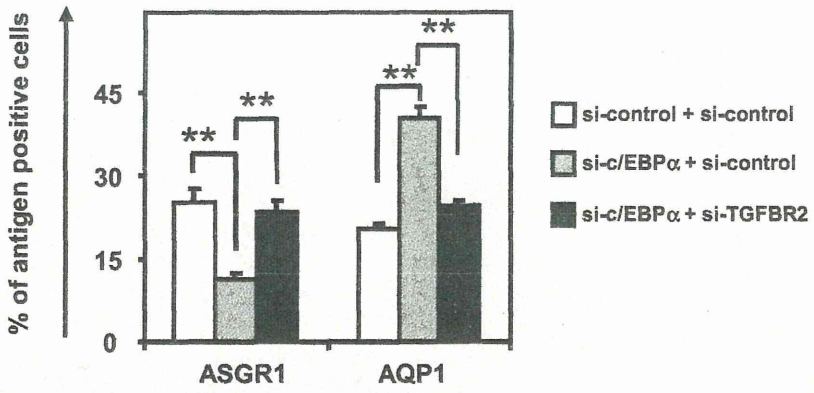

Fig. S8 Inhibition of hepatocyte differentiation by si-c/EBPα transfection was rescued by si-TGFBR2 transfection.

The HBCs were transfected with 50 nM of each of si-control + si-control, si-c/EBPα + si-control, or si-c/EBPα + si-TGFBR2 and cultured with the differentiation hESF-DIF medium for 10 days. The efficiency of hepatocyte or cholangiocyte differentiation was measured by estimating the percentage of ASGR1- or AQP1-positive cells, respectively, using FACS analysis. $*P<0.05$; $**P<0.01$.

