Supporting Information

Tseng and Li 10.1073/pnas.1119684109

SI Text

Selecting the Surface Attributes of a Protein. To characterize a protein surface, we used the surface attributes, which are geometrical and physicochemical features, including the number of amino acid residues on the surface pocket, solvent-accessible area, and shape of the surface (Table 1). We focused on the residue compositions on the binding sites, captured their topological shapes, and analyzed their physicochemical properties in a large-scale manner. For the 28,986 identified functional surfaces, we used the surface attributes to study functional diversity. Assessing these attributes in a high-dimensional space allows us to characterize a protein in an effective manner.

To integrate the surface attributes, we present protein global and local surfaces in a scatter matrix (i.e., a display of related plots) that depicts 2D values for a set of attributes. In Fig. S2, the plots along the diagonal show the attribute distributions, and the display gives a comprehensive view of the properties of protein surface in relation to one another. For example, the functional surface of a protein has a mean number of 31 residues but its shape is diverse. Most shapes are angular because the mean sphericity (*sph*) is only 0.57. The larger binding surfaces are closer to the mass center as shown in the *len-d* plot, where *len* refers to length (i.e., the number of residues in a pocket) and *d* measures the anisotropic distance (1) (*SI Geometrical Features*). However, they usually have a lower value of sphericity, indicating a weaker compactness in the *len-sph* plot. Moreover, the total area occupied by hydrophobic (apolar) residues on the binding surface has a mean of 960.51 Å^2, whereas that of a hydrophilic (polar) surface has a smaller area of 586 Å^2. The ratio of hydrophobic to hydrophilic area is ~10:6 on a typical binding surface of ~1,600 Å^2. For the attribute distributions, a panel of kernel-density plots, for example, *len*, *Wsph*, and *len-Wsph* in Fig. S3, where *Wsph* denotes the global sphericity of a protein, can be used to capture the shape and texture of a protein surface. Figs. S2 and S3 also reveal the complexity of protein surfaces because of weak global trends and wide ranges of attribute values. Although attribute distributions in general are Gaussian (Fig. S2), distinct surface types may share overlapped values of surface attributes, potentially increasing the challenges of surface classification.

Determining the Number of Subtypes in a Surface Type. One way to estimate the number of subtypes is to use the number of modes of a distribution as an initial estimate of the number of subtypes n. As an example, we computed the pairwise distances of the binding surfaces of 41 oxidoreductases to obtain a histogram and approximated it by a continuous distribution (Fig. S4*B*). The number of modes of the distribution can be taken as the initial estimate of the number of subtypes. For example, the number of modes in Fig. S4*B* is 5, so an initial estimate of n is 5. Then, using an automated progressive search, we obtained an optimal value of $n = 3$ when the Tanimoto coefficient between this classification and its actual Enzyme Commission (EC) (2) annotations reaches the maximum value (see *SI Materials and Methods* for the definition of the Tanimoto coefficient). Similar to a hierarchical clustering, this subgrouping approach is to specify the number of surface subtypes for obtaining an optimal classification in terms of function. In Fig. S4*C*, we cluster the 41 oxidoreductases into three subtypes ($n = 3$) because it gives an optimal Tanimoto coefficient. After the surface subtypes are determined, we can compute the equatorial radii of the ellipsoid boundaries that separate different surface subtypes. This basic inference naturally gives rise to a classification approach in which a specific ellipsoid contains as many related members as possible. Therefore, identifying the effective attributes of a functional surface is crucial for characterizing a functional surface.

SI Materials and Methods

Collecting the Functional Pockets of Bound Structures. We collected ~68,000 structures from the Protein Data Bank (PDB) that were then grouped into a bound group and an unbound group. The surface of a functional pocket (3, 4) of a bound form is "canonical," because it has a standard (fixed) shape associated with its binding substrate and molecular function. This is important, because the canonical shapes of two functional pockets can be superimposed and then compared in a reliable manner. The functional surfaces of bound forms have been identified previously (5, 6). In this study, a total of 28,986 functional pockets of bound forms were collected and used for the surface classification based on pairwise shape similarities.

Geometric Matching to Compare the Shapes of Two Pockets. In shape analysis, the two aligned pocket fragments are superimposed to calculate the atomic coordinate root-mean-square deviation (rmsd). Because the pocket residues of a fragment are sequence-ordered, we used the nonpermutable measure of rmsd to assess the shape similarity between two pockets (http://pocket.uchicago.edu/fpop).

Denote the kth pocket sequence by $S_k = (r_1, r_2, \ldots, r_p)$, where p is the number of residues in the pocket sequence and $r_i \in \Re^3$ represents the coordinates of the C_α in the ith residue in the pocket sequence. Denote the k'th pocket sequence by $S'_{k'} = (r'_1, r'_2 \ldots\ldots, r'_q)$. Let n be the length of the superimposed alignment and denote the aligned subsequences as $\{g_i\}$ and $\{g'_i\}$. Let R be an orthonormal matrix for the linear transformation R: $\Re^3 \rightarrow \Re^3$. The rmsd between S_k and $S'_{k'}$ is minimized by optimizing the transformation matrix R (7), using the singular value decomposition, so that R is represented by a form of translation-rotation-translation 4×4 matrix:

$$R = (r_{ij}) = \begin{bmatrix} a & b & c & u_x \\ d & e & f & u_y \\ g & h & i & u_z \\ v_x & v_y & v_z & 1 \end{bmatrix},$$

where $r_{ij} \in \Re$ and $i, j \in [1,2,3,4]$, $\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$ is a 3×3 rotation matrix, (u_x, u_y, u_z) is a translation vector before rotation, and (v_x, v_y, v_z) is a translation vector after rotation.

