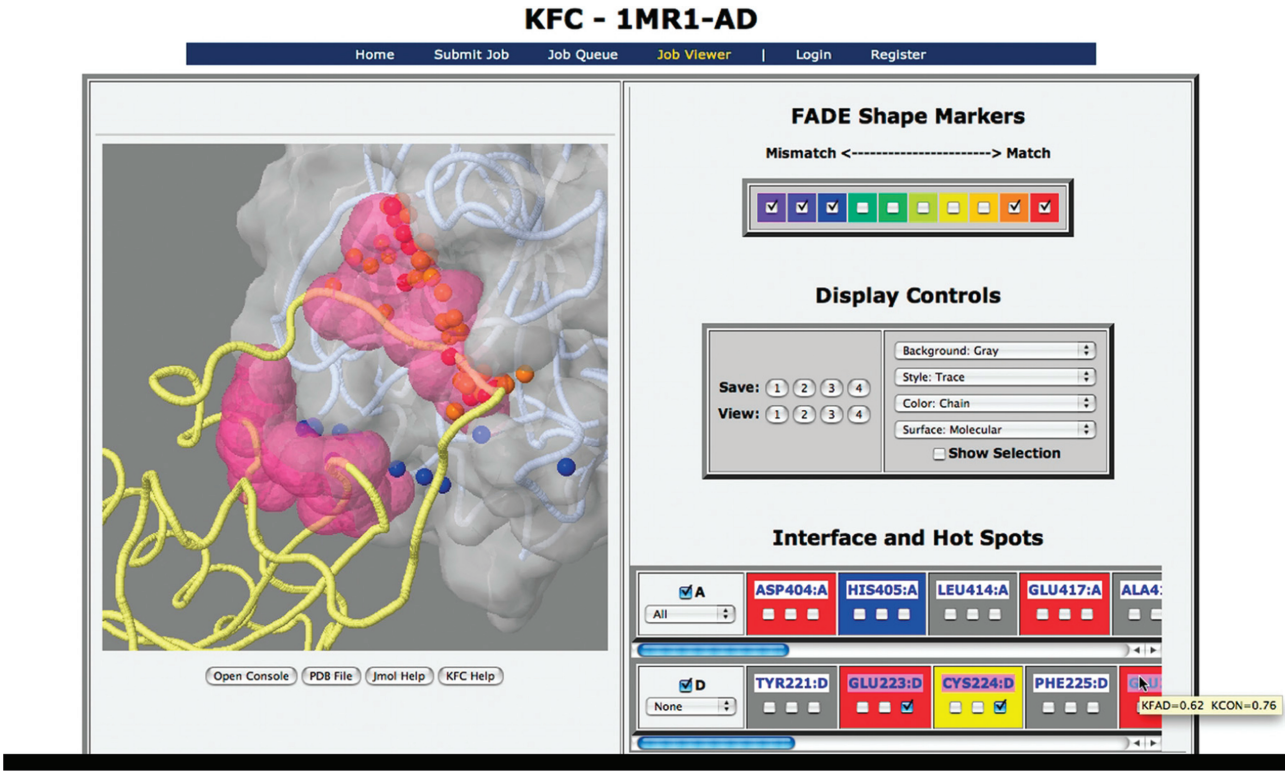

Figure 1. The major components of the job viewer are: the molecular viewer, the FADE shape marker controls, the display controls and the interface and hot spot controls. Displayed is the KFC analysis of the Smad4/Ski protein complex (PDB: 1MR1) and the control panel configuration used to generate the image. Molecular surfaces surround Smad4 (white) and the predicted hot spots in Ski (pink). This representation clearly shows two distinct hot spot clusters, one which is strongly associated with favorable shape specificity (red and orange spheres). Regions of mismatched shape specificity (blue and violet spheres) flank both clusters. In the case of Glu268, both the K-FADE and K-CON models predict this residue is a hot spot.

is defined as a residue with at least one atom within 4 Å of the opposite binding partner. First, a FADE analysis is performed and the radial distribution of shape specificity markers is calculated about each residue. If an error occurs at this point, it is likely that the submitted chain identifiers are incorrect or the selected binding partners do not form a valid interface.

Next, each residue's intermolecular contacts and hydrogen bonds are tabulated. The assignment of hydrogen bonds is automated by PDB2PQR (27), which produces comparable results to the WHAT IF (28) program used to create KFC training data. Finally, the K-FADE and K-CON models are applied to the calculated features and the putative hot spots are selected. The results file lists the predicted classification of each residue by both models (namely 'hot spot' or 'interface residue') and a score indicating the confidence of each prediction (its worst value is 0 and its best value is 1).