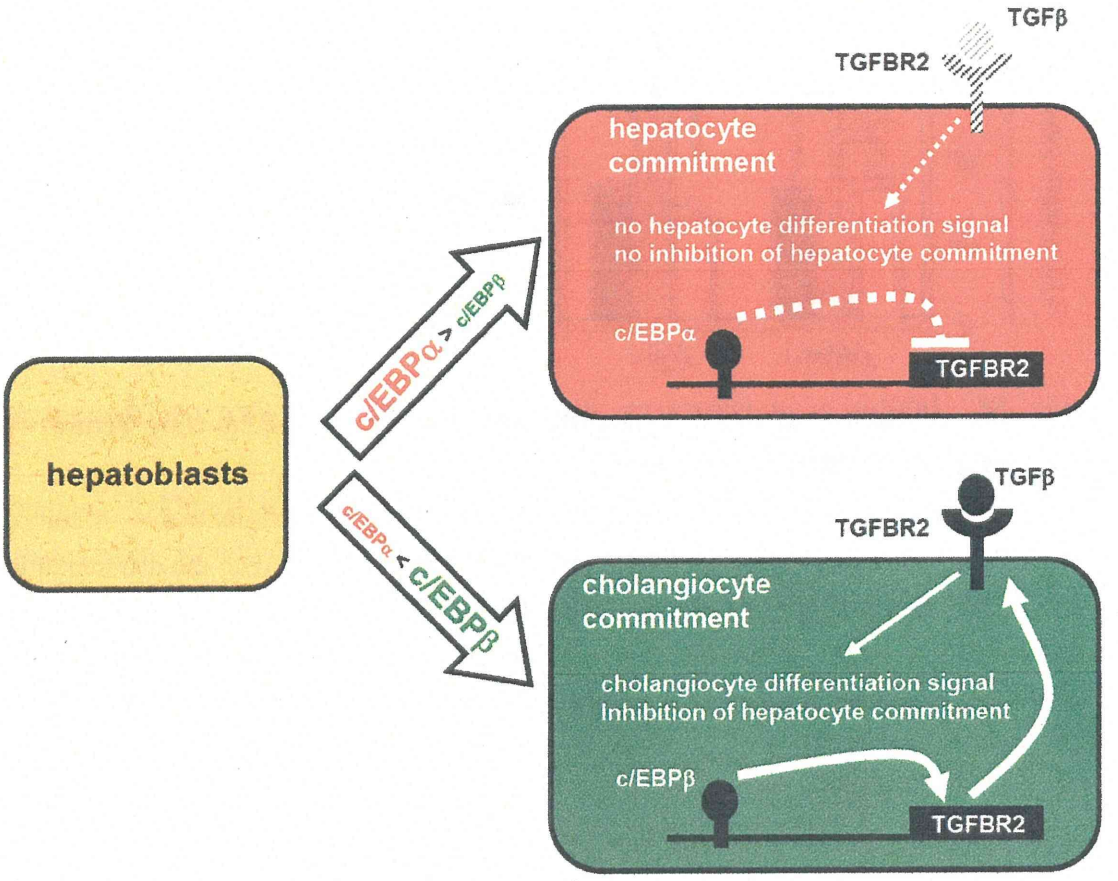

Fig. S9 The lineage segregation of hepatoblasts might be explained by c/EBP-mediated control of TGFBR2 expression.

In hepatocyte differentiation from hepatoblasts, c/EBPα promotes hepatocyte differentiation via negative regulation of TGFBR2 expression. On the other hand, c/EBPβ promotes cholangiocyte differentiation via positive regulation of TGFBR2 expression in cholangiocyte differentiation.

Supplemental Table 1 The primary antibodies used in this study

Antigen	Species	Company (catalog number)	Dilution
CK19	rabbit	Abcam (ab52625)	1/250
AFP	mouse	Cell Signaling (#3903)	1/100
c/EBPβ	rabbit	Santa Cruz Biotechnology (sc-150AC)	1/50
ALB (ELISA)	goat	Bethyl Laboratories (E80-129)	
ALB (FCM)	rabbit	Abcam (ab135575)	1/40
ALB (IHC)	goat	Santa Cruz Biotechnology (sc-46293)	1/200
c/EBPα	rabbit	Abcam (ab40764)	1/50
HNF4α	abcm	Abcam (ab36175)	1/100
TGFBR2	mouse	Santa Cruz Biotechnology (sc-17799)	1/50
ASGR1	goat	Santa Cruz Biotechnology (sc-13467)	1/50
CYP3A4	goat	Santa Cruz Biotechnology (sc-27639)	1/200
AQP1	mouse	Abcam (ab9566)	1/40
EpCAM	mouse	Militenyi Biotec (130-091-254)	1/50

Supplemental Table 2 The secondary antibodies used in this study

Antigen	label	Company	Species	Dilution
rabbit IgG	alexa fluor 488	Molecular Probes	goat	1/1000
rabbit IgG	alexa fluor 488	Molecular Probes	chicken	1/1000
mouse IgG	alexa fluor 488	Molecular Probes	rabbit	1/1000
goat IgG	alexa fluor 488	Molecular Probes	rabbit	1/1000
rabbit IgG	alexa fluor 594	Molecular Probes	mouse	1/1000
goat IgG	alexa fluor 594	Molecular Probes	mouse	1/1000
goat IgG	alexa fluor 594	Molecular Probes	chicken	1/1000
goat IgG	alexa fluor 594	Molecular Probes	donkey	1/1000
mouse IgG	alexa fluor 594	Molecular Probes	chicken	1/1000