Surface Comparisons by Geometric Matching. We conducted an all-against-all surface comparison to assess the structural relationships between any two functional pockets (surfaces), using the technique of geometric matching described above. Essentially, these structural relationships yield a comprehensive network to increase the knowledge of surface classification. Therefore, we computed the surface rmsds in an exhaustive manner. We selected all rmsds with a significant P value (smaller than a defined threshold) and then ranked rmsds by their P values. Subsequently, we constructed a look-up table of pairwise relationships for each surface. After computing all tables of surface similarity, we were able to effectively cluster similar surface types with an agglomerative technique.

Clustering Algorithm of a Coarse Surface Classification. Denote by F the set of members, R the set of all possible pairwise relationships, and T the set of surface classifications (functional types) in the

SplitPocket database (5). Let m_i be the center by measuring the highest degree of significant structural relationships with the smallest mean rmsd within the same protein type T_i. T_i is created with m_i whenever $m_i \notin T_k$, $k = 1 \cdots i-1$. Subsequently, each T_i is represented by m_i. The center of a functional type can be found as follows.

For a functional surface s_q in T_i, define the local similarity D_q as $D_q = \frac{\sum_{t=1}^{n_q} RMSD(s_q, s_t)}{n_q}$, where $s_t \in T_i$, $RMSD(s_q, s_t)$ is the rmsd value between s_q and s_t, and n_q is the degree of graph connectivity of s_q within type T_i. By comparing D_a with D_b for all b in T_i, one can obtain the center $m_i = s_q$ with D_q the smallest among the D_a's for every member a in type T_i.

Surface Characteristics of a Functional Pocket. ***Geometrical features.*** For each functional pocket, we computed its residue composition, surface-accessible area A, and molecular volume V at the atomic level (8, 9). Then, the shape of a pocket can be depicted by the sphericity, defined as $\Psi \equiv \frac{\pi^{1/3}(6V)^{2/3}}{A}$. Note that Ψ is sensitive to the measurements of V and A, which can be analytically computed by an exact algorithm (4, 9). A perfect round shape (i.e., sphere) has the highest value of $\Psi = 1$, the octahedron has $\Psi \sim 0.864$, whereas an angular shape such as a tetrahedron has a smaller value of $\Psi \sim 0.671$.

We also computed the anisotropic shape proposed by Nicola and Vakser (1), which is a measure of the distance $d_k = |\Delta_k - \Delta_{mc}|$ between the mass center Δ_k of a surface and that Δ_{mc} of the protein P. The centers of mass can be easily computed by

$$\Delta_k \equiv \frac{\delta_i \cdot (x_i, y_i, z_i)}{\sum_{i \in k} \delta_i},$$

$$\Delta_{mc} \equiv \frac{\delta_i \cdot (x_i, y_i, z_i)}{\sum_{i \in P} \delta_i}$$

where δ_i is the weight of atom type $i \in \{C, N, O, S\}$ and (x_i, y_i, z_i) is its corresponding 3D coordinates. In a large-scale study of proteins, we found that the functional pocket has the shortest distance $|\Delta_k - \Delta_{mc}|$. In addition, the weight of a surface is defined as $SD_k \equiv \frac{\Delta_k}{A_k}$ to compute a surface density.

In addition to the surface characteristics of a functional pocket, we followed Ballester and Richards (10) to compute the moments μ_k of a distribution, which can be used to depict the asymmetry of a protein shape and the outliers of a large set of atoms. The kth moment is defined as

$$\mu_k \equiv E\left((X - \mu)^k\right),$$

where μ is the mean of the variable X. Specifically, the third standardized moment is the skewness and the fourth standardized moment is the kurtosis, which are used to obtain an initial assessment of the protein shape for screening the similarity of surface characteristics. The empirical distribution X of the distances from N atoms to the mass center of a protein is calculated as $X = (x_1, x_2, \ldots, x_N)$. Skewness g_1 and kurtosis g_2 ($g_1, g_2 \in \Re$) are computed by

$$g_1 \equiv \frac{\mu_3}{\sigma^3} = \frac{\sum_{i=1}^{N} (x_i - \mu)^3}{(N-1)\sigma^3},$$

$$g_2 \equiv \frac{\mu_4}{\sigma^4} - 3 = \frac{\sum_{i=1}^{N} (x_i - \mu)^4}{(N-1)\sigma^4} - 3,$$

where σ is the sample standard deviation.

Skewness is a measure of the asymmetry of protein shape. The skewness of a perfectly symmetric distribution of atoms is 0. Kurtosis is a measure of how outlier-prone a distribution of atoms is, that is, a measure of peak and tail of the distribution. The kurtosis of the normal distribution of atoms is 3; a protein shape with a high value of kurtosis tends to have a segment(s) away from the center of mass.

We computed these geometrical attributes as a shape profile for each protein. Using the shape profile, we are able to characterize a protein surface and relate one surface to another.

Physicochemical features. The solvent-accessible area of a residue is classified into hydrophobic (apolar) and hydrophilic (polar). The solvent-accessible areas of a residue at the atomic level are analytically computed by the Volbl package (4, 9), so that they can be divided into hydrophobic and hydrophilic with respect to the whole structure (global) and the functional pocket (local). This calculation of solvent accessibility is highly accurate for separating surface areas to assess hydrophobicity strength and charge concentration. The ratio of hydrophobic to hydrophilic areas gives a means of surface comparison. That is, a comparison of surface hydrophobicity between two proteins can reveal similar functions across superfamilies.

Evolutionary features. The residue compositions of a pocket sequence domain (PSD) (11) are used to assess protein divergence in terms of sequence identity. Moreover, the evolutionary conservation of a functional pocket is computed with a surface conservation index (SCI) ranging from 0 to 1 (5). In a protein, the functional pocket usually has the highest value of SCI, including 85% of all functional surfaces. Ranking the SCI values of the putative pockets on a protein surface, we are able to filter out nonfunctional pockets and identify the canonical surface of a protein.