Viewing hot spot predictions

The interactive job viewer enables the user to quickly highlight predicted hot spots and surrounding structural features. Its customized interface is built around the Jmol molecular viewer (http://www.jmol.org) and requires the Sun Java Runtime Environment (version 1.4 or later, http://www.java.com) and a Javascript enabled web browser to function. The job viewer has two major components: the molecular viewer and the control panel (Figure 1). Users can directly interact with the molecular viewer or use the control panel to affect the display. This section describes the functionality of the control panel components in detail.

FADE shape markers. At the top of the control panel is a group of check boxes that affect the display of FADE-calculated shape specificity markers. Users can highlight different degrees of shape specificity within the interface by clicking on the corresponding color-coded boxes. Red and orange represents regions with well-matched shape fit, yellow and green represent flat surfaces and blue and violet represent mismatched surfaces. When boxes are selected, corresponding spheres will appear in the molecular viewer to highlight the different degrees of shape specificity.

Display controls. The display controls are located in the middle of the control panel and they alter the appearance of the selected atoms. By default, all protein atoms in the complex are selected. Different sets of atoms can be selected using the pop-up menus located in the 'Interface and Hot Spot' panel (next section). Advanced users may also change the atom selection by using the Jmol scripting language. To help users refine their atom selections, the 'Show Selection' check box can be used to highlight the current atom selection.

The panel also contains four pop-up menus which simplify the execution of several molecular viewer commands. The background menu controls the color of the background. The style menu manipulates the representation of the selected backbone or sidechain atoms. Its styles include several stick, wireframe, ribbon and space-filling representations. The color menu applies different color schemes to the selected atoms, including coloring by chain, secondary structure, temperature factor and chemical properties. Lastly, the surface menu allows the user to render a molecular or solvent accessible surface around the selected molecules.

In addition to these functions, users can also save up to four different views of their modeling session. Clicking one of the 'Save' buttons will record the current state of the display, and clicking the appropriate 'View' button will restore the viewer to the saved state. This uncommon feature is relatively simple to build into Jmol applets, and we have found it tremendously useful when doing detailed explorations of new protein interfaces.

Interface and hot spots. The bottom panel controls the display of each interface residue and predicted hot spot, and provides summary information about each residue. The residues are grouped by chain, and the appearance of each sidechain is controlled by a set of three check boxes. The first check box highlights the residue with a space filling representation, the second box shows the residue in stick form and the third box adds a surface around the residue. Also, the coloring within each table cell encodes information about that particular residue. The background color describes the chemical type of the amino acid. Hydrophobic residues are gray, polar are yellow, acidic are red and basic are blue. Second, pink highlighting around a residue's name indicates that it is a putative hot spot.

Holding the mouse over the highlighted name activates a display box containing KFC confidence scores for that hot spot. Scores marked as 'K-FADE' mean that the hot spot was predicted based on its shape specificity features, and those marked 'K-CON' are predicted based on biochemical contacts. In some cases, both methods will predict the same sidechain, suggesting that both geometry and physics play a strong role in its favorability. It is important to note that KFC makes predictions for all residues within the protein interface. As such, residues whose sidechains are directed into the protein core may be predicted as hot spots. Mutation of these residues can indirectly disrupt the protein complex by destabilizing the monomeric protein. To identify this scenario, putative hot spots with only intermolecular backbone contacts are flagged as 'backbone' in the results file, and the pop-up box of scores will distinguish these cases using 'KFADE-bb' and 'KCON-bb'.

In addition to these individualized controls, the lower panel contains controls that are applied to an entire chain. The check box next to the chain name toggles whether the chain is displayed or hidden. Additionally, the pop-up menu beneath the chain name determines which subset of atoms is selected for action by the 'Display Controls'. Note that selections made with the Jmol scripting language will override any mouse-driven selection and display controls.

Incorporating external data

As noted in the description of the KFC model, combining different prediction methods can lead to an overall improvement in hot spot predictions. In order to allow users to visualize hot spot predictions generated using other methods, and to allow easy comparison between methods, three types of external data can be uploaded. Because we have used Robetta-Ala service as a basis for comparison with KFC (23), the server supports uploading Robetta's predictions. Sidechains for which alanine substitution is predicted to generate a $\Delta\Delta G$ greater than 2 kcal/mol will be marked as hot spots.

Two additional types of external input further extend the utility of our hot spot visualization. Because sequence conservation at interface residues can indicate functional importance, we also include the ability to upload output from the ConSurf server (21,22). Sidechains with conservation class greater than 7, as determined by ConSurf, are listed as conserved in the output. Finally, we allow the user to upload a file containing hot spots that are experimentally determined or otherwise known. The user indicates whether each included residue is a known hot spot, and score describing its effect on binding. This score is a variable field and can contain any preferred type of score, such as $\Delta\Delta G$ values or the percentage decrease in binding affinity.