Supplemental Table 3 The primers used for real-time RT-PCR in this study

Genes	Primers (forward/reverse, 5' to 3')
CK7	AGACGGAGTTGACAGAGCTG/GGATGGCCCGGTTCATCTC
CK19	CTCCCGCGACTACAGCCACT/TCAGCTCATCCAGCACCCTG
HES1	ATGGAGAAAAATTCCTCGTCCC/TTCAGAGCATCCAAAATCAGTGT
SOX9	TTTCCAAGACACAAACATGA/AAAGTCCAGTTTCTCGTTGA
integrin β4	GCAGCTTCCAAATCACAGAGG/CCAGATCATCGGACATGGAGTT
TO	GGCAGCGAAGAAGACAAATC/TCGAACAGAATCCAACTCCC
αAT	ACTGTCAACTTCGGGGACAC/CATGCCTAAACGCTTCATCA
ALB	GCACAGAATCCTTGGTGAACAG/ATGGAAGGTGAATGTTTTCAGCA
TGFBR2	GGAAACTTGACTGCACCGTT/CTGCACATCGTCCTGTGG
c/EBPα	TTCACATTGCACAAGGCACT/GAGGGACCGGAGTTATGACA
c/EBPβ	CGTGTACACACGCGTTCAG/CTCTCTGCTTCTCCCTCTGC
HNF6	CAAACCCTGGAGCAAACTCAA/TGTGTTGCCTCTATCCTTCCC
HNF1β	ACCAAGCCGGTCTTCCATACT/GGTGTGTCATAGTCGTCGCC
CYP2D6	CTTTCGCCCCAACGGTCTC/TTTGGAAGCGTAGGACCTTG
TIR	TCATCGTCTGCTCCTCCTCT/AGGTGTCATCAGCAGCCTTT
HNF1α	AACACCTCAACAAGGGCACTC/CCCCACTTGAAACGGTTCCT
CYP3A4	AAGTCGCCTCGAAGATACACA/AAGGAGAGAACACTGCTCGTG
mouse αAT	TTGCTCGACACAACATGGAAT/ACGTCCCAGTTGACATCTCT
mouse CYP7A1	GCTGTGGTAGTGAGCTGTTG/GTTGTCCAAAGGAGGTTCACC
mouse AQP1	AGGCTTCAATTACCCACTGGAACTTTGGGCCAGAGTAGCGAT
mouse integrin β4	AGAGCTGTACCGAGTGCATC/TGGTGTCGATCTGGGTGTTCT

Supplemental Table 4 The primers used for ChIP assay in this study

	Primers (forward/reverse; 5' to 3')
c/EBP binding site A	TCACAACTTTCTAAGTCCCAATTTT/ACTGAGGCAGGGACTGTGTC
c/EBP binding site B	AACTGAAATGTCTTCCTTTTTCAA/CAGGAGGAGTAGAGCCAGCA
c/EBP binding site C	GCCACATTGTGTTTTCAGGA/ TTAGCCGAGAATGATGTCACC
c/EBP binding site D	CCAGAGGGCTGTACAGAATCA/ CCAGATTTGCCCAAGACATT
c/EBP binding site E	TGCCTACTGGGTGCTAGAGG/AACCTTCAGAGACAGCGATCA
β-actin	CCGGCGGGGTCTTTGTCTGACC/GGGCCGGCCGCGTTATTACCA

OPEN ACCESS Freely available online

PLOS | ONE

Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests

Chioko Nagao[1]*, Nozomi Nagano[2], Kenji Mizuguchi[1]*

1 National Institute of Biomedical Innovation, Ibaraki, Osaka, Japan, 2 Computational Biology Research Center, AIST, Koto-ku, Tokyo, Japan

Abstract

Determining enzyme functions is essential for a thorough understanding of cellular processes. Although many prediction methods have been developed, it remains a significant challenge to predict enzyme functions at the fourth-digit level of the Enzyme Commission numbers. Functional specificity of enzymes often changes drastically by mutations of a small number of residues and therefore, information about these critical residues can potentially help discriminate detailed functions. However, because these residues must be identified by mutagenesis experiments, the available information is limited, and the lack of experimentally verified specificity determining residues (SDRs) has hindered the development of detailed function prediction methods and computational identification of SDRs. Here we present a novel method for predicting enzyme functions by random forests, EFPrf, along with a set of putative SDRs, the random forests derived SDRs (rf-SDRs). EFPrf consists of a set of binary predictors for enzymes in each CATH superfamily and the rf-SDRs are the residue positions corresponding to the most highly contributing attributes obtained from each predictor. EFPrf showed a precision of 0.98 and a recall of 0.89 in a cross-validated benchmark assessment. The rf-SDRs included many residues, whose importance for specificity had been validated experimentally. The analysis of the rf-SDRs revealed both a general tendency that functionally diverged superfamilies tend to include more active site residues in their rf-SDRs than in less diverged superfamilies, and superfamily-specific conservation patterns of each functional residue. EFPrf and the rf-SDRs will be an effective tool for annotating enzyme functions and for understanding how enzyme functions have diverged within each superfamily.