Furthermore, biological information such as the structural positions of catalytic residues from UniProt-KB/SwissProt (12) is mapped onto a functional pocket. It gives support to our claim that our predicted pockets are actually functional. In a systematic manner, the geometrical, biological, physicochemical, and evolutionary features of a functional pocket are extracted from its structural coordinates and then recorded as structural attributes. The selected features of the functional pocket of a bound structure are accessible at SplitPocket (5) (http://pocket.uchicago.edu) and PSD (http://pocket.uchicago.edu/psd).

Cosine Similarity and Tanimoto Coefficient. We used the cosine transformation to compare the two sets of surface attributes. In this way, we are able to assess the structural similarity of two surfaces. Denote the n attributes of surface type T_1 by $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and those of surface type T_2 by $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$. Define the similarity between T_1 and T_2 as $\cos\theta = \frac{\alpha \cdot \beta}{\|\alpha\|\|\beta\|}$, where $\cos\theta \in [-1,1]$ and α,β are attribute vectors.

Moreover, the cosine measure is extended to the Tanimoto coefficient (13) when attributes are binary. The Tanimoto coefficient is defined as $\Gamma = \frac{N_c}{N_a + N_b - N_c}$, where N_a is the number of properties in surface a, N_b is the number of properties in surface b, and N_c is the number of properties in the intersection set c (i.e., number of matched properties).

Assessment of Statistical Significance for a Functional Surface Alignment. Because the functional surface alignment computed by the Smith–Waterman algorithm follows the model of extreme value distribution (14), it allows estimating the statistical significance of functional similarity scores at P values.

$$p(Z > z) = 1 - \exp\left(-e^{\frac{z\cdot\pi}{\sqrt{6}} - \Gamma'(1)}\right),$$

where an observed z value is calculated by $z = \frac{(\Phi' - \mu)}{\sigma}$ through μ and σ, which represent the mean and standard deviation of sampled scores Φ', respectively.

The empirical estimation was performed by matching a query against the basic set of 28,986 functional surfaces. Sampled μ and σ are estimated for each query after a matching process. With the equation above, a functional similarity (z value) can be converted to a probability (P value). To evaluate the statistical significance of a functional similarity score, we chose a P value of 5×10^{-4} as the threshold.

Performance Evaluation. Our classification focuses on the characteristics of functional surfaces associated with protein function. To evaluate performance, we used the EC classification as our gold standard because protein structures with EC annotations have been experimentally verified with a panel of enzymatic assays. A positive is defined when a classified and unique label matches its EC annotation. Let P be the number of positives and N be the number of negatives. We then compared the results of surface classification with EC classification by constructing a contingency table where TP denotes the number of true positives, TN denotes the number of true negatives. Based on the contingency table, we calculated the accuracy (i.e., rand index) defined as $\frac{TP+TN}{P+N}$, which is identical to an assessment by the Tanimoto coefficient.

1. Nicola G, Vakser IA (2007) A simple shape characteristic of protein-protein recognition. *Bioinformatics* 23:789–792.
2. Webb EC (1992) *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes* (Academic, San Diego).
3. Edelsbrunner H, Facello M, Liang J (1998) On the definition and the construction of pockets in macromolecules. *Discrete Appl Math* 88:83–102.
4. Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci* 7:1884–1897.
5. Tseng YY, Dupree C, Chen ZJ, Li WH (2009) SplitPocket: Identification of protein functional surfaces and characterization of their spatial patterns. *Nucleic Acids Res* 37(Web Server issue):W384–W389.
6. Tseng YY, Li WH (2009) Identification of protein functional surfaces by the concept of a split pocket. *Proteins* 76:959–976.
7. Umeyama S (1991) Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans Pattern Anal Mach Intell* 13:376–380.
8. Edelsbrunner H, Facello M, Fu P, Liang J (1995) Measuring proteins and voids in proteins. *Proceedings of the 28th Annual Hawaii International Conference on System Sciences* (IEEE Comput Soc, Los Alamitos, CA), pp 256–264.
9. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S (1998) Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins* 33(1):1–17.
10. Ballester PJ, Richards WG (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. *J Comput Chem* 28:1711–1723.
11. Tseng YY, Li WH (2011) Evolutionary approach to predicting the binding site residues of a protein from its primary sequence. *Proc Natl Acad Sci USA* 108:5313–5318.
12. Wu CH, et al. (2006) The Universal Protein Resource (UniProt): An expanding universe of protein information. *Nucleic Acids Res* 34(Database issue):D187–D191.
13. Godden JW, Xue L, Bajorath J (2000) Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *J Chem Inf Comput Sci* 40(1):163–166.
14. Binkowski TA, Adamian L, Liang J (2003) Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J Mol Biol* 332:505–526.

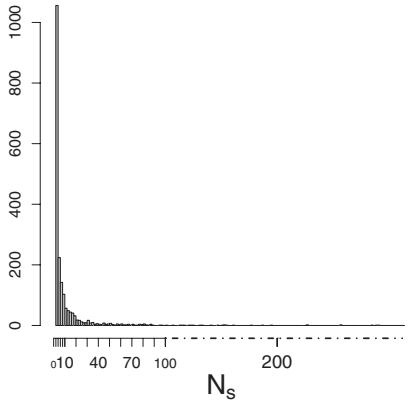

Fig. S1. The distribution of the number of members in a surface type (N_s). In the basic set of 1,974 functional surface types, most surface types have $N_s < 10$.

