Published online 16 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D191–D194***
doi:10.1093/nar/gkn716

SDR: a database of predicted specificity-determining residues in proteins

Jason E. Donald[1,*] and Eugene I. Shakhnovich[2]

[1]Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA and
[2]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA

Received August 6, 2008; Revised September 25, 2008; Accepted September 30, 2008

ABSTRACT

The specificity-determining residue database (SDR database) presents residue positions where mutations are predicted to have changed protein function in large protein families. Because the database precalculates predictions on existing protein sequence alignments, users can quickly find the predictions by selecting the appropriate protein family or searching by protein sequence. Predictions can be used to guide mutagenesis or to gain a better understanding of specificity changes in a protein family. The database is available on the web at http://paradox.harvard.edu/sdr.

INTRODUCTION

The specificity-determining residue database (SDR database) (http://paradox.harvard.edu/sdr) presents residue positions that are predicted to have changed protein function when mutations occurred during evolution. There are two distinctive features of the SDR database. First, it predicts specificity-determining residue positions for large protein families that likely have multiple protein functional specificities. Second, it pre-computes these predictions so that they can be accessed quickly, without requiring the user to input multiple sequence alignments or know the function of proteins within the alignment.

In the past decade, there has been increased interest in the development of methods to predict specificity-determining residues. In general, the predictions of these positions depend on a logical assumption: specificity-determining residues will be conserved by groups of proteins with the same function, but often different between groups of proteins with different functions (1).

These methods can be divided into two classes based on the number of subfamilies that are considered in the calculation. The first class of methods considers protein families that can be divided into two subfamilies (2–5). For example, if a set of homologous proteins (also known as a protein family) has two distinct functional specificities, which residues are responsible for the change in specificity?

The second class of methods, including the method underlying the SDR database, considers more than two subfamilies at a time (1,6–15). For a protein family that includes proteins with many different functions, one would like to determine which residues have been mutated repeatedly during evolution to modify protein function. As an example, if a position is alanine in proteins with the first function, leucine for proteins with the second function and arginine for proteins with the third function, etc. such a position is likely to be determining the function of proteins in the family.

Several years ago, Mirny and Gelfand developed a statistical model that found these residues when a multiple sequence alignment could be divided into groups based on function (1). An existing webpage builds off this method to make predictions of specificity-determining residues based on a user-provided multiple sequence alignment and functional categorization of the proteins (7). While additional careful comparisons are needed, a recent comparison of different methods found that this method (7) predicted many specificity-determining residues in three families with several functions (8).

Understandably, predictions have often focused on smaller protein families where most protein sequences can be functionally categorized. The SDR database combines the statistical method of Mirny and Gelfand (1) with an automated, approximate method of grouping proteins by their functional specificities (16) that allows prediction of specificity-determining residues for large protein families. This method makes accurate predictions of the specificity-determining residue positions in two large families of transcription factors (17). In the SDR database, we now used this method to pre-compute specificity-determining residue predictions for all large families found in a comprehensive protein database, PFAM (18).

In summary, the SDR database offers several advantages. First, the SDR database requires minimal input from the user (see below). Using the automated functional

*To whom correspondence should be addressed. Tel: +617 893 8713; Fax: +215 573 7229; Email: jdon@mail.med.upenn.edu
Correspondence may also be addressed to Eugene I. Shakhnovich. Tel: +617 495 4130; Fax: +617 384 9228;
Email: shakhnovich@chemistry.harvard.edu

specificity grouping means that users do not have to create multiple sequence alignments or functionally categorize the proteins. As can be expected, such categorization is very time consuming and approximate because of limited experimental data. Second, the SDR database focuses on large protein families, such as GPCRs, many of which are of great biological interest. The larger number of sequences in these families provides better statistics, which should lead to more accurate calculations, as was found for transcription factors (17). Finally, by being automated, the method is consistent in the way that it treats protein alignments. This could be an important advantage for studies that compare predicted specificity-determining residues in different protein families.