We have created a demonstration area on the server with a KFC analysis for each protein complex in the training and test sets described in the original KFC manuscript (23) augmented with Robetta-Ala predictions, ConSurf conservation scores and experimental results for the listed residues. These examples illustrate how quickly users can identify convergent hot spot predictions (suggesting a particularly important residue for binding) and simultaneously view the predictions from the different models within the context of the displayed structure. Additionally, these results are downloadable and can be used as a benchmarking set for the development of other binding hot spot prediction methods. This integration of external data results in a powerful way for a user to intuitively investigate how chemical, physical, energetic and conservational properties affect the importance of residues within a binding interface.

CONCLUSIONS

The importance of a user-friendly interface for computational tools cannot be over emphasized. In this spirit, the KFC Server streamlines the submission process, automates the analysis and integrates its predictions into a visual framework. Our hope is that the KFC Server is adopted by experimenters to guide molecular recognition experiments. Given a known structure, its most straightforward application is to predict residues whose mutation can significantly disrupt an interaction. Currently, we are also using the KFC Server to design a protein with *improved* affinity for its binding partner. Using the shape

specificity analysis to highlight shape mismatches, we are performing substitutions in these regions to optimize shape fit and biochemical contacts. In addition, we are screening modeled mutations with KFC to predict if they introduce new binding hot spots.

ACKNOWLEDGEMENTS

This work was supported by the NIH-NLM (5T15LM 007359 to SJD) and U.S. Department of Energy (DE-FG02-04ER25627 to JCM).

Conflict of interest statement. None declared.

REFERENCES

1. Clackson,T. and Wells,J. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386.
2. Jones,S. and Thornton,J. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
3. Bogan,A. and Thorn,K. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.
4. Moreira,I., Fernandes,P. and Ramos,M. (2007) Hot spots—a review of the protein-protein interface determinant amino-acid residues. *Proteins*, **68**, 803–812.
5. Thorn,K. and Bogan,A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285.
6. Fischer,T., Arunachalam,K., Bailey,D., Mangual,V., Bakhru,S., Russo,R., Huang,D., Paczkowski,M., Lalchandani,V., Ramachandra,C. *et al.* (2003) The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*, **19**, 1453–1454.
7. Arkin,M. and Wells,J. (2004) Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat. Rev. Drug Discov.*, **3**, 301–317.
8. Thanos,C., DeLano,W. and Wells,J. (2006) Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc. Natl Acad. Sci. USA*, **103**, 15422–15427.
9. Xu,D., Tsai,C. and Nussinov,R. (1997) Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.*, **10**, 999–1012.
10. Sheinerman,F., Norel,R. and Honig,B. (2000) Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.*, **10**, 153–159.
11. Chakrabarti,P. and Janin,J. (2002) Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–343.
12. Li,X., Keskin,O., Ma,B., Nussinov,R. and Liang, J. (2004) Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J. Mol. Biol.*, **344**, 781–795.
13. Massova,I. and Kollman,P. (1999) Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *J. Am. Chem. Soc.*, **121**, 8133–8139.
14. Guerois,R., Nielsen,J. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
15. Gao,Y., Wang,R. and Lai,L. (2004) Structure-based method for analyzing protein-protein interfaces. *J. Mol. Model.*, **10**, 44–54.
16. Kortemme,T., Kim,D. and Baker,D. (2004) Computational alanine scanning of protein–protein interfaces. *Sci. STKE*, **2004**, pl2.
17. del Sol,A. and O'Meara,P. (2005) Small-world network approach to identify key residues in protein-protein interaction. *Proteins*, **58**, 672–682.
18. Li,L., Zhao,B., Cui,Z., Gan,J., Sakharkar,M. and Kangueane,P. (2006) Identification of hot spot residues at protein-protein interface. *Bioinformation*, **1**, 121–126.
19. Ofran,Y. and Rost,B. (2007) Protein-protein interaction hotspots carved into sequences. *PLoS Comput. Biol.*, **3**, e119.
20. Guney,E., Tuncbag,N., Keskin,O. and Gursoy,A. (2008) HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res.*, **36**, D662–D666.
21. Glaser,F., Pupko,T., Paz,I., Bell,R., Bechor,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
22. Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben-Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.
23. Darnell,S., Page,D. and Mitchell,J. (2007) An automated decision-tree approach to predicting protein interaction hot spots. *Proteins*, **68**, 813–823.
24. Mitchell,J., Kerr,R. and Ten Eyck,L. (2001) Rapid atomic density methods for molecular shape characterization. *J. Mol. Graph. Model.*, **19**, 325–330.
25. Berman,H., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T., Weissig,H., Shindyalov,I. and Bourne,P. (2000) The Protein Data Bank. *Nucleic Acids Res.*, 28, 235–242.
26. Kortemme,T. and Baker,D. (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
27. Dolinsky,T., Nielsen,J., McCammon,J. and Baker,N. (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.*, **32**, W665–W667.
28. Vriend,G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.*, **8**, 52–56.