RESEARCH PAPER

Intrinsically Disordered Proteins 1, e25463; 2013; © 2013 Landes Bioscience

Structural phylogeny by profile extraction and multiple superimposition using electrostatic congruence as a discriminator

Sandeep Chakraborty,[1,*] Basuthkar J. Rao,[1] Nathan Baker[2] and Bjarni Ásgeirsson[3]

[1]Department of Biological Sciences; Tata Institute of Fundamental Research; Mumbai, India; [2]Pacific Northwest National Laboratory; Richland, WA USA; [3]Science Institute; Department of Biochemistry; University of Iceland; Reykjavik, Iceland

Keywords: computational biology, phylogenetics, active site prediction, finite difference Poisson-Boltzmann (FDPB), APBS, evolution, multiple sequence alignment, multiple superimposition, directed evolution, structural alignment

Abbreviations: MSA, multiple sequence alignment; CLASP, CataLytic Active Site Prediction; APBS, Adaptive Poisson-Boltzmann Solver; STEEP, structure and electrostatic potential based multiple sequence alignment; MBL, Metallo-β-lactamase; SBL, Serine-β-lactamase; PBP, Penicillin binding protein; STP, signal transduction protein

Phylogenetic analysis of proteins using multiple sequence alignment (MSA) assumes an underlying evolutionary relationship in these proteins which occasionally remains undetected due to considerable sequence divergence. Structural alignment programs have been developed to unravel such fuzzy relationships. However, none of these structure based methods have used electrostatic properties to discriminate between spatially equivalent residues. We present a methodology for MSA of a set of related proteins with known structures using electrostatic properties as an additional discriminator (STEEP). STEEP first extracts a profile, then generates a multiple structural superimposition providing a consolidated spatial framework for comparing residues and finally emits the MSA. Residues that are aligned differently by including or excluding electrostatic properties can be targeted by directed evolution experiments to transform the enzymatic properties of one protein into another. We have compared STEEP results to those obtained from a MSA program (ClustalW) and a structural alignment method (MUSTANG) for chymotrypsin serine proteases. Subsequently, we used PhyML to generate phylogenetic trees for the serine and metallo-β-lactamase superfamilies from the STEEP generated MSA, and corroborated the accepted relationships in these superfamilies. We have observed that STEEP acts as a functional classifier when electrostatic congruence is used as a discriminator, and thus identifies potential targets for directed evolution experiments. In summary, STEEP is unique among phylogenetic methods for its ability to use electrostatic congruence to specify mutations that might be the source of the functional divergence in a protein family. Based on our results, we also hypothesize that the active site and its close vicinity contains enough information to infer the correct phylogeny for related proteins.

Introduction

DNA sequencing technologies have provided a quantitative foundation for our understanding of evolution, which was previously based on logical, yet empirical, observations.[1] The chronology of the development of computational techniques has closely followed innovations in biotechnology. Pairwise alignment algorithms of nucleotide sequences, both global[2] and local,[3] were enhanced to incorporate multiple sequences from related proteins.[4-7] Such multiple sequence alignment (MSA) methods enabled visualization of evolutionary pathways through phylogenetic trees.[8,9] While considerable divergence in sequence often resembles noise and masks true relationships, structural conservation in such cases have provided the basis for evolutionary kinship. For instance, MSA techniques are not applicable to the serine and metallo-β-lactamase superfamilies due to significant sequence divergence.[10-14] Lately, rapid strides in crystallization techniques have fueled progress in structural alignment methods, both for pairwise[15-20] and multiple[21-28] proteins.

The program MAPS (an extension of the program TOP),[28] which has been used for the structural analysis of metallo-β-lactamases,[12] first superimposes the proteins and then computes the phylogeny based on structural similarity of the main and side-chain atoms. A widely used methodology for structural alignment (MUSTANG) uses a simple dynamic programming algorithm for all pairs of structures and applies a robust scoring

*Correspondence to: Sandeep Chakraborty; Email: sanchak@gmail.com
Submitted: 06/19/13; Accepted: 06/19/13
http://dx.doi.org/10.4161/idp.25463
Citation: Chakraborty S, Rao BJ, Baker N, B Ásgeirsson. Structural phylogeny by profile extraction and multiple superimposition using electrostatic congruence as a discriminator. Intrinsically Disordered Proteins 2013; e25463

Table 1. Potential and spatial congruence of the active site residues in proteins from the chymotrypsin superfamily

PDB		ab	ac	bc
2ALP	D	4.7	3.1	6.2
	PD	−13.6	−86.5	−72.9
1SGT	D	5.5	3	8
	PD	4.2	−120.4	−124.5
1TGS	D	5.2	2.6	7.3
	PD	31.6	−85.8	−117.4
2SGA	D	4.6	3	6.2
	PD	59.1	−123.6	−182.6
1PPF	D	5.4	2.5	7.3
	PD	−29	−103.7	−74.6
3EST	D	4.6	3.2	6.4
	PD	−3.7	−124	−120.3
3RP2	D	5	3.1	6.5
	PD	−51.9	−136.4	−84.5
1TPP	D	5.5	2.7	7.6
	PD	−83.9	−162.5	−78.6

The active site atoms are HIS57NE2 (a), ASP102OD1 (b) and SER195OG (c). D = Pairwise distance in Å. PD = Pairwise potential difference. The electrostatic potential is in dimensionless units of kT/e where k is Boltzmann's constant, T is the temperature in K and e is the charge of an electron.

scheme obviating the need for troublesome gap penalties.[22] A recent method uses many informative features (torsion angles, secondary structure, residue type, surface accessibility, etc.) to guide the alignment.[29] An innovative technique for alignment allows local flexibility between fragments which might be physically impossible under rigid body transformations and restores geometric consistency at the end.[30] Another multiple protein alignment method (MISTRAL) uses the minimization of an empirical energy function of the relative rotations and translations of the molecules.[31] However, such methods have not addressed the problem of identifying residues which, although spatially equivalent, have diverged from a stereochemical and electrostatic perspective resulting in functional plasticity.

In the current work, we present a methodology for generating the MSA of a set of related proteins with known structures, using electrostatic properties as an additional discriminator - *St*ructure and *el*ectrostatic *p*otential based multiple sequence alignment (STEEP). We demonstrate that residues identified by comparing the alignments obtained by including and excluding electrostatic properties can be targeted by directed evolution experiments to transform the enzymatic properties of one protein into another. We also show that the active site vicinity contains enough information to infer correct kinship in a set of related proteins.

Previous work by our group has established the spatial and electrostatic congruence in cognate residue pairs of the active site in proteins with the same functionality - *C*ata*L*y*t*ic *A*ctive *S*ite *P*rediction (CLASP).[32] CLASP was used to unravel a serine protease scaffold in alkaline phosphatases,[32] and a scaffold recognizing a β-lactam (imipenem) in a cold-active *Vibrio* alkaline phosphatase.[33,34] STEEP superimposes the proteins based on the active site motif specified in one of the proteins by extracting matching scaffolds using CLASP, thus pruning out unrelated proteins which are known to affect the quality of MSA results.[35] It then considers the reactive atoms of the residues in the superimposed cluster while matching the distance, and as an additional option uses electrostatic criteria to prune out non-congruent residues, and emits the MSA for the set of proteins. Such a constrained alignment highlights the conserved residues from an electrostatic perspective as well. Comparison of these alignments could form the basis of mutations in directed evolution experiments that intend to endow the desired protein with certain enzymatic properties.[36]

We have compared results obtained with STEEP to those obtained from a sequence based MSA program (ClustalW),[4] and a structural alignment method (MUSTANG)[22] for a set of chymotrypsin serine proteases. We have also generated phylogenetic trees for the serine and metallo-β-lactamase superfamilies from the STEEP generated MSA using PhyML,[8] and corroborated the accepted relationships of proteins in these two superfamilies.[10-14] Interestingly, using electrostatic congruence as a discriminator led to a functional classification instead of a true evolutionary relationship. We observe that Trp154 in class D serine β-lactamases (SBL) and signal transduction proteins is spatially equivalent to Glu166 in class A SBL but lacks electrostatic congruence. Although this critical Trp154 has been mutated to Gly, Ala, and Phe with resulting poor catalytic efficiencies and reduced stability,[37] we propose that a mutation to Glu might show functional similarity to class A SBLs by mimicking the Glu166.[38] In summary, STEEP is a multi-faceted methodology that generates evolutionary and functional relationships in a set of related proteins, a multiple superimposition and proposes mutations based on electrostatic properties that might endow the enzymatic functionality of one protein to another. Thus, it helps in narrowing down critical mutations that would be expected to shape the functional plasticity of any given enzyme superfamily, especially in cases where sequence divergence has left little traces of any relationship.