Citation: Nagao C, Nagano N, Mizuguchi K (2014) Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests. PLoS ONE 9(1): e84623. doi:10.1371/journal.pone.0084623

Editor: Valerie de Crécy-Lagard, University of Florida, United States of America

Received June 27, 2013; **Accepted** November 15, 2013; **Published** January 8, 2014

Funding: This study was supported by the Industrial Technology Research Grant Program in 2007 (grant number 07C46056a) from New Energy and Industrial Technology Development Organization (NEDO) of Japan, Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology (grant numbers 25430186 and 25293079) and from the Ministry of Health, Labor, and Welfare to K.M., and also by Grant-in-Aid for Publication of Scientific Research Results (grant numbers 238048 and 248047) from Japan Society for the Promotion of Science (JSPS) to N.N. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: chio@nibio.go.jp (CN); kenji@nibio.go.jp (KM)

Introduction

Almost all chemical reactions in living organisms are catalyzed by enzymes [1]. For a thorough understanding of cellular processes, it is essential to determine enzyme functions, i.e., what types of reactions are catalyzed, and what chemical compounds are utilized as substrates or cofactors. Prediction of enzyme function is a longstanding problem and many methods have been developed. The targeted functional details range from the broadest classification level such as enzyme/non-enzyme discrimination to a highly specific scheme such as the four-digit Enzyme Commission (EC) numbers [2]. Also, different types of features have been used, such as sequence/structural similarities, physico-chemical properties of amino acids, specific sequence/structural motifs, and their combinations [3–12]. Furthermore, many methods have been proposed recently for large-scale prediction of protein functions defined by Gene Ontology (GO) terms [13]. However, the most widely used method for functional annotation remains the simplest one: the transfer of functions based on sequence similarity calculated by BLAST/PSI-BLAST [14,15], despite its known limitations [16–19]. Moreover, predicting a precise enzyme function is still a significant challenge, as only a few methods currently available can predict the full four-digit EC numbers. The knowledge of such detailed functions can help determine true substrates for disease-related enzymes and design specific inhibitors for drug targets.

Enzymes in a protein family are considered to be evolutionary related. In many cases, these enzymes have similar but different functions. Divergence of sequences and functions are different in each family. Some enzymes, which share the sequence identity of over 90%, have different functions and differ in the first-digit of their EC numbers [16–19]. On the other hand, some enzymes, the sequence identity of which is below 30%, share all four digits of the EC numbers. This nonlinear correlation between function and sequence similarity makes the identification of detailed functions of enzymes such a difficult task.

One solution to overcome this problem is to use the information about functionally critical residues. The construction and use of sequence motifs can be considered an example of this approach [20,21]. Residues critical for functions, mutations of which bring drastic changes in the catalytic efficacy or substrate specificity, are sometimes called specificity determining residues (SDRs) or

function determining residues (FDRs). Proper information about SDRs is expected to improve the ability to distinguish enzyme functions [22–24]. However, such information is limited, because SDRs are determined by mutagenesis experiments. Therefore, most prediction methods use other properties serving as a proxy for SDRs [4,6,23–26]: catalytic residues, ligand binding sites or residues conserved in a functional subfamily. The lack of information about SDRs has hindered the development of computational methods for identifying SDRs [27–30] as well as predicting detailed functions.

Some machine learning methods can construct classifiers from a large number of attributes and calculate contributions from each attribute. Random forests [31] are one of the most accurate machine learning algorithms used for many applications, including the analysis of microarray data [32,33] and prediction of protein-protein interactions [34,35]. For enzyme function prediction, random forests have been applied for assigning the first or second digit of the EC numbers [7,8,36,37]. These methods used several hundreds of physico-chemical features calculated from only the full-length sequences and thus, provided no information about the importance of each residue for discriminating different functions.

In this study, we applied random forests, for the first time, for predicting the four-digit EC numbers (rather than only the first or second digit) in each homologous superfamily and also for obtaining a putative set of SDRs at the same time by using residue position specific attributes. We focus on a problem of discriminating detailed enzyme functions within a single protein family, since methods for assigning a protein sequence to an existing family have been well established. Thus, we assume that a functionally unknown protein has been already classified into a known protein family by sequence similarity. Given this framework, our objectives were two-fold; first, we aimed to develop a method that can predict the full four-digit EC number for a given protein. Second, we aimed to define putative SDRs as the most highly contributing positions used in our prediction model. Characterizing these "computational defined SDRs" in a systematic manner should mitigate the lack of experimentally defined SDRs.