THE DATABASE

On the website, we present the statistical significance of the predictions in the form of *Z*-scores (1,17) (the number of standard deviations the calculated mutual information is away from the expected value), and we display the most significant positions both on a multiple sequence alignment and on existing protein structures. The predicted specificity-determining residues can be conveniently accessed by selecting the family from the search page or using a query protein sequence to be matched to an existing family using the program HMMER (18). Selecting a family leads to a webpage containing the family name, the number of sequences used, a graph of the size of the giant component (used for automated functional grouping) and links to pages where significant predictions are highlighted on a subset of the multiple sequence alignment or a protein structure, where available. Other links lead to text files containing the *Z*-score for each position and the full multiple sequence alignment organized by predicted functional specificities. From the highlighted multiple sequence alignment, logo plots from WebLogo (19) are available for each predicted position, showing which residues are found at that position in the different functional groups. A compressed file that contains all data files can be downloaded for further analysis.

The database contains 1346 protein families from the PFAM 20.0 (18) and the GPCRDB 10.0 (20) databases. As new sequences are added from newer releases of PFAM and GPCRDB, the number of families covered by the database should increase. A family was selected from PFAM for prediction of specificity-determining residues if it contained at least 500 sequences in the full multiple sequence alignment and if the family contained position with not >30% of the position is made up of gaps. From the GPCRDB, five families of G-protein coupled receptors (GPCR) were selected: the full class A family and four class A subfamilies (amine, olfactory, opsin and peptide). Class A is the largest GPCR class, and the four selected subfamilies are the largest subfamilies in this class. Loops were excluded from the GPCRDB calculations because of the much lower sequence identity in these regions. Because the sequence alignments in the transmembrane regions should be more accurate, the calculations are for the transmembrane regions alone. For all families with more than 5000 sequences, 5000 sequences were randomly selected from the multiple sequence alignment for the calculation because of computer memory constraints. This is similar to using an older sequence database where fewer sequences are available.

Further details and frequently asked questions are available at the website.

STATISTICAL ANALYSIS OF PREDICTIONS

Because functional residues are often clustered in three dimensions, we expect that specificity-determining residues will be in contact with one another more often than other pairs of residues. As a test of quality of the predictions, we randomly selected 800 proteins, each from a different PFAM family, where the 3D structure is known. For both pairs of predicted specificity-determining residues and pairs of residues not predicted, we calculated the fraction of pairs in within a given C_β distance (Figure 1A). To ensure that the contacts between residues are not due to proximity in sequence, we only considered pairs separated by at least eight residues. Notably, predicted specificity-determining residues are much more likely to be in contact than positions that are not predicted, supporting the quality of the predictions. As a result, the predicted positions may also aid protein structure prediction.

AN EXAMPLE

As an example, we selected the transmembrane region of the amine-binding subfamily of class A GPCRs (Table 1). After the predictions were made, several crystal structures of amine-binding GPCRs have become available. Several of these structures are of an inverse agonist, carbazol, bound to human β2 adrenergic receptor. It is expected that carbazol uses the same binding pocket as the native agonists.

While the specificity-determining residues in this family are not known experimentally, the residues that form the binding pocket are known. Some of these positions are likely too important for the protein to mutate and remain functional, but other positions are likely specificity-determining residues determining ligand binding specificity. Therefore, several of the predicted specificity-determining positions should be near the ligand.

As a test of the prediction sensitivity of the method for this family, we considered how many of the predictions are close to carbazol in the solved structure. Four out of 10 predicted positions have C_β atoms within 5.5 Å of the carazolol (21). The predicted positions in close proximity are likely to be very important for amine-binding GPCR ligand-binding specificity. The proximity of the predicted residues further supports the predictions found in the SDR database.