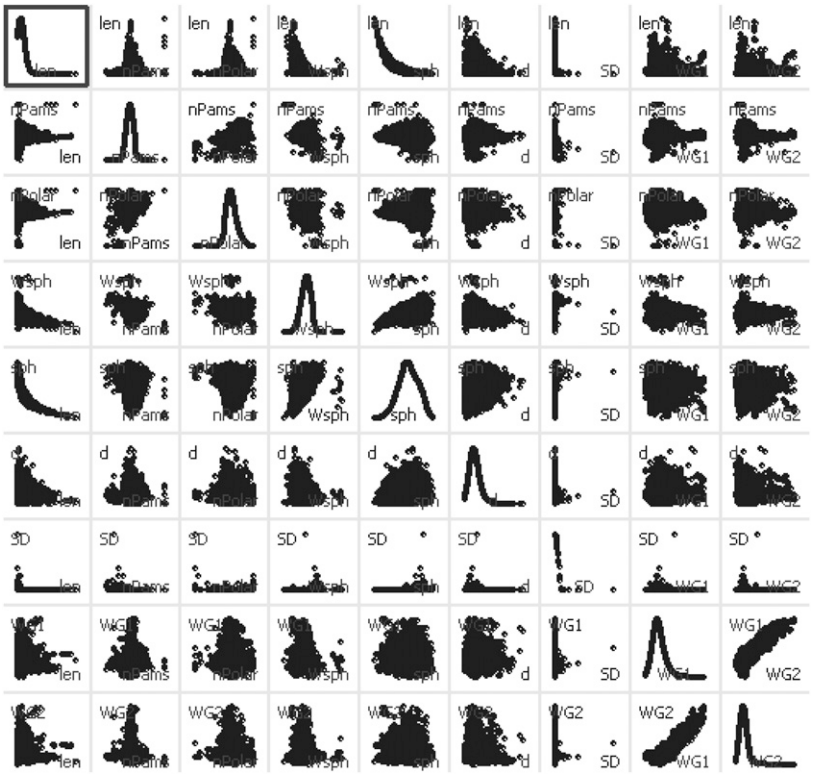

Fig. S2. Distributions of selected geometrical and physiochemical attributes in the 28,986 identified functional surfaces.

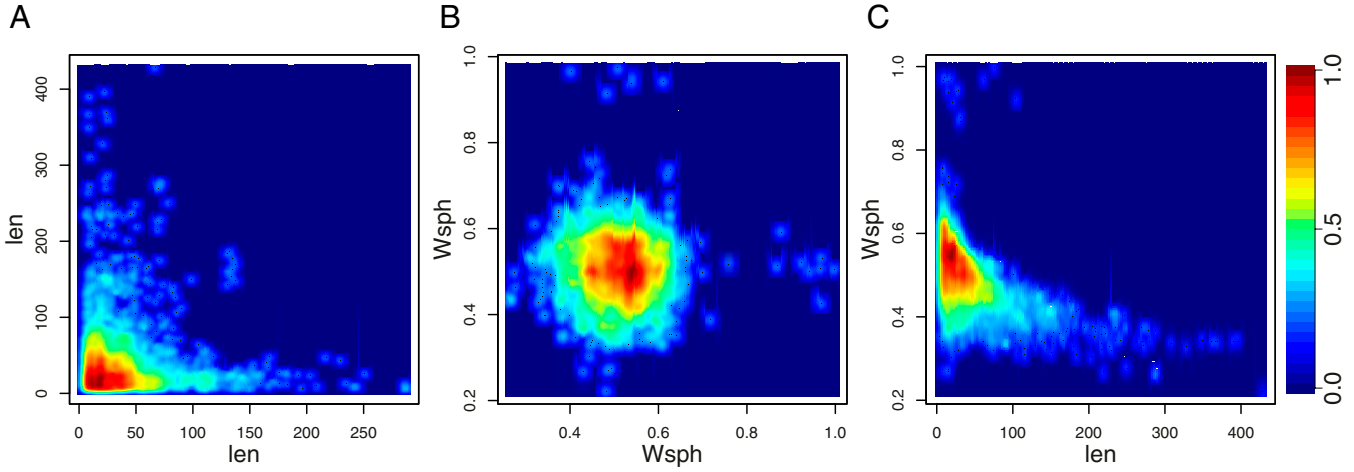

Fig. S3. The kernel densities of the length and global sphericity attributes. The panels highlight higher-density areas (red), the areas colored according to the ratio of the spectrum on the right. A binding surface typically has a mean number of 31 amino acid residues (*len*) (*A*) and a global sphericity (*Wsph*) of 0.51 (*B*). The global shape of a protein with a smaller binding surface tends to be round. (*C*) The correlation between *len* and *Wsph*.

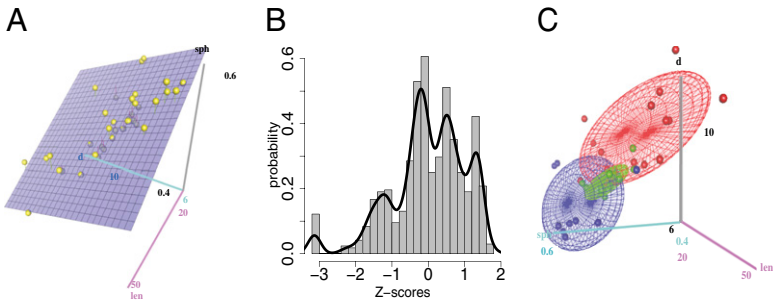

Fig. S4. Classification of surface subtypes of oxidoreductase. (*A*) The 41 oxidoreductases (yellow) lie scattered around the hyperplane (blue) in the plot of the selected attributes *len*, *sph*, and *d*. (*B*) Five distinct peaks appear in the distribution of pairwise distances of binding surfaces. The number of modes (peaks) is used to obtain an initial estimate of the number of surface subtypes. (*C*) The members colored red, green, and blue in EC labels are grouped into three subtypes, each of which is approximately bounded by an ellipsoid when we applied the surface attributes to their function identifications.