Our analysis is based on the CATH domain classification [38]; we created a dataset from the UniProtKB/Swiss-Prot database [39] by selecting the enzymes, which had complete four-digit EC numbers and for which CATH homologous superfamilies were assigned by Gene3D [40]. For each enzyme in each superfamily, binary predictors were constructed by random forests with full-length sequence similarities and the residue similarities for active sites, ligand binding sites and conserved sites as input attributes. From the most highly contributing attributes, we obtained a set of putative SDRs and termed them random forests derived SDRs (rf-SDRs). The predictors (EFPrf) showed a performance comparable to that of a related method currently available and the rf-SDRs included many residues, for which functional importance had been verified by experimental studies. This study revealed a general tendency that functionally diverged superfamilies tend to include more active site residues (ASRs) in their rf-SDRs than in less diverged superfamilies. From the analysis of selected superfamilies, we also made superfamily-specific observations that conserved residues across enzymes, even if functionally important, tend not to be selected as rf-SDRs.

Results and Discussion

Overview of the enzyme function prediction

Figure 1A describes an overview of the enzyme function prediction method by random forests (EFPrf). A query to the system is a domain sequence pre-assigned to a CATH homologous superfamily (indicated as CATH X.X.X.X in the figure) by Gene3D. We chose a CATH homologous superfamily as a unit of protein family because a structure-based classification scheme can capture more distant proteins than a sequence-based one. In CATH X.X.X.X superfamily, binary predictors for each enzyme have been developed (Figure 1B). In each predictor, the query is aligned to the representative sequence by the FUGUE software [41] with the structure environment-specific substitution tables (ESSTs). Based on the alignment, the similarity scores for the full-length sequence and at the functional sites are calculated for the input to the predictor.

Dataset construction

We selected the enzyme sequences from the UniProtKB/Swiss-Prot database, for which complete EC numbers are assigned, and obtained their CATH domain regions from the Gene3D database. After removing redundancies, predictors have been constructed for the enzymes that had ten or more sequences and had at least one other enzyme in the superfamily (with a total of ten or more sequences) as negative data (Figure 2; see Materials and Methods for more details). Thus, we have built predictors for 1121 enzymes distributed over 306 CATH superfamilies. The representative structures for each enzyme were selected from the CATH S-level representatives with the longest sequence length and the highest resolution. In each superfamily, 3.7 enzymes were selected for constructing predictors on average. In 89 superfamilies, a single predictor was constructed. Fifteen superfamilies contained more than ten enzyme predictors and the largest superfamily was the NAD(P)-binding Rossmann-like domain superfamily (CATH 3.40.50.720) with 65 predictors (Table S1 and Figure S1). All the superfamilies, for which at least one predictor was created, were included in the analysis below.

Additional information to BLAST score improved the precision of the prediction

To investigate whether the use of the information about functional residues improves prediction performance or not, we built two types of predictors. First, we created simple decision trees by C4.5 with the BLAST bit score for the top hit in each enzyme as an attribute ("the simple model"). Because BLAST scores are the most widely used measure for function transfer, the simple model served as our baseline for predicting enzyme functions. Next, we constructed a second set of predictors by random forests (EFPrf) with more attributes. Three scoring matrices, BLOSUM62 [42], position specific scoring matrices (PSSM) [43] and ESST-based structural profiles, were used to calculate the scores at the active site residues (ASRs), ligand binding residues (LBRs) and conserved residues (CSRs), in addition to the full-length scores. The resulting 12 ($=3\times4$) attributes and the BLAST score were used as input to the system.