Other predicted positions may be involved in GPCR function ways other than ligand binding, such as protein–protein interaction or degree of basal activation. In the related opsin subfamily of class A GPCRs, several of the predicted specificity-determining residues appear

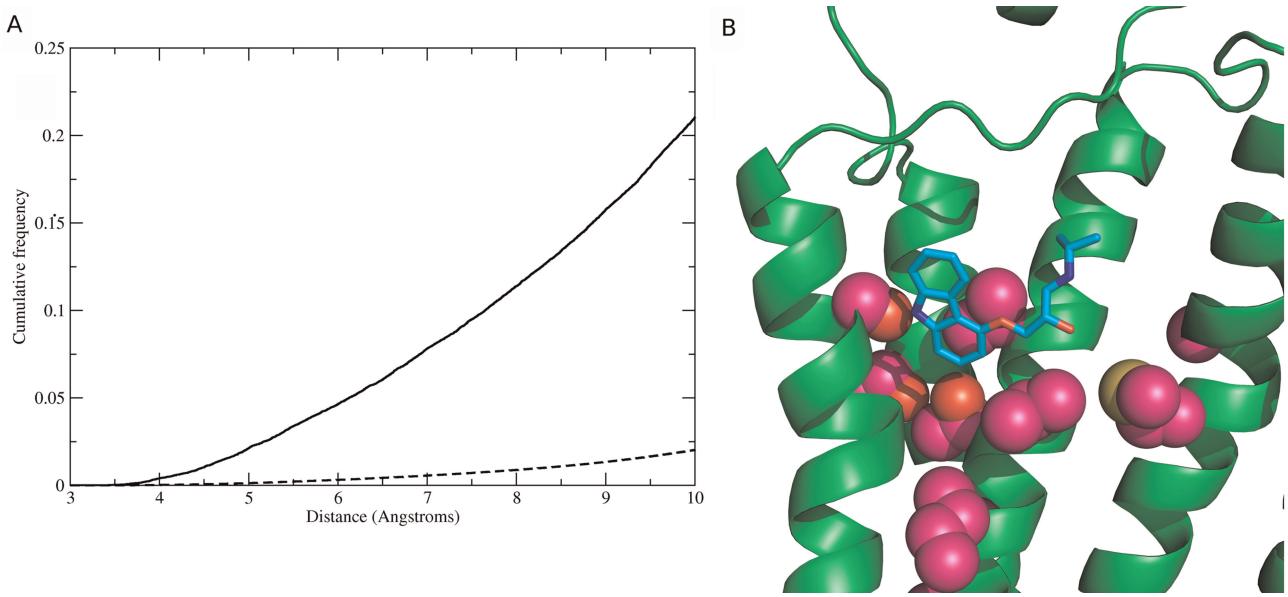

Figure 1. Specificity-determining residues are in proximity to each other and to the ligand-binding site of an example. (**A**) Cumulative distribution of observing two C_β atoms within a distance cutoff. The distribution for specificity-determining residues is shown in a solid line and the distribution for all other positions is shown in a dashed line. (**B**) The predicted specificity-determining residues of the amine subfamily of class A GPCRs are shown on the human β2 adrenergic receptor (21) (pdb code 2RH1). The inverse agonist (carbazol) is shown in cyan, and the predicted side chains are shown in magenta. Helices six and seven of the GPCR have been removed for clarity. The figure was made using PyMOL (25).

Table 1. The predicted specificity-determining residues of the amine subfamily of class A GPCRs

Z-score	SDR alignment numbering	GPCRDB numbering	Most common residues	2RH1 residue and numbering	C_β distance to carbazol (Å)
7.5	104	419	VAT	V157	14.0
7.4	74	326	CSV	V117	4.0
6.9	83	335	AVI	V126	16.8
6.7	78	330	IVL	I121	8.3
6.6	71	323	VIY	V114	4.2
6.4	125	512	STG	S203	3.8
6.4	44	227	VLI	M82	8.5
6.2	111	426	IVL	T164	8.1
6.2	75	327	TNY	T118	5.2
6.1	47	230	LFT	A85	12.0

The three most common residues are shown using single letter abbreviations. The most common residue is listed first, followed by the second and third most common residues. The distance of each residue's C_β atom to the closest atom on the inverse agonist, carbazol, in the recent β2 adrenergic receptor structure (pdb code 2RH1), is also listed.

to play a role in G-protein coupling in the well-studied protein bovine rhodopsin. Residue L226 is found at the edge of intracellular loop 3, known to be important for interaction with G-proteins (22), E113 and A117 are in contact with residues where naturally occurring mutations cause constitutively activity (G90D, A292E, K296E) (23) and residue E134 is part of the D/ERY motif that appears to play an important role in constitutive activation (24).