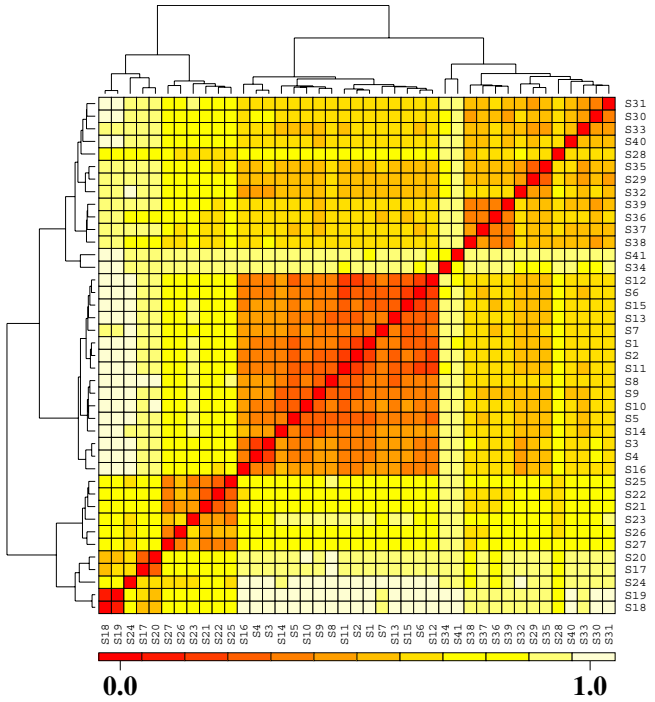

Fig. S5. The shape–function relationships of the bound forms of oxidoreductases. The dissimilarity matrix in a heatmap suggests the subtypes of binding surfaces for a fine classification.

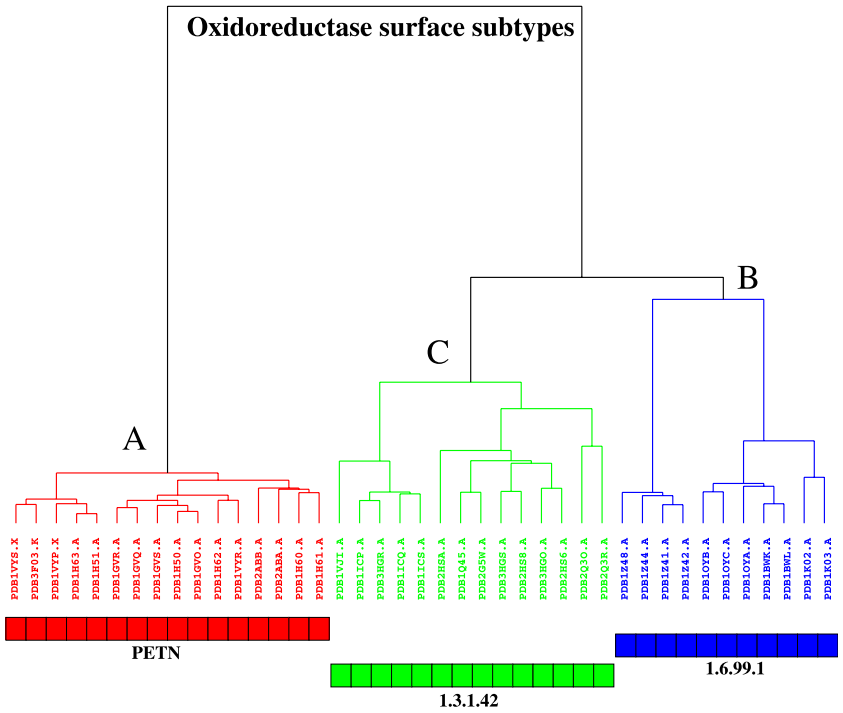

Fig. S6. The three identified subtypes of oxidoreductases exactly match their EC annotations: PETN (pentaerythritol tetranitrate), EC 1.6.99.1, and EC 1.3.1.42. Among the 41 surface members, surface subtype A contains 16 members with no EC assignment, whereas subtypes B and C are composed of 11 and 14 members identified by EC 1.6.99.1 and EC 1.3.1.42, respectively.

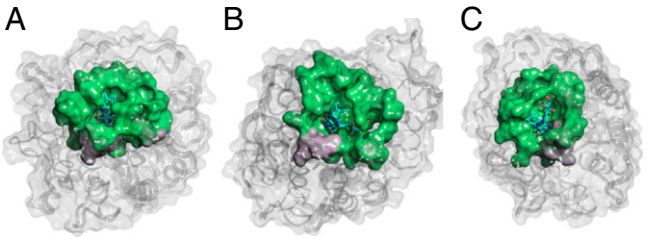

Fig. S7. Subtypes of oxidoreductase across different species with the same fold of Aldolase class I have similar binding pockets (green; http://pocket.uchicago.edu) with key residues (violet) to accommodate the cofactor flavin mononucleotide. The subtype binding surface on (*A*) PDB1gvr.A of *Enterobacter cloacae* (EC not assigned; PETN) contains 33 amino acids with a solvent-accessible area of 675.44 Å^2 and a molecular volume of 874.78 Å^3, whereas that on (*B*) PDB1z41.A of *Bacillus subtilis* (EC 1.6.99.1) has 34 amino acids with a solvent-accessible area of 634.99 Å^2 and a molecular volume of 893.65 Å^3 and that on (*C*) PDB2hs6.A of *Solanum lycopersicum* (EC 1.3.1.42) has 29 amino acids with a solvent-accessible area of 473.50 Å^2 and a molecular volume of 857.85 Å^3.