In a cross-validated benchmark assessment (see Materials and Methods), we followed a previous study [4] and calculated the maximal test to training sequence identity (MTTSI) for each query, and evaluated the prediction performance for eight different MTTSI ranges separately. Figure 3 and Table S2 show recall and precision averaged in each of the eight MTTSI ranges. (The average was taken by using only the enzymes, for which precision or recall was defined in the given MTTSI range.) In Figure 3A, recall in all ranges shows no significant differences between the simple model and EFPrf. On the other hand, precision improved significantly by EFPrf, especially in the lowest MTTSI range, where distinguishing functions by sequence similarity alone is known to be difficult (Figure 3B). This result

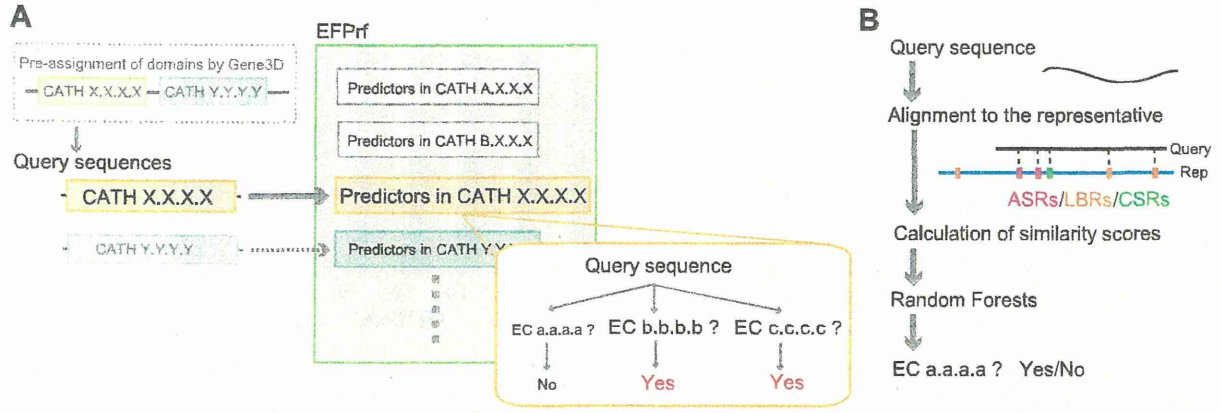

Figure 1. Outline of the EFPrf system (A) and the predictor for each enzyme constructed by Random Forests (B). A query to the system is a domain sequence pre-assigned to a CATH homologous superfamily by Gene3D. For each CATH superfamily, binary predictors, each for a known enzyme, process the query and return their results (A). In each predictor, the query is aligned to a representative sequence by the FUGUE software. Based on the alignment, similarity scores for the full-length sequence and at the functional sites are calculated for the input to the predictor (B).
doi:10.1371/journal.pone.0084623.g001

indicates that the additional information about functionally important residues is useful for discriminating detailed functions. Table 1 shows the prediction performance averaged over the 1121 enzyme predictors (see Table S3 for the individual values). Although a general trade-off between recall and precision was observed, the statistically significant increase in the F-measure achieved by EFPrf over the simple model also suggested the usefulness of the additional attributes of ASRs/LBRs/CSRs.

Because of differences in the training and test datasets, a direct comparison of performance with other methods is difficult but the prediction performance of EFPrf (recall = 0.30, precision = 0.78 in MTTSI <30%) is comparable to or better than that of EFICAz^2 [4,5] (recall = 0.23, precision = 0.74 in MTTSI <30%), which combines FDRs recognition, sequence similarity and support vector machine (SVM) models. Moreover, EFICAz^2 and EFPrf achieved an average precision of above 0.9 for MTTSI ≥40%, which is considered to be a "non trivial achievement" [4,17].

General properties of the random forest derived SDRs

In constructing the EFPrf, importance scores for each attribute were also calculated. We selected the top $3\times\sqrt{n}$ attributes as "highly contributing attributes", where n is the number of input attributes for each enzyme, and defined the residue positions in the highly contributing attributes (except for the full-length sequence similarity score) as the "random forests derived SDRs" (rf-SDRs) (Table S4). (In all enzymes, the full-length sequence similarity score was included in the highly contributing attributes, consistent with the result that the simple model was a modestly successful predictor.) On average, 8.4 residue positions were selected as the rf-SDRs for each enzyme. Among the position specific attributes calculated with different scoring matrices, the most frequently selected were those with PSSMs, suggesting that PSSMs may represent the amino acid differences among enzymes having similar structures/functions more clearly than the other scoring matrices (Table S5).