CONCLUSIONS

As shown in the example, the SDR database predicts likely specificity-determining residues not only for DNA-binding proteins, but also for proteins with other functions, such as ligand binding. Predicted positions are more likely to be in contact than other positions in a protein. We expect that the database will provide useful targets for experimental mutagenesis as well as the design and modification of protein function. The predictions also should lead to a better understanding protein function of large protein families.

ACKNOWLEDGEMENTS

J.E.D. thanks Gevorg Grigoryan for a helpful discussion.

FUNDING

National Institutes of Health; National Institutes of Health Hemostasis and Thrombosis training grant (T32 H107971 to J.E.D.). Funding for open access charge: National Institutes of Health. The grant number is GM068670 (a grant to the senior author, E.I.S.).

Conflict of interest statement. None declared.

REFERENCES

1. Mirny,L.A. and Gelfand,M.S. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.*, **321**, 7–20.
2. Feenstra,K.A., Pirovano,W., Krab,K. and Heringa,J. (2007) Sequence harmony: detecting functional specificity from alignments. *Nucleic Acids Res.*, **35**, W495–W498.
3. Pirovano,W., Feenstra,K.A. and Heringa,J. (2006) Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res.*, **34**, 6540–6548.

4. Abhiman,S. and Sonnhammer,E.L. (2005) Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins*, **60**, 758–768.
5. Abhiman,S. and Sonnhammer,E.L. (2005) FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res.*, **33**, D197–D200.
6. Kalinina,O.V., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, **13**, 443–456.
7. Kalinina,O.V., Novichkov,P.S., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.*, **32**, W424–W428.
8. Chakrabarti,S., Bryant,S.H. and Panchenko,A.R. (2007) Functional specificity lies within the properties and evolutionary changes of amino acids. *J. Mol. Biol.*, **373**, 801–810.
9. Casari,G., Sander,C. and Valencia,A. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.
10. Lichtarge,O., Yamamoto,K.R. and Cohen,F.E. (1997) Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.*, **274**, 325–337.
11. Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.
12. Pei,J., Cai,W., Kinch,L.N. and Grishin,N.V. (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, **22**, 164–171.
13. Reva,B., Antipin,Y. and Sander,C. (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.*, **8**, R232.
14. Marttinen,P., Corander,J., Toronen,P. and Holm,L. (2006) Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics*, **22**, 2466–2474.
15. Ye,K., Anton Feenstra,K., Heringa,J., Ijzerman,A.P. and Marchiori,E. (2008) Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. *Bioinformatics*, **24**, 18–25.
16. Donald,J.E. and Shakhnovich,E.I. (2005) Determining functional specificity from protein sequences. *Bioinformatics*, **21**, 2629–2635.
17. Donald,J.E. and Shakhnovich,E.I. (2005) Predicting specificity-determining residues in two large eukaryotic transcription factor families. *Nucleic Acids Res.*, **33**, 4455–4465.
18. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
19. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
20. Horn,F., Bettler,E., Oliveira,L., Campagne,F., Cohen,F.E. and Vriend,G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.
21. Cherezov,V., Rosenbaum,D.M., Hanson,M.A., Rasmussen,S.G., Thian,F.S., Kobilka,T.S., Choi,H.J., Kuhn,P., Weis,W.I., Kobilka,B.K. *et al.* (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science*, **318**, 1258–1265.
22. Wong,S.K. (2003) G protein selectivity is regulated by multiple intracellular regions of GPCRs. *Neurosignals*, **12**, 1–12.
23. Seifert,R. and Wenzel-Seifert,K. (2002) Constitutive activity of G-protein-coupled receptors: cause of disease and common property of wild-type receptors. *Naunyn Schmiedebergs Arch. Pharmacol.*, **366**, 381–416.
24. Cohen,G.B., Yang,T., Robinson,P.R. and Oprian,D.D. (1993) Constitutive activation of opsin: influence of charge at position 134 and size at position 296. *Biochemistry*, **32**, 6111–6115.
25. DeLano,W.L. (2002) *The PyMOL Molecular Graphics System.* DeLano Scientific, Palo Alto, CA, USA.