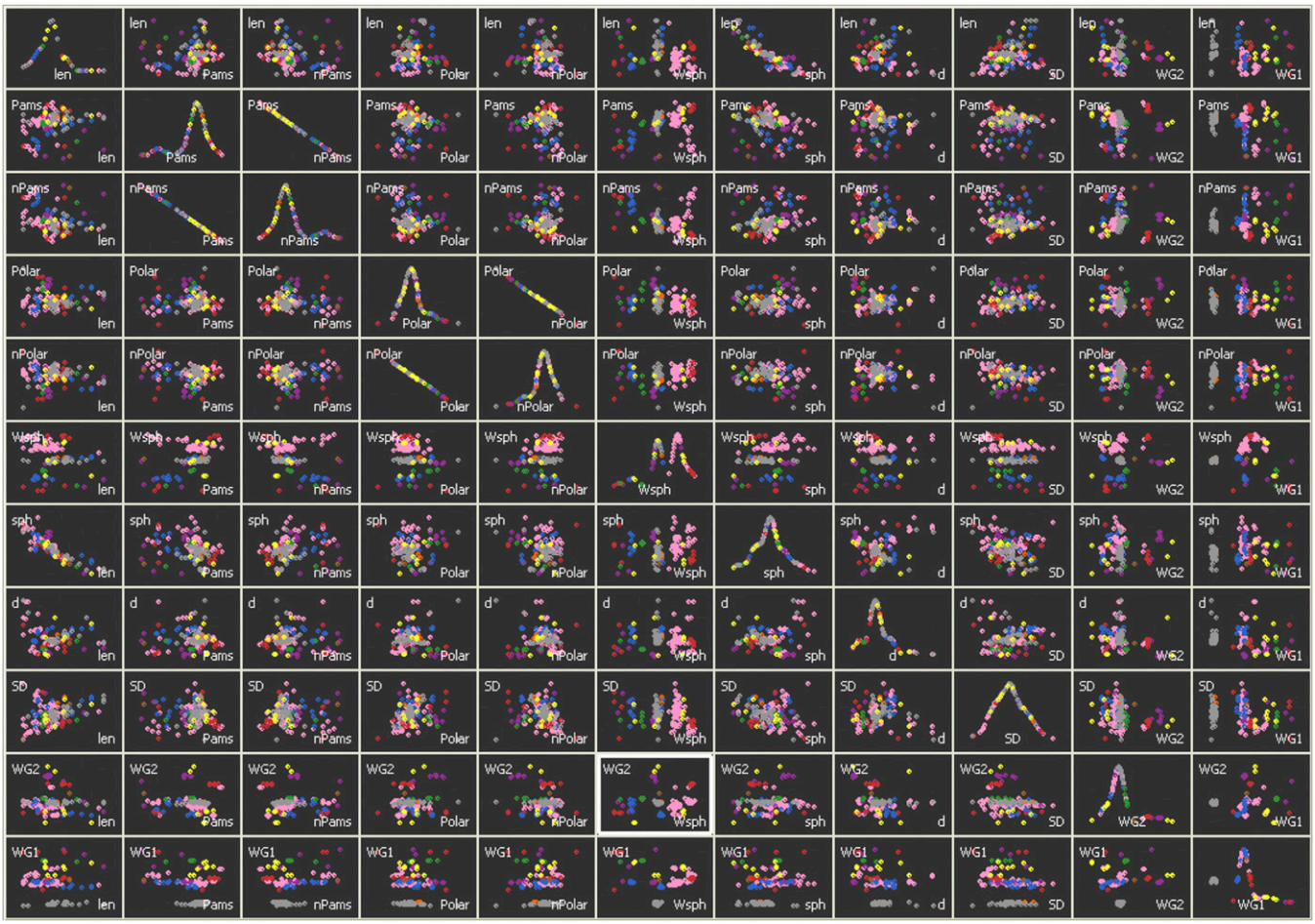

Fig. S8. The scatter matrix of surface attributes for identifying the surface subtypes of glycosidase. The nine EC labels are painted in different colors. In general, members with the same color are clustered in a subtype. High-dimensional attributes are required to subdivide the surface type into subtypes.