Results

Chymotrypsin serine proteases. Serine proteases are grouped based on structural homology and are then further sub-grouped into families with similar sequences.[39,40] The two major families, chymotrypsin and subtilisin, are a classical example of convergent evolution where the catalytic Ser-His-Asp triad shows very similar geometry in the structurally different chymotrypsin and subtilisin families.[41] We chose a set of eight proteins (PDBids: 2ALP, 1SGT, 1TGS, 2SGA, 1PPF, 3EST, 3RP2 and 1TPP) for analysis based on previous work on serine proteases,[42] barring one (PDBid:5CHA) which did not complete APBS electrostatic analysis (**Tables S1 and S2**). The motif from a trypsin protein (PDBid:1SBT)—(His57, Asp102, Ser195)—was chosen for representing serine proteases. CLASP analysis using this query motif detected significantly congruent scaffolds in each of the proteins (**Table 1**).

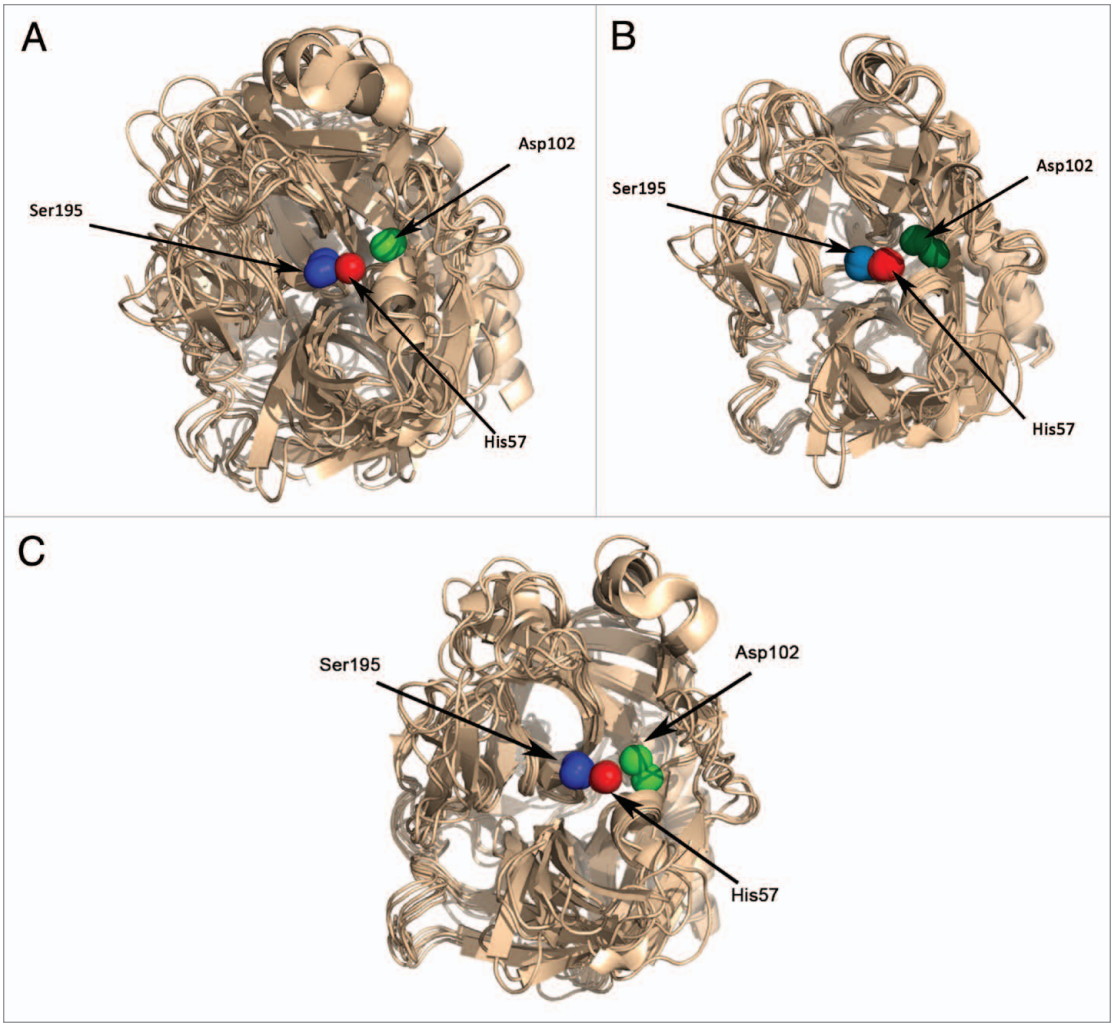

Figure 1. Superimposing multiple proteins based on the homologous active site scaffolds for trypsin serine proteases. (**A**) STEEP generated superimposition, where each amino acid is represented by a user defined reactive atom. (**B**) MUSTANG generated superimposition. It can be seen that MUSTANG generates a better overall superimposition, but the active site residues are less dispersed after the superimposition by STEEP. (**C**) STEEP generated superimposition, where each amino acid is represented by the Cα atom.

The structural profile was used to generate a multiple superimposition of the proteins (see Materials and Methods), and provided a single frame of reference for comparing the proteins with STEEP (**Fig. 1A**). Since all the structures could be superimposed, we proceeded to finding the residues from other proteins in the set which were spatially close to residues of the template protein. **Figure 2A** and **Figure 2C** show the alignment and the cladogram using only spatial constraints. The alignment (**Fig. 2B**) and the cladogram (**Fig. 2D**) taking electrostatic congruence into consideration resulted in a different phylogeny from the one generated using just spatial constraints. Later, we show for serine β-lactamases (SBL) that this relationship suggests functional relationships rather than true evolutionary kinship. For example, classes A and C SBLs appear as sister taxa when using electrostatic congruence as a discriminator, and it is known that both classes A and C SBLs have the ability to hydrolyze cephalosporins. This is a reasonable finding considering that electrostatic fields have a direct bearing on specificity and functionality.

We next used ClustalW to generate the alignment (**Fig. 2E**) and the phylogenetic tree (**Fig. 2G**) for the same set of proteins. We used the sequences obtained from the PDB files, and not the complete fasta sequence, to ensure a fair comparison with STEEP and MUSTANG. MUSTANG results also showed similar alignments (**Fig. 2F**) and phylogenetic trees (**Fig. 2H**). For example, all three methods suggest that the protein groups (2ALP-2SGA), (3EST-3RP2) and (1SGT-1TGS-1TPP-1PPF) are closely related. Qualitatively, the structural alignment obtained from MUSTANG (**Fig. 1B**) was better than that from STEEP (**Fig. 1A**), but the active site residues in STEEP were less dispersed since STEEP aligns the proteins based on the active site residues. **Table 2** shows the RMSD values for the pairwise comparison of a protein (PDBid:1TGS) with all other proteins. Cα atoms that are within 2 Å of each other are considered to be equivalent. However, it is seen that STEEP comparisons resulted in a better overall fit when each amino acid was represented by the Cα atom rather than the reactive atom (a much smaller RMSD and more