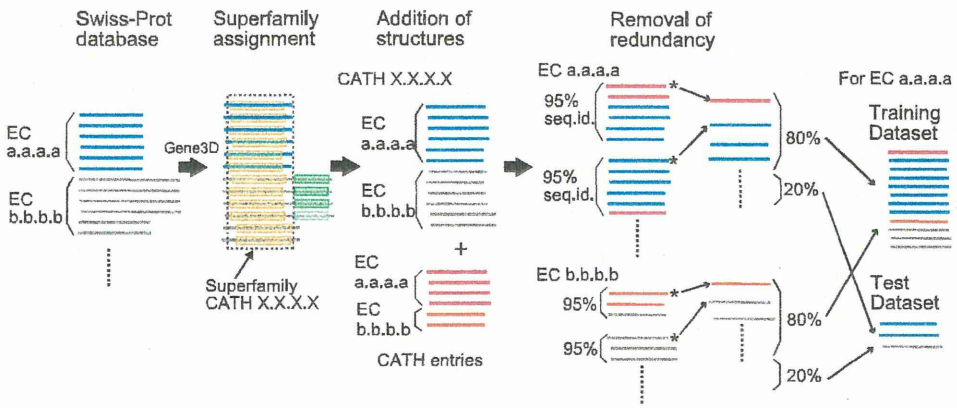

Figure 2. Outline of dataset construction. From the UniProtKB/Swiss-Prot database, the enzyme sequences, for which complete EC numbers are assigned, were obtained and their CATH domain regions from the Gene3D database were selected. After adding CATH entries and removal of redundancies, the enzymes having less than ten sequences were removed. The representative structures for each enzyme were selected from the CATH S-level representatives. In the remaining sequences, a predictor was constructed for an enzyme, which has sufficient numbers of positive and negative sequences (see Materials and Methods for more details). Randomly selected 80% of the sequences were used for training. The remaining 20% of the sequences were used as a test dataset.
doi:10.1371/journal.pone.0084623.g002

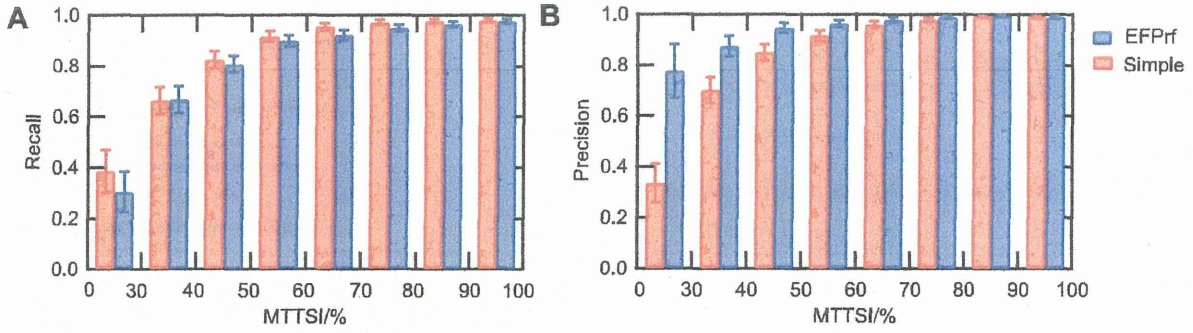

Figure 3. Prediction performance of EFPrf. The recall (A) and precision (B) at each level of the maximal test to training sequence identity (MTTSI) are plotted for the simple model (red) and the EFPrf (blue). Error bars represent 95% confidence intervals in each MTTSI range.
doi:10.1371/journal.pone.0084623.g003