Table S1. Surface attributes and EC annotations of oxidoreductases

ID	*len*	*Pams*	*aPams*	*Polar*	*aPolar*	*Wsph*	*sph*	*d*	*SD*	*WG1*	*WG2*	*EC*
1gvr.A	33	0.432	0.568	0.493	0.507	0.560	0.543	8.117	0.687	−0.383	−0.090	PETN
1h50.A	34	0.433	0.567	0.484	0.516	0.564	0.544	8.046	0.686	−0.389	−0.099	PETN
1h60.A	31	0.434	0.566	0.467	0.533	0.560	0.560	8.024	0.785	−0.392	−0.103	PETN
1h61.A	31	0.435	0.565	0.456	0.544	0.556	0.565	8.177	0.770	−0.397	−0.105	PETN
1h62.A	35	0.430	0.570	0.472	0.528	0.560	0.530	8.512	0.711	−0.387	−0.095	PETN
1h63.A	33	0.432	0.568	0.472	0.528	0.558	0.542	7.978	0.659	−0.384	−0.092	PETN
1vys.X	33	0.426	0.574	0.486	0.514	0.556	0.539	7.739	0.614	−0.370	−0.070	PETN
1vyr.A	37	0.436	0.564	0.484	0.516	0.557	0.526	8.420	0.663	−0.355	−0.203	PETN
1gvo.A	38	0.432	0.568	0.477	0.523	0.559	0.510	8.727	0.678	−0.405	−0.101	PETN
1gvq.A	37	0.431	0.569	0.485	0.515	0.560	0.538	8.585	0.745	−0.391	−0.096	PETN
1gvs.A	34	0.438	0.562	0.481	0.519	0.560	0.544	8.006	0.683	−0.402	−0.100	PETN
1h51.A	33	0.432	0.568	0.473	0.527	0.564	0.543	7.878	0.658	−0.405	−0.098	PETN
1vyp.X	35	0.437	0.563	0.473	0.527	0.554	0.535	8.159	0.649	−0.403	−0.101	PETN
2aba.A	38	0.426	0.574	0.476	0.524	0.568	0.511	9.392	0.753	−0.438	−0.089	PETN
3f03.K	32	0.428	0.572	0.473	0.527	0.560	0.541	8.167	0.626	−0.425	−0.128	PETN
2abb.A	30	0.424	0.576	0.434	0.566	0.569	0.553	7.953	0.684	−0.439	−0.111	PETN
1z41.A	34	0.432	0.568	0.452	0.548	0.528	0.532	8.222	0.713	−0.056	0.252	1.6.99.1
1z42.A	21	0.429	0.571	0.505	0.495	0.529	0.596	8.649	0.682	−0.059	0.238	1.6.99.1
1z44.A	21	0.426	0.574	0.512	0.488	0.525	0.592	8.736	0.668	−0.065	0.268	1.6.99.1
1z48.A	31	0.427	0.573	0.462	0.538	0.528	0.544	8.164	0.707	−0.053	0.254	1.6.99.1
1bwk.A	31	0.403	0.597	0.370	0.630	0.550	0.587	6.370	0.610	−0.136	0.065	1.6.99.1
1bwl.A	30	0.399	0.601	0.380	0.620	0.550	0.581	6.783	0.594	−0.132	0.063	1.6.99.1
1k02.A	27	0.409	0.591	0.398	0.602	0.537	0.588	6.378	0.576	−0.096	0.147	1.6.99.1
1k03.A	41	0.410	0.590	0.371	0.629	0.530	0.512	8.765	0.606	−0.036	0.291	1.6.99.1
1oya.A	31	0.402	0.598	0.383	0.617	0.551	0.595	6.776	0.604	−0.109	0.129	1.6.99.1
1oyb.A	37	0.367	0.633	0.311	0.689	0.550	0.552	7.047	0.657	−0.152	0.108	1.6.99.1
1oyc.A	33	0.371	0.629	0.336	0.664	0.554	0.575	6.982	0.562	−0.161	0.108	1.6.99.1
2hsa.A	44	0.390	0.610	0.404	0.596	0.529	0.473	12.322	0.691	−0.267	0.192	1.3.1.42
3hgs.A	34	0.389	0.611	0.418	0.582	0.538	0.537	8.696	0.707	−0.436	−0.054	1.3.1.42
2hs8.A	44	0.392	0.608	0.408	0.592	0.549	0.484	10.240	0.617	−0.497	−0.090	1.3.1.42
3hgo.A	43	0.386	0.614	0.408	0.592	0.547	0.489	11.104	0.722	−0.493	−0.107	1.3.1.42
2hs6.A	29	0.398	0.602	0.434	0.566	0.563	0.566	8.233	0.754	−0.480	−0.090	1.3.1.42
1q45.A	44	0.403	0.597	0.441	0.559	0.547	0.501	10.674	0.719	−0.354	−0.020	1.3.1.42
2q3o.A	28	0.515	0.485	0.457	0.543	0.589	0.559	10.378	0.510	−0.360	−0.006	1.3.1.42
2g5w.A	33	0.402	0.598	0.400	0.600	0.541	0.543	9.237	0.720	−0.369	−0.015	1.3.1.42
1icp.A	37	0.419	0.581	0.403	0.597	0.529	0.508	8.936	0.586	−0.316	0.038	1.3.1.42
1icq.A	39	0.416	0.584	0.388	0.612	0.536	0.499	8.882	0.665	−0.342	−0.013	1.3.1.42
1ics.A	43	0.413	0.587	0.405	0.595	0.539	0.478	10.256	0.626	−0.309	0.062	1.3.1.42
3hgr.A	44	0.419	0.581	0.398	0.602	0.546	0.491	9.099	0.585	−0.373	−0.039	1.3.1.42
1vji.A	52	0.407	0.593	0.427	0.573	0.540	0.443	10.799	0.675	−0.227	−0.013	1.3.1.42
2q3r.A	26	0.476	0.524	0.343	0.657	0.941	0.546	9.504	0.551	−0.213	0.022	1.3.1.42

aPams, global apolar solvent accessible area (Å^2); *aPolar*, local apolar solvent accessible area (Å^2); *d*, anisotropic (Å); *len*, number of residues in a pocket (aa); *Pams*, global polar solvent accessible area (Å^2); *Polar*, local polar solvent accessible area (Å^2); *SD*, local surface density (g/mol Å^2); *sph*, local sphericity; *WG1*, global skewness; *WG2*, global kurtosis; *Wsph*, global sphericity.

Table S2. Representative canonical surfaces from 39 surface types with selected attributes

PDB ID	EC annotation	PSC ID	CATH ID
2j30.A	3.4.22.56	ST178	3.40.50.1460
1nms.A	3.4.22.56	ST178	3.40.50.1460
2j31.A	3.4.22.56	ST178	3.40.50.1460
2j32.A	3.4.22.56	ST178	3.40.50.1460
1nmq.A	3.4.22.56	ST178	3.40.50.1460
3h0e.A	3.4.22.56	ST178	NA
2j33.A	3.4.22.56	ST178	3.40.50.1460
3deh.B	3.4.22.56	ST178	3.40.50.1460
3dej.B	3.4.22.56	ST178	3.40.50.1460
3dek.B	3.4.22.56	ST178	3.40.50.1460
1cp3.A	3.4.22.56	ST178	3.40.50.1460
3dei.B	3.4.22.56	ST178	3.40.50.1460
1rhm.A	3.4.22.56	ST178	3.40.50.1460
1rhq.A	3.4.22.56	ST178	3.40.50.1460
2c1e.B	3.4.22.56	ST178	3.30.70.1470
2c2k.B	3.4.22.56	ST178	3.30.70.1470
2c2m.B	3.4.22.56	ST178	3.30.70.1470
2c2o.B	3.4.22.56	ST178	3.30.70.1470
2cdr.B	3.4.22.56	ST178	3.30.70.1470
2cjy.B	3.4.22.56	ST178	3.30.70.1470
2cnk.B	3.4.22.56	ST178	3.30.70.1470
2cnl.B	3.4.22.56	ST178	3.30.70.1470
2cnn.B	3.4.22.56	ST178	3.30.70.1470
2dko.B	3.4.22.56	ST178	3.30.70.1470
2h5i.B	3.4.22.56	ST178	3.30.70.1470
2cno.B	3.4.22.56	ST178	3.30.70.1470
2h65.B	3.4.22.56	ST178	3.30.70.1470
3edq.B	3.4.22.56	ST178	3.30.70.1470
1nme.B	3.4.22.56	ST178	3.30.70.1470
3gjq.B	3.4.22.56	ST178	3.30.70.1470
3gjr.B	3.4.22.56	ST178	3.30.70.1470
1rhu.B	3.4.22.56	ST178	3.30.70.1470
3gjt.B	3.4.22.56	ST178	3.30.70.1470

CATH, class, architecture, topology, homologous superfamily; NA, not assigned; PSC, protein surface classification.