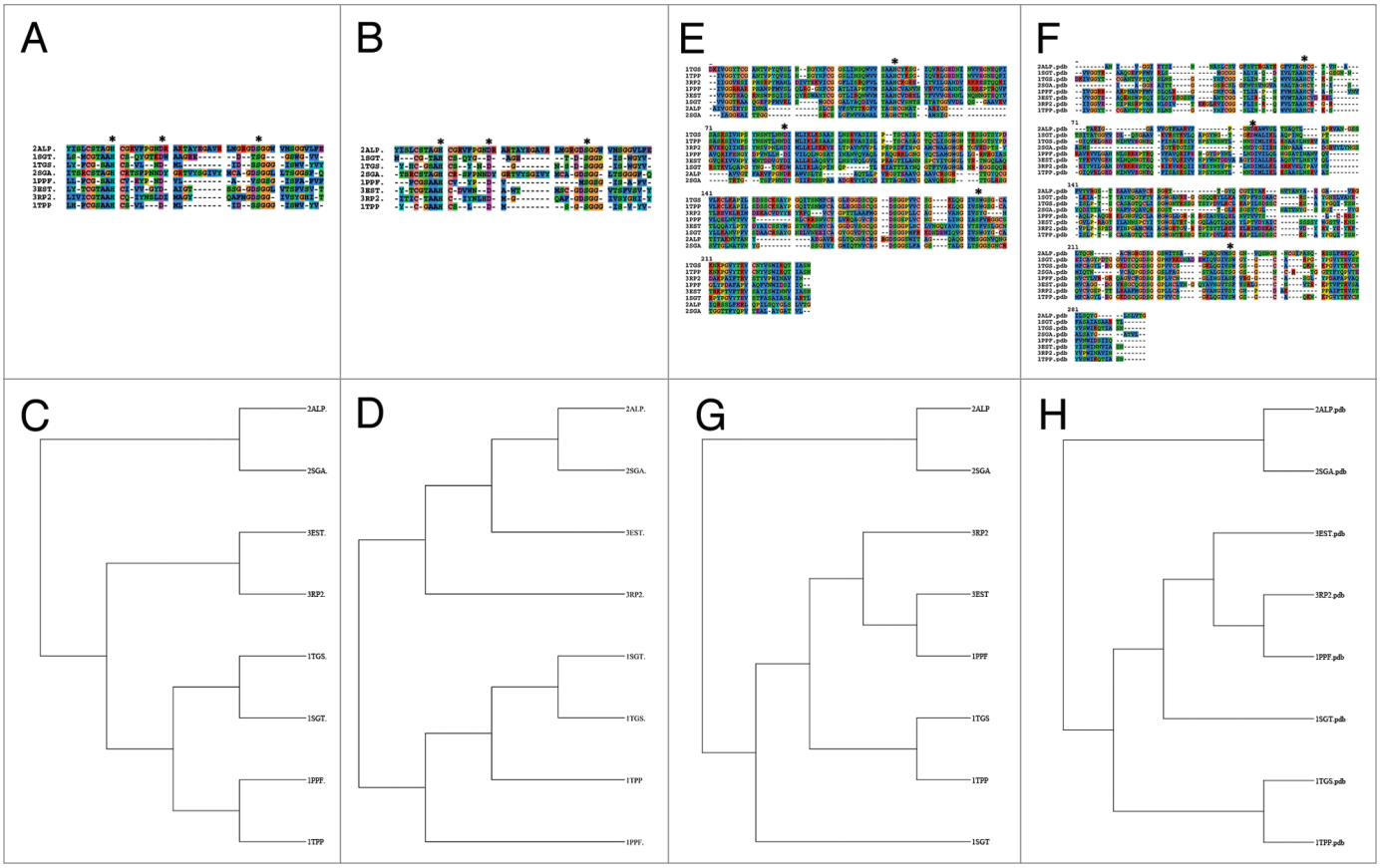

Figure 2. Multiple sequence alignments using STEEP, ClustalW or MUSTANG, and phylogenetic trees generated using PhyML for the chymotrypsin superfamily. The active site motif is marked as '*'. The residues used to initiate STEEP were within a radius of 9 Å from the specified active site residues. (**A**) Alignment using spatial proximity using STEEP. (**B**) Alignment using spatial proximity and electrostatic congruence using STEEP. (**C**) Cladogram generated from (A). (**D**) Cladogram generated from (B). (**E**) Alignment using ClustalW. (**F**) Alignment using MUSTANG. (**G**) Cladogram generated from (E) (ClustalW). (**H**) Cladogram generated from (F) (MUSTANG).

equivalent residues (**Fig. 1C**; **Table 2**). Since the STEEP methodology is directed at identifying active site residue equivalence, it does not intend to obtain the best global superimposition. Thus, by default the amino acids are represented by their reactive atoms (when this applies).

To summarize, we show that STEEP generates similar phylogenies as obtained from sequence alignment (ClustalW) and structural alignment (MUSTANG) programs by considering residues in the vicinity of the active site, and also generates a superimposition comparable to the one generated by MUSTANG by simply aligning the active site residues.

Serine and metallo-β-lactamase superfamilies. β-lactamases inactivate antibiotics by hydrolyzing the amide bond of the β-lactam ring. The Ambler classification has four classes—classes A, C and D have a nucleophilic serine at the active site (SBL),[43] while MBLs or class B β-lactamases are metallo-enzymes requiring zinc for their activity, and have been further divided into three subgroups - B1, B2, and B3 - based on sequence homology.[44]

SBLs are characterized by three conserved motifs [SXXK, (S/Y)X(N/V) and K(T/S)G].[43] We constructed the active site motif (Ser70, Lys73, Ser130, Lys234) by choosing at least one residue from each of the three motifs from a class A SBL (PDBid:1E25).[45] While searching for matches, Ser130 was matched with either Ser or Tyr to accommodate the variability seen in various SBLs. The set of proteins analyzed consisted of three structures from each of the classes A, C and D of SBLs and penicillin binding proteins (PBP), and two structures from signal transduction proteins (**Tables S3 and S4**). CLASP queried the set of proteins using the active site motif, and detected significantly congruent scaffolds in each of the proteins (**Table 3**). Thus, these residues represent a structural profile for the serine β-lactamase superfamily.

In MBLs, classes B1 and B3 possess a binuclear active site that requires one or two Zn^{2+} ions (Zn1 and Zn2 site) for full activity.[44] Subclass B2 enzymes are catalytically active with one Zn^{2+} ion,[46] while the binding of the second zinc ion has been shown to have inhibitory effects.[47] The active site profile for MBLs was created from two residues each from the Zn1 (His118 and His196) and Zn2 (Asp120 and His263) ligands. The set of proteins analyzed consisted of three structures from each of the classes B1 and B3, a structure each from the class B2, glyoxalase II and a methyl parathion hydrolase (**Tables S5 and S6**). As expected, we detected significantly congruent scaffolds in each of the proteins (**Table 4**). It is this spatial and electrostatic congruence that has

been used to identify a scaffold recognizing a β-lactam (imipenem) in a cold-active *Vibrio* alkaline phosphatase.[34]

Figure 3A shows the superimposition of the three class A SBLs, **Figure 3B** shows the superimposition of one structure each of the classes (A, C and D) SBLs, while **Figure 3C** shows the superimposition of a class A SBL, a PBP and a signal transduction protein. Likewise for the MBL superfamily, **Figure 3D** shows the superimposition of the three class B1 MBLs, **Figure 3E** shows the superimposition of one structure each of the classes (B1, B2 and B3) MBLs, while **Figure 3F** shows the superimposition of a class B3 MBL, a human glyoxalase II and a methyl parathion hydrolase.

It can be seen from these superimpositions that the active site shape is conserved, while the proteins accommodate much greater structural changes in the peripheral regions. The superimposition of all proteins is shown in **Figure S1**. It was noted here that aligning only three residues has the effect of aligning the complete protein, highlighting that the protein sequence accepts only those mutations that do not violate the conserved structure (and electrostatic properties) of the active site. Thus, it is only logical to compare these proteins based on the conserved residues in the vicinity of the active site.

Figure 4A and **Figure 4C** in SBLs (**Fig. 4E and 4G** in MBLs) show the alignment and the cladogram, respectively, in the case when we ignore electrostatic congruence with the residues in the template protein. In the scenario where potential difference congruence is used as discriminator, we obtained the alignment shown in **Figure 4B** and the cladogram of **Figure 4D** in SBLs (**Fig. 4F and 4H** in MBLs) as the alignment and the cladogram.

It has been shown previously that class A and class D SBLs are sister taxa, and the divergence of the class C SBL predated the bifurcation of classes A and D SBLs.[10] **Figure 4C** corroborates this hypothesis. Simultaneously, **Figure 4C** conforms to the known relationship between class D SBL and signal transduction proteins.[48] Interestingly, the expected similarity in class A enzymes and some penicillin binding proteins (PDBid:1NZ0) is not apparent from the cladogram. The deletion of a segment from the sequence close to the active site in these PBPs makes it difficult for even structural programs to identify such relationships.[49]

Interestingly, when we constrained the MSA using electrostatic congruence criteria, a different relationship emerged (**Fig. 4D**) which suggests that classes A and C SBLs are sister taxa. This dichotomy is explained by the fact that electrostatic homology often implies functional similarity—and it is known that both classes A and C SBLs have the ability to hydrolyze cephalosporins, unlike class D SBLs which are specialized oxacillanases.[10] Thus, **Figure 4D** ought to be interpreted as indicating functional relationship along with sequence/structural homology. A similar observation reveals PBPs and signal transduction proteins closer in **Figure 4D** as compared with **Figure 4C**, highlighting their functional similarity, namely their inability to hydrolyze β-lactams.

It has been shown that the B3 subclass of MBLs is distinct from the B1/B2 subclass based on sequence alignment.[13] Extending this work, it was proposed that functionality in B1/B2 evolved approximately one billion years ago, whereas subclass B3 evolved about two billion years ago before Gram-positive and Gram-negative eubacteria had diverged.[14] The culmination of this work was achieved by applying structural methods to generate a phylogeny which corroborated the above hypotheses, and also included other proteins from the MBL superfamily (like human glyoxalase II and methyl parathion hydrolase).[12] Furthermore, human glyoxalase II and methyl parathion hydrolase were shown to be closely related to subclass B3. **Figure 4E, F, G and H** demonstrate that the STEEP results corroborate these hypotheses. The MBLs show much more electrostatic homogeneity than SBLs in the related classes (**Fig. 4F** and **Figure 4H**). The inhibitory effect of the binding of the second zinc ion in subclass B2 enzymes is also highlighted by the fact that Asn116 has spatial equivalence (**Fig. 4E**), but lacks electrostatic congruence with the corresponding histidine (His116) in the other subclasses B1 and B3 (**Fig. 4F**).

The MUSTANG generated phylogenetic tree for SBLs did not reflect the accepted relationship in the superfamily (**Fig. S2A**), since class A and class C enzymes were seen to be sister taxa, rather than of class A and class D enzymes.[10] In fact, this relationship is similar to the functional relationship detected by

Table 2. Comparing results obtained with STEEP or MUSTANG for serine proteases

	PDB	RMSD	Residues matched (out of 222)
STEEP Cα atoms	1PPF	1	170
	2ALP	1.1	97
	1SGT	1.1	170
	2SGA	1.2	97
	3EST	0.9	187
	3RP2	1.1	179
	1TPP	1.2	160
STEEP reactive atoms	1PPF	1.4	89
	2ALP	1.5	39
	1SGT	1.4	62
	2SGA	1.4	39
	3EST	1.5	51
	3RP2	1.5	48
	1TPP	0.5	206
MUSTANG	1PPF	0.9	176
	2ALP	1.3	96
	1SGT	0.9	177
	2SGA	1.3	100
	3EST	0.9	187
	3RP2	0.9	182

The RMSD obtained for superimposing one protein (PDBid:1TGS, 222 amino acids) with all other proteins are shown. Cα atoms that are within 2 Å of each other are considered to be equivalent. The number of residues matched is another important metric, since an inferior superimposition might have an equivalent RMSD, but align fewer residues. It is seen that when each amino acid is represented by the Cα atom rather than the reactive atom STEEP results in much smaller RMSD and more equivalent residues.

Table 3. Potential and spatial congruence of the active site residues in proteins from the Serine β-lactamase superfamily

PDB	Active site atoms (a,b,c,d)		ab	ac	ad	bc	bd	cd
1E25	Ser70OG, Lys73NZ, Ser130OG, Lys234NZ,	D	2.8	3.2	4.7	3.6	5.6	2.9
	Class A Serine β-lactamase	PD	−125.6	22.4	−189.1	148.1	−63.5	−211.5
1I2S	Ser70OG, Lys73NZ, Ser130OG, Lys234NZ,	D	2.7	3.2	4.5	3.1	5	2.8
	Class A Serine β-lactamase	PD	−166.4	−35.5	−219.5	130.9	−53.1	−184
1BSG	Ser70OG, Lys73NZ, Ser130OG, Lys234NZ,	D	2.8	3.4	4.7	3.3	5.3	2.9
	Class A Serine β-lactamase	PD	−178.3	−31.4	−188.6	146.8	−10.3	−157.1
2WZX	Ser90OG, Lys93NZ, Tyr177OH, Lys342NZ,	D	3.5	3	4.5	2.6	5	2.8
	Class C Serine β-lactamase	PD	−161.5	−56.9	−153.1	104.6	8.4	−96.2
1KE4	Ser64OG, Lys67NZ, Tyr150OH, Lys315NZ,	D	2.9	3	4.6	3.4	5.6	2.8
	Class C Serine β-lactamase	PD	−228	−10.4	−187.1	217.6	40.9	−176.7
1FR6	Ser64OG, Lys67NZ, Tyr150OH, Lys315NZ,	D	2.9	3.3	4.6	2.4	5	3.1
	Class C Serine β-lactamase	PD	−132.1	−18.4	−164.2	113.7	−32.1	−145.8
1K57	Ser67OG, Lys70NZ, Ser115OG, Lys205NZ,	D	2.8	2.6	4.7	3.1	5.6	3.8
	Class D Serine β-lactamase	PD	−162.9	51.7	−184.7	214.6	−21.8	−236.4
3ISG	Ser67OG, Lys70NZ, Ser115OG, Lys212NZ,	D	3.3	3.9	4.3	4.8	5.3	2.2
	Class D Serine β-lactamase	PD	−246.3	−13.9	−231.5	232.4	14.8	−217.7
1K38	Ser67OG, Lys70NZ, Ser115OG, Lys205NZ,	D	3.1	3.7	4.9	4.7	5.6	2.7
	Class D Serine β-lactamase	PD	−292.5	−50.7	−309.8	241.8	−17.3	−259.1
1QME	Ser337OG, Lys340NZ, Ser395OG, Lys547NZ,	D	2.9	3.2	4.5	2.7	5	3
	Penicillin binding protein	PD	−211.5	−38.2	−242	173.3	−30.5	−203.8
1NZO	Ser44OG, Lys47NZ, Ser110OG, Lys213NZ,	D	3.1	4.2	6.3	5.1	6.8	2.7
	Penicillin binding protein	PD	−241.6	−68.8	−277.9	172.8	−36.2	−209.1
2EX2	Ser62OG, Lys65NZ, Ser306OG, Lys417NZ,	D	2.9	3	4.3	3.3	5	2.9
	Penicillin binding protein	PD	−213.6	−84	−264.8	129.6	−51.2	−180.8
1XA1	Ser59OG, Lys62NZ, Ser107OG, Lys196NZ,	D	2.6	3.5	4.7	3.8	5.8	2.9
	Signal transducer BlaR1	PD	−126.2	73.7	−175.8	199.9	−49.6	−249.5
1NRF	SER402OG, LYS405NZ, SER450OG, LYS539NZ,	D	2.7	3.6	4.7	4.9	6.1	2.8
	Signal transducer BlaR1	PD	−249.6	2.1	−217.7	251.7	31.9	−219.8

D = Pairwise distance in Å. PD = Pairwise potential difference. The electrostatic potential is in dimensionless units of kT/e where k is Boltzmann's constant, T is the temperature in K and e is the charge of an electron.

STEEP by using electrostatic pruning (**Fig. 4D**). From the complete set used by STEEP, MUSTANG was unable to process one protein each from class A, PBP and signal transduction proteins. The MUSTANG inferred phylogeny in MBLs concurred with the accepted relationship, and with the one detected by STEEP (**Fig. 4F**; **Fig. S2B**).

As can be done after any MSA, we extracted a profile from the STEEP generated MSA, and extended the initial profile provided as the input motif. The extended profiles for the SBL and MBL superfamilies were created from **Figure 4B** and **Figure 4F**, respectively, by choosing columns that have less than 75% gaps (10 in case of SBLs, 6 in the case of MBLs) (**Table 5**). These extended profiles can be considered as a better representative of the superfamilies.

It is possible to identify residues that lack electrostatic congruence by comparing these two alignments, which can be subjected to site directed mutagenesis techniques designed to mirror the specificity of the desired protein.[36] Previously, we have noted that Leu153 is the best candidate for mimicking Glu166 when we superimposed the class A SBL and the PBP-5 (PDBid:1NZO).[50] We proposed that the L153E PBP-5 mutant might provide greater success in replicating β-lactamase enzymatic efficiency in PBPs than achieved through a similar mutation in a PBP-A from *T. elongatus*.[51]

Figure 4A corroborates the Leu153-Glu166 spatial equivalence, while **Figure 4B** shows that there is no electrostatic congruence in these two residues. Another observation is that the spatially equivalent Trp154 from class D SBLs and signal transduction protein, and Glu166 in class A SBLs (**Fig. 4A**) also lack electrostatic similarity (**Fig. 4B**), although both of them are critical for catalysis.[37,38]

Discussion

We present a three-dimensional structure-based method for generating a multiple sequence alignment (MSA) of a set of proteins

Table 4. Potential and spatial congruence of the active site residues in proteins from the metallo- β-lactamase superfamily

PDB	Active site atoms (a,b,c,d)		ab	ac	ad	bc	bd	cd
1ZNB	HIS101NE2, ASP103OD1, HIS162NE2, HIS223NE2,	D	6.3	4.9	9.1	5.8	4.7	6
	Class B1	PD	124.5	152.2	168.3	27.7	43.8	16.1
1DD6	HIS79NE2, ASP81OD1, HIS139NE2, HIS197NE2,	D	6.8	5	9.4	6.1	5.2	6
	Class B1	PD	97	98.6	47.3	1.6	−49.7	−51.3
1M2X	HIS118NE2, ASP120OD1, HIS196NE2, HIS263NE2,	D	6.9	5	9.3	6.1	4.9	6.1
	Class B1	PD	59.9	100.8	6.1	40.9	−53.8	−94.8
3F9O	HIS118NE2, ASP120OD1, HIS196NE2, HIS263NE2,	D	6.8	4.8	10	5.6	5	6.2
	Class B2	PD	−109.3	−180.8	−74.4	−71.6	34.8	106.4
1JT1	HIS118NE2, ASP120OD1, HIS196NE2, HIS263NE2,	D	8	4.5	9.4	6.6	3.1	6.5
	Class B3	PD	245.3	65	152.3	−180.2	−93	87.3
1SML	HIS86NE2, ASP88OD1, HIS160NE2, HIS225NE2,	D	6.3	4.7	9.3	6	4.9	6.1
	Class B3	PD	140.3	93.7	147.2	−46.6	6.9	53.5
3LVZ	HIS103NE2, ASP105OD1, HIS177NE2, HIS242NE2,	D	6.5	4.4	9.4	6.1	4.9	6.3
	Class B3	PD	131	110.6	104.5	−20.4	−26.5	−6.1
1QH5	HIS56NE2, ASP58OD1, HIS110NE2, HIS173NE2,	D	6.5	4.5	9.4	6.6	5	6.6
	glyoxalase II	PD	246.3	249.1	254.4	2.9	8.1	5.2
1P9E	HIS149NE2, ASP151OD1, HIS234NE2, HIS302NE2,	D	6.3	4.4	9.1	6.8	5.1	6.7
	methyl parathion hydrolase	PD	14.3	−72.8	27.4	−87	13.1	100.2

D = Pairwise distance in Å. PD = Pairwise potential difference. The electrostatic potential is in dimensionless units of kT/e where k is Boltzmann's constant, T is the temperature in K and e is the charge of an electron.

which additionally incorporates electrostatic properties of the residues in the matching algorithm (STEEP). STEEP requires that the proteins have known structures, and at least one of the proteins has known active site residues. This active site motif extracts a structural profile that is then used for the comparison of proteins (e.g., in data banks). The congruence in cognate pairs, seen across various structures within the same protein superfamily (**Tables 1, 3, and 4**), is non-trivial and is an innate property of the enzymatic function. Subsequently, we applied geometrical transformations to generate a multiple superimposition of the proteins and emit a MSA based on a parameterized distance from the catalytic site. The matching algorithm can either include or exclude electrostatic considerations, resulting in two distinct alignments.

Such a technique, applied to distantly related proteins, gives better results when confined to the active site and its close neighborhood rather than including all the residues. We have shown that the chosen distance of 9 Å, which typically includes 30–50 residues, gives an equivalent phylogeny as determined by a larger number of residues.[12] This is because the active site and its vicinity have the highest 'inertia' when it comes to mutations, and thus preserves the largest information of its lineage. A comparison of the alignments that either include or exclude electrostatic congruence provides key residues that are possible candidates for mutations intending to transfer the functionality of one protein into another by directed evolution strategies.[36]

STEEP can easily be incorporated in the PALI database which provides structure-based sequence alignments for homologous proteins with known structures.[52] Currently, the PALI database uses DALI[53] to implement pairwise superimpositions and MUSTANG to superimpose multiple structures. This will extend the database to display functional relationships in the homologous protein sets.

Proteases have evolved to use different mechanisms for proteolysis.[39,54-58] Serine proteases, the most abundant class, cut peptide bonds in proteins using a well-known catalytic triad (His57, Asp102, Ser195).[39] Though His57, Asp102 and Ser195 are far apart in their primary sequence, they converge in the 3D structure to form the active site. We have compared the results obtained from STEEP to those obtained from a sequence based MSA program (ClustalW) and a structure based alignment program (MUSTANG) for a set of proteins from the chymotrypsin superfamily. While the MUSTANG superimpositions of the complete proteins was superior to the one generated by STEEP (**Fig. 1**), it should be noted that the dispersion in the active site residues in the STEEP superimposition is less. Since, STEEP generates the MSA based on the residues in the vicinity of the active site, the alignments are based on a better spatial overlap. The results obtained from all three programs generated almost equivalent phylogenies (**Fig. 2**).

The evolution of β-lactamases and the prevalence of antibiotic resistance is the subject of intense research and speculation.[59,60] We have applied STEEP to generate the phylogenetic trees for the serine (SBL) and metallo-β-lactamase (MBL) superfamilies. These relationships have been studied previously.[10-14] Class C SBLs were hypothesized to have evolved separately from class A or class D proteins.[11] Also, structural comparison has revealed a common fold between signal transduction proteins and class D enzymes.[48] For MBLs, the B3 subclass has been shown to have an independent origin as compared with that B1 and B2

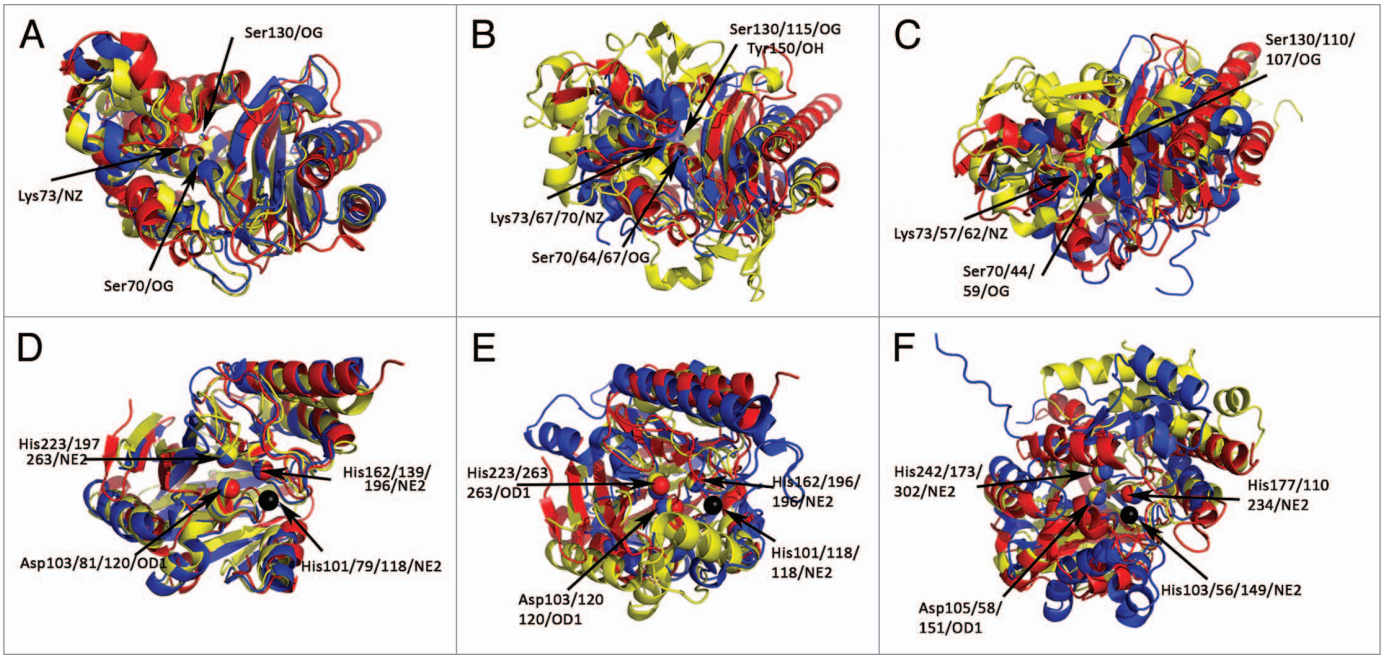

Figure 3. Superimposing multiple proteins based on the homologous active site scaffolds for serine and metallo-β-lactamases (SBL, MBL). SBL motif = (Ser70, Lys73, Ser130, Lys234), MBL motif = (His118, His196, Asp120 and His263). Ser70 and His118 are colored black and are at the center of the coordinate axes (X = 0, Y = 0, Z = 0) for SBLs and MBLs, respectively. The proteins are colored red, yellow and blue respectively in order of appearance. (**A**) Three class A SBLs - PDBids:1E25, 1I2S and 1BSG. (**B**) A class A (PDBid:1E25), a class C (PDBid:1KE4) and a class D (PDBid:3ISG) SBL. (**C**) A class A SBL (PDBid:1E25), a penicillin binding protein (PDBid:1NZO) and a signal transducer BlaR1 protein (PDBid:1XA1). (**D**) Three class B1 MBLs—PDBids:1ZNB, 1DD6 and 1M2X. (**E**) A class B1 (PDBid:1ZNB), a class B2 (PDBid:3F9O) and a class B3 (PDBid:1JT1) MBL. (**F**) A class B3 MBL (PDBid:3LVZ), a human glyoxalase II (PDBid:1QH5) and a methyl parathion hydrolase (PDBid:1P9E).

subclasses, based on both sequence and structural phylogeny. Other proteins from the MBL superfamily like human glyoxalase II and methyl parathion hydrolase are more related to the class B3 enzymes.[12,13] We demonstrate here a confirmation of the expected structural homologies in these proteins (**Fig. 3**), and our results have concurred with the hypothesized relationships in these sets of proteins (**Fig. 4**). However, the cladogram did not indicate the expected similarity in class A enzymes and certain penicillin binding proteins (PDBid:1NZ0) possibly due to the deletion of a segment from the sequence close to the active site.[49] Since MSA techniques are not applicable to the serine and metallo-β-lactamase superfamilies due to significant sequence divergence, we compared STEEP results for these two superfamilies to those generated by MUSTANG. While both methods agree on the MBL phylogenies, the cladogram generated for SBLs by MUSTANG differed from STEEP generated cladogram (the latter showing the accepted relationship).

Previous work by various groups have elucidated the discriminating powers of electrostatic properties and proposed methods for identifying residues that determine specificity of homologous proteins from different species. For example, the molecular dipole of the binding site of a ligand free structure has been used to discriminate between adenine and guanine binding sites in proteins.[61] Another work has applied electrostatic similarity indices[62] to 100 members of the Pleckstrin homology (PH) domain family, and demonstrated that the "electrostatic properties of the PH domains are generally conserved despite the extreme sequence divergence".[63] Such conservation in protein superfamilies has been established by other groups as well.[64] The electrostatic similarity index has also been applied to identify residues that are responsible for differing selectivity in the dihydrofolate reductase protein taken from different species (PIPSA).[65] This feature is similar to the one we have described in the current work. Both STEEP and PIPSA however are dependent on being able to obtain a relevant superimposition of the target protein. A superimposition independent method has been proposed to functionally classify protein structures based on properties that are invariant of affine transformations.[66,67] Such a method is particularly applicable to cases where there is little global similarity.

The inability of CLASP to distinguish between mirror images is typical of methods that use RMSD. The symmetry in the potential difference in the active site of the β-lactamase superfamily is highlighted in **Table S7**. Here, the mirror images of the correct scaffold had marginally better CLASP scores. We filtered out such images by ensuring the proper sequence order between the querying motif and the matched residues. A caveat in inferring phylogenetic relationship through structural similarity is the phenomenon of convergent evolution that achieves the same fold through a different evolutionary pathway.[39,41] The presence of a convergently evolved protein in the set might result in the detection of a homologous scaffold, but subsequently produce irrelevant results. This limitation is shared by almost all programs generating MSA of proteins, and thus requires manual inspection in pruning out unrelated proteins. Another limitation of STEEP

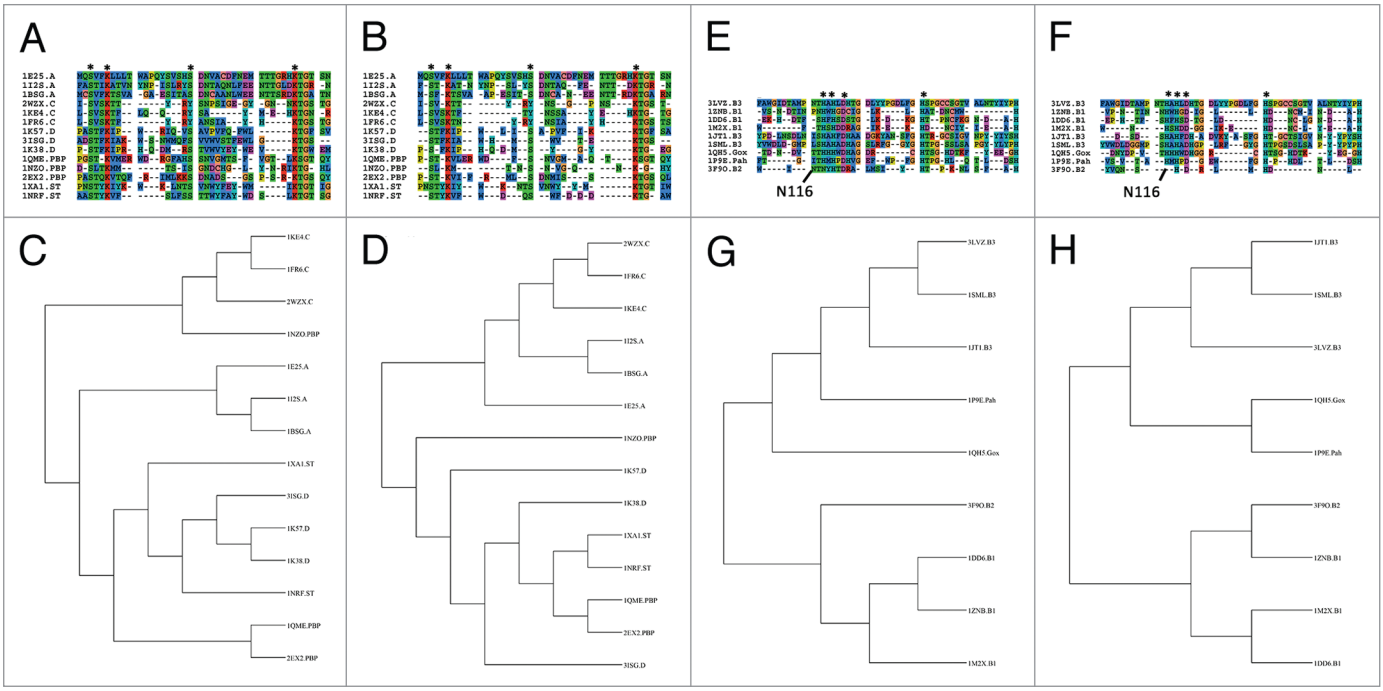

Figure 4. Multiple sequence alignments obtained using STEEP, and phylogenetic trees generated using PhyML for serine and metallo-β-lactamases (SBL, MBL). The active site motif is marked as '*'. The residues are within a radius of 9 Å from the specified active site residues. AS = alignment using spatial proximity. ASE = alignment using spatial proximity and electrostatic congruence. (**A**) AS for SBLs. (**B**) ASE for SBLs. (**C**) Cladogram generated from (A). (**D**) Cladogram generated from (B). (**E**) AS for MBLs. (**F**) ASE for MBLs. (**G**) Cladogram generated from (E). (**H**) Cladogram generated from (F).

(and other structure-only based methods) when compared with sequence based MSA methods is the requirement that the structure of the protein is to be previously known. One can use a structure prediction method to generate a likely structure to circumvent this limitation.[68] However, the accuracy of the tool used to predict the structure needs to be kept in mind when assessing the results of such a work-flow. A quantitative comparison with such structural methods is made difficult by the lack of good metrics for benchmarking structural alignments, although a recent method proposes a mathematical framework for protein structure comparison.[69] It is also to be noted that the STEEP methodology involves the residues in the active site, a demanding constraint that leads to non-optimal results as the distance from the active site increases. Thus, it does not fare as well as other methods (MUSTANG, as compared using RMSD values) that apply global and flexible constraints while superimposing, although the results improve considerably when amino acids are represented by Cα instead of the reactive atom. The approach adopted by STEEP is necessary in order to ensure the optimal superimposition of the active site, even at the cost of non-optimal results in other domains. By doing so, the identification of residues that are to be mutated is proper. Comparisons with sequence alignment methods suffer for the same reason, as well as the fact that the benchmarking suites have been shown to have been inadequately represented by structural information.[70-72] Finally, it is to be noted that STEEP is less automated than the other methods compared in the current work (ClustalW and MUSTANG), and requires a priori knowledge of the active site residues. Any comparison metric that favors STEEP should take this into consideration.

Materials and Methods

STEEP takes a set of related proteins with known structures, such that the catalytic site is known for at least one protein (the template protein) and that the chemically active side chain consists of three or more residues. These residues are used to create a query motif for analysis using CLASP. The underlying theoretical foundation for CLASP is the non-triviality of the spatial and electrostatic congruence in cognate pairs seen across various structures of the same catalytic function. CLASP extracts matching scaffolds in these related proteins, which are then superimposed. Thus, we obtain a consolidated spatial reference frame for the set of proteins. We proceed to align the residues from the template protein that are within a certain (parameterized) radius from the residues in the active site motif other spatially close residues in other proteins, providing the user an option to use electrostatic congruence as a discriminator. These steps are now described in details.

Extracting the partial scaffolds. STEEP takes as an input a set of M related proteins (**Eq. 1**) with known structures and a motif consisting of N (> = 3) residues (**Eq. 2**) from the catalytic site of one of the proteins (P1). Every amino acid is represented by a user defined atom. This is the atom whose electrostatic potential will be representative of that particular residue in the protein, just as the Cα atom represents the spatial coordinates while doing a RMSD analysis. Also, each position of the motif has a set of amino acids (**Eq. 3**) specified to allow for stereochemically equivalent matches at that particular position, such that matching amino acids of type R_i should belong to $Group_i$.

Table 5. Extending the profile

SBL			MBL		
Index	**Count**	**Amino acid types**	**Index**	**Count**	**Amino acid types**
1	10	(F/M/P/I/L)	2	6	(A/Y/E/V)
3	14	(S)	3	7	(S/W/D/P/V)
4	13	(F/T/L/V)	5	7	(H/N/I/Y/L/V)
6	14	(K)	8	7	(S/A/T/D/G)
7	14	(A/M/T/I/L/V)	12	6	(S/T/N)
8	13	(A/F/S/T/N/P/Y/I/L)	13	8	(H)
20	14	(S/Y)	14	8	(H/S/F/A/M/W)
22	10	(N)	15	9	(H)
23	13	(S/W/T/P/Y/V/M/C)	16	8	(A/F/S/W/D/P/G/L)
24	12	(F/S/A/I/G/Y/V)	17	9	(D)
30	11	(S/Q/M/D/K/P/Y/E)	19	6	(T/I/G)
37	14	(K)	20	9	(A/R/P/G)
38	13	(S/T)	22	8	(W/I/L/V)
39	14	(G)	29	6	(F/M/Y/L)
40	10	(F/S/A/T/R)	31	9	(H)
41	12	(A/S/T/E/H/Q/R/I)	32	8	(S/T/D)
42	12	(A/S/W/N/G/L/Y)	36	8	(H/T/D/N/C)
			37	8	(S/D/C)
			38	6	(T/M/I/G/L)
			39	6	(S/T/K/G/L)
			41	8	(A/T/N/P/Y)
			43	8	(D/N/L/Y/E)
			47	8	(A/D/L/Y)
			49	8	(H)

Consensus residues in the SBL and MBL superfamily with respect to spatial location and electrostatic properties. Indexing is with reference to the sequence alignment shown in **Figure 4A** and **Figure 4E** for SBL and MBL, respectively. Count is the number of proteins which have a certain amino acid in that index in the alignment. The profile is extended if there are less than 75% gaps. For SBLs, the complete set has 14 proteins, so the required count is 10. For MBLs, the complete set has 9 proteins, so the required count is 6.

All sets of N residues with the above mentioned constraints are obtained in each protein P_i using an exhaustive search procedure similar to the one used in SPASM[73] (**Eq. 4**). The pairwise distances and potential differences are computed in each match $Match^{Pi}_j$ for each protein P_i ($i \neq 1$), and are furthermore compared with the active site motif Φ^{P1}_{ASM} using a scoring function (Cscore),[32] resulting in a score which defines an ordering of the matches.

$$\Phi_{proteins} = \{P_1 \ldots P_M\} \quad (1)$$

$$\Phi^{PI}_{ASM} = \{R_1 \ldots R_N\},\ N >= 3 \quad (2)$$

$$\Phi_{groups} = \{Group_1 \ldots Group_N\} \quad (3)$$

$$\Phi^{Pi}_{matches} = \{Match_1^{Pi} \ldots Match_K^{Pi}\},$$
$$\forall(j = 1 \ldots K)[Match_j^{Pi} = \{r_1, r_2 \ldots r_N\},\ \forall(p = 1 \ldots N)[\text{AminoAcidType}(r_p) \in Group_p]],$$
$$[CScore(Match_1^{Pi}) < CScore(Match_1^{Pi}) < CScore(Match_1^{Pi}) \ldots],$$
$$CScore(Match_1^{Pi}) < Sthresh \quad (4)$$

Matches below a user defined threshold score (Sthresh) are discarded. In cases where the best match has a score of more than Sthresh, the protein is discarded under the assumption that it is not related to other proteins. The scaffolds for a protein P_i is defined as the motif with the least CScore – $Match^{Pi}_1$. The pseudocode for this function is shown in **Figure S3A**.

Superimposing the scaffolds. The scaffolds from all the M proteins are now superimposed extending the technique described previously for a pair of proteins[50] to include multiple proteins. In order to superimpose two scaffolds, $Match^{P1}_1$ and $Match^{Pi}_1$, we apply both linear and rotational transformations for all atoms in P1 and Pi such that the first three atoms {a1, a2, a3} in $Match^{P1}_1$ and $Match^{Pi}_1$ lie on the same plane (Z = 0), a1 atoms are at the center, and a2 atoms lie on the Y axis. We iterate the pairwise superimposition for the template protein with all other proteins to obtain a multiple superimposition. This superimposition is now outputted as a Pymol formatted file, and can be viewed using Pymol. The set of proteins now have a consolidated spatial reference frame. The pseudocode for this function is shown in **Figure S3B**.

Generating the alignment, and proposing mutations. Finally, we proceed to align the residues from the template protein which are within a certain (parameterized) radius Rdist from the active side residues. The set of these residues is Φ^{P1}_{align} (**Eqn. 5**). The choice of the radial distance that encompasses interacting residues has to be evaluated based on the enzymes being investigated. A small radius will not include enough residues, while a large one will include irrelevant ones. We have seen that a distance greater than 4 Å gives comparable cladograms (**Fig. S4**). The residues in the template protein within a distance of 9 Å constitute the sequence that are used for alignment in all the examples in the current work.

$$\Phi^{PI}_{align} = \{ra_1^{PI} \ldots ra_X^{PI}\}, \forall(i = 1 \ldots X, \exists j = 1 \ldots N)[dist(ra_i^{PI}, R_j) <= Rdist] \quad (5)$$

Next, for each protein P_i ($i \neq 1$), we identify residues that are in the vicinity of each of the residues in Φ^{P1}_{align}, choosing the closest residue as the alignment (**Eqn. 6**). This is possible since we have a consolidated spatial reference frame for the set of proteins ($NRes_{P_i}$ = number of residues in protein P_i). At this stage, we provide the option to use electrostatic congruence as a discriminator (**Eqn.** 7). The subroutine potcon evaluates whether the two atoms have potential congruence.

$$Align^{Pi}_{Distance} = \{raD_1^{Pi} \ldots raD_X^{Pi}], [\forall(j = 1 \ldots X), raD_j^{Pi} = \forall(q = 1 \ldots N\,\mathrm{Res}_{Pi}) mindist\,(ra_q^{Pi}, ra_j^{P1})] \quad (6)$$

$$Align^{Pi}_{Potential} = \{raP_1^{Pi} \ldots raP_X^{Pi}], [\forall(j = 1 \ldots X), raP_p^{Pi} = raD_p^{Pi} \wedge potcon(ra_q^{Pi}, ra_p^{P1})] \quad (7)$$

Two kinds of alignments are obtained by either ignoring potential congruence or filtering out residues that do not have electrostatic congruenc. The pseudocode for this function is shown in **Figure S3C**. A comparison of these alignments identifies residues which lack electrostatic congruence, even though they occupy a spatially equivalent position in the structure. These can be the basis of mutations in directed evolution experiments designed to mirror the desired protein and its functionality/specificity.

Implementation details and third party software. The STEEP package is written in Perl on Ubuntu. Hardware requirements are modest - all results here are from a simple workstation (2GB RAM) and runtimes were a few minutes at the most. The source code and manual are made available at www.sanchak.com/steep.html.

Adaptive Poisson-Boltzmann Solver (APBS) and PDB2PQR packages were used to calculate the potential difference between the reactive atoms of the corresponding proteins.[74,75] The APBS parameters and electrostatic potential units were set as described previously in.[32] The invariance in the electrostatic features (measured in structures that have been solved independently over many years) also speaks highly of the reliability of the APBS/PDB2PQR implementation.

All protein structures were rendered by PyMol (http://www.pymol.org/). The alignment and cladograms images were generated using Seaview.[76] We have used PHYML to generate phylogenetic trees from these alignments, which uses the method of maximum likelihood.[8] The method searches for a tree with the highest probability or likelihood that would give rise to the observed data set, given a proposed model of evolution and the hypothesized history. The LG model is the chosen evolutionary model providing the amino acid replacement matrices, and is the default setting in PHYML.[77] Although, such methods are computationally intensive, they are robust to the choice of the evolutionary model and outperform alternative techniques methods (parsimony or distance methods).[78]

Availability of supporting data. The source code and manual are made available at www.sanchak.com/steep.html.

Conclusions

To summarize, we propose a MSA methodology that generates both evolutionary and functional relation-ships, eliminates unrelated proteins from the computation, emits a multiple superimposition of the related proteins and demonstrate that the active site vicinity contains enough information to infer correct kinship in a set of related proteins. A unique feature of STEEP is the ability to identify residues that can be targeted by directed evolution experiments in order to endow enzymatic functionality of one member of a superfamily to another.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

This work was funded by the Tata Institute of Fundamental Research (Department of Atomic Energy), and the Department of Science and Technology (JC Bose Award Grant). BA extends gratitude to the Icelandic National Research Council (RANNIS) and the University of Iceland Research Found for supporting the project financially. NAB was supported by NIH grants R01 GM069702 and P41 RR0860516.

Authors' Contributions

SC, BJR, and BA helped designing the study and analyses. NB participated in the analysis of data. SC did the programming in the study. All authors participated in writing the paper and they all approved the final manuscript.

References

1. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008; 9:387-402; PMID:18576944; http://dx.doi.org/10.1146/annurev.genom.9.081307.164359.
2. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970; 48:443-53; PMID:5420325; http://dx.doi.org/10.1016/0022-2836(70)90057-4.
3. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981; 147:195-7; PMID:7265238; http://dx.doi.org/10.1016/0022-2836(81)90087-5.
4. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007; 23:2947-8; PMID:17846036; http://dx.doi.org/10.1093/bioinformatics/btm404.
5. Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, et al. T-Co ee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res 2011; 39:W13-17; http://dx.doi.org/10.1093/nar/gkr245.
6. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 2005; 15:330-40; PMID:15687296; http://dx.doi.org/10.1101/gr.2821705.

7. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004; 5:113; PMID:15318951; http://dx.doi.org/10.1186/1471-2105-5-113.
8. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online-a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 2005; 33:W557-559; http://dx.doi.org/10.1093/nar/gki352.
9. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 2001; 17:754-5; PMID:11524383; http://dx.doi.org/10.1093/bioinformatics/17.8.754.
10. Hall BG, Barlow M. Evolution of the serine β-lactamases: past, present and future. Drug Resist Updat 2004; 7:111-23; PMID:15158767; http://dx.doi.org/10.1016/j.drup.2004.02.003.
11. Hall BG, Barlow M. Structure-based phylogenies of the serine β-lactamases. J Mol Evol 2003; 57:255-60; PMID:14629035; http://dx.doi.org/10.1007/s00239-003-2473-y.
12. Garau G, Di Guilmi AM, Hall BG. Structure-based phylogeny of the metallo-β-lactamases. Antimicrob Agents Chemother 2005; 49:2778-84; PMID:15980349; http://dx.doi.org/10.1128/AAC.49.7.2778-2784.2005.
13. Hall BG, Salipante SJ, Barlow M. The metallo-β-lactamases fall into two distinct phylogenetic groups. J Mol Evol 2003; 57:249-54; PMID:14629034; http://dx.doi.org/10.1007/s00239-003-2471-0.
14. Hall BG, Salipante SJ, Barlow M. Independent origins of subgroup