Figure 4 shows the amino acid propensity for the rf-SDRs. The propensity of amino acid *i* was obtained as the fraction of amino acid *i* in the rf-SDRs divided by the fraction of amino acid *i* in all representative enzyme domains. In general, polar or charged residues were overrepresented in the rf-SDRs and non-polar residues were underrepresented. In polar, aromatic and charged residues, Trp, Tyr, Cys, Asn, Arg and His had a particularly high propensity value and in non-polar hydrophobic residues, Ala, Val, Leu and Ile had a low propensity value. In charged residues, Lys and Glu were underrepresented. This biased distribution of charged residues suggests that the delocalized charge in the guanidino group of Arg may be better utilized for SDRs than the charge in Lys, as observed in protein-protein interactions [44], and that the short side chain of Asp, with a smaller degree of freedom than that for Glu, is more suitable to form specific interactions. Some of the propensity values are different from those observed in the Catalytic Site Atlas (CSA) [45]; Asn favored for non-catalytic sites in the CSA [46], was overrepresented in the rf-SDRs and Lys and Glu, favored for catalytic sites in the CSA, were underrepresented. These differences are likely due to different definitions of functional residues, because the rf-SDRs were selected from not only catalytic sites but also ligand binding and conserved sites.

To analyze the relationships between functional diversity and the residues important for distinguishing functions, we classified superfamilies based on the functional entropy, defined by using the number of distinct EC numbers up to the third- and forth-digit levels (see details in Materials and Methods; Table S6). In the third-digit level classification, the three classes defined, the low-, medium- and high-degrees of functional diversity, approximately corresponded to having one, two to four, and more than four distinct EC numbers at the third-digit level within each superfamily. In the fourth-digit level classification, the low-, medium- and high-degrees of diversity corresponded to having one to five, six to ten and more than ten distinct EC numbers at the fourth-digit level within each superfamily. The prediction performance for the most diverged class was shown to be lower than that for the other classes in both the third- and fourth-digit based classification schemes (Tables S7 and S8).

Table 1. Prediction performance.

Model	Precision	Recall	F-measure
Simple	0.94	0.91	0.92
EFPrf	0.98 (<2.2e-16)	0.89 (1.3e-5)	0.93 (0.009)

The values in the parentheses represent the p-values calculated against the simple model by paired t-test.
doi:10.1371/journal.pone.0084623.t001

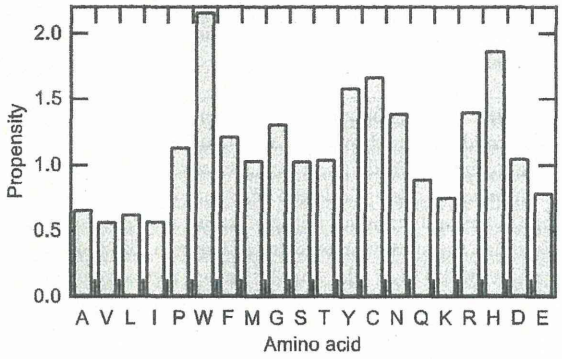

Figure 4. Amino acid propensities for the rf-SDRs. The propensity of amino acid *i* was calculated as the fraction of amino acid *i* in the rf-SDRs divided by the fraction of amino acid *i* in all representative enzyme domains.
doi:10.1371/journal.pone.0084623.g004

We then decided to examine what proportion of the ASRs or LBRs were selected as rf-SDRs in each superfamily. We excluded the CSRs from this analysis, because the ASRs and LBRs should be more directly linked to enzyme functions, whereas the identification of CSRs depended on the number of available sequences. If we consider all the superfamilies, the rf-SDRs included either no ASRs, about half of them or all of them (corresponding to peaks at zero, 0.5 and one in Figure S2), while in many superfamilies, about half of the LBRs were selected to be rf-SDRs (a peak around 0.5). We next examined these quantities as a function of functional diversity. Figure 5 and Table S9 showed that the proportion of ASRs to be selected as rf-SDRs increased with functional diversity, as defined by numbers of the third-digit EC number level functions. Although this tendency was weak (with moderate statistical significance for the difference; p-value = 0.019 for the superfamilies with low and medium functional diversity, and p-value = 0.017 for those with low and high functional diversity by the Wilcoxon rank sum test), it is consistent with the notion that enzymes in a superfamily with low functional diversity often have similar active sites and similar catalytic mechanisms and thus, ASRs generally do not distinguish different functions. On the other hand, the proportion of LBRs to be selected as rf-SDRs decreased slightly from medium to high functional diversity