Table S3. Thirty-three entries in the family of cysteine 3 endopeptidase (EC 3.4.22.56) used for the comparison of similarities among EC, PSC, and CATH

PDB ID	*len*	*Pams*	*nPams*	*Polar*	*nPolar*	*Wsph*	*sph*	*d*	*SD*	*WG1*	*WG2*	*EC*
1bwk.A	31	0.403	0.597	0.370	0.630	0.550	0.587	6.370	0.610	−0.136	0.065	1.6.99.1
2prl.A	47	0.427	0.573	0.313	0.687	0.509	0.496	9.839	0.666	−0.379	−0.292	1.3.5.2
1al7.A	34	0.387	0.613	0.404	0.596	0.521	0.521	10.307	0.631	−0.047	0.075	1.1.3.15
1ag9.A	13	0.439	0.561	0.428	0.572	0.607	0.620	13.500	0.572	−0.126	0.089	NA
1ds7.A	27	0.371	0.629	0.302	0.698	0.476	0.466	13.818	0.570	−0.043	−0.482	1.5.1.34
1c7e.A	13	0.476	0.524	0.408	0.592	0.632	0.597	14.124	0.469	−0.252	−0.011	NA
1g28.A	20	0.383	0.617	0.384	0.616	0.605	0.605	4.736	0.768	0.288	0.342	NA
1e5d.A	15	0.390	0.610	0.392	0.608	0.516	0.612	35.413	0.423	0.056	−0.669	NA
1ja0.B	51	0.418	0.582	0.445	0.555	0.436	0.440	4.521	0.655	−0.146	−0.625	1.6.2.4
1ci0.A	22	0.406	0.594	0.350	0.650	0.489	0.520	11.279	0.612	0.133	−0.563	1.4.3.5
1yrx.A	16	0.384	0.616	0.357	0.643	0.502	0.645	5.520	0.683	1.384	2.465	NA
1ea0.A	42	0.413	0.587	0.401	0.599	0.402	0.536	21.480	0.724	−0.289	−0.310	1.4.1.13
1yrh.A	16	0.387	0.613	0.469	0.531	0.530	0.578	13.783	0.540	0.068	−0.130	NA
1bkj.A	52	0.374	0.626	0.323	0.677	0.464	0.424	8.995	0.594	−0.075	−0.192	NA
1djn.A	25	0.424	0.576	0.353	0.647	0.430	0.607	9.349	0.847	−0.072	−0.445	1.5.8.2
1e20.A	21	0.392	0.608	0.354	0.646	0.539	0.557	15.493	0.483	0.303	0.169	4.1.1.36
1t5b.A	12	0.362	0.638	0.362	0.638	0.541	0.692	14.307	0.466	0.312	−0.266	NA
1n07.A	43	0.396	0.604	0.393	0.607	0.538	0.492	9.140	0.579	−0.282	−0.155	2.7.1.26
2d36.A	30	0.364	0.636	0.362	0.638	0.529	0.513	8.987	0.632	0.077	0.434	NA
2h8x.A	39	0.407	0.593	0.377	0.623	0.556	0.530	9.024	0.728	0.136	0.585	NA
2ohh.A	11	0.395	0.605	0.344	0.656	0.509	0.650	35.726	0.399	0.163	−0.597	NA
2zru.A	43	0.381	0.619	0.401	0.599	0.520	0.489	11.354	0.621	−0.189	−0.182	5.3.3.2
1yw3.A	15	0.386	0.614	0.368	0.632	0.464	0.517	16.149	0.509	0.492	0.417	NA
2z6i.A	70	0.391	0.609	0.385	0.615	0.469	0.434	8.069	0.663	0.378	0.463	1.3.1.9
3gbh.A	22	0.408	0.592	0.394	0.606	0.445	0.516	12.332	0.492	0.005	−0.526	NA
1bvy.A	70	0.403	0.597	0.344	0.656	0.467	0.430	10.115	0.690	−0.230	−0.336	1.14.14.1; 1.6.2.4
1he4.A	33	0.414	0.586	0.440	0.560	0.563	0.504	8.223	0.762	−0.363	−0.049	1.3.1.24; 1.5.1.30
1t6y.A	37	0.415	0.585	0.390	0.610	0.462	0.501	13.735	0.590	0.001	−0.814	NA
1y56.B	62	0.388	0.612	0.413	0.587	0.522	0.460	4.536	0.727	−0.070	−0.429	NA
2i02.A	12	0.430	0.570	0.305	0.695	0.557	0.644	8.917	0.599	−0.099	−0.551	NA
2isj.A	63	0.394	0.606	0.362	0.638	0.458	0.381	11.502	0.704	0.122	−0.108	NA
2j09.A	47	0.405	0.595	0.397	0.603	0.485	0.511	8.592	0.737	−0.044	−0.450	4.1.99.3
2pia.A	42	0.447	0.553	0.451	0.549	0.483	0.475	5.866	0.673	−0.151	−0.454	NA
2vbv.A	33	0.380	0.620	0.342	0.658	0.559	0.541	7.306	0.630	−0.025	0.189	2.7.1.161
3fgc.A	37	0.400	0.600	0.365	0.635	0.469	0.485	9.348	0.704	−0.145	−0.373	1.14.14.3
3g5a.A	43	0.419	0.581	0.435	0.565	0.484	0.471	11.561	0.751	−0.099	0.016	NA
3iam.1	35	0.400	0.600	0.292	0.708	0.496	0.501	7.032	0.736	0.023	−0.214	1.6.99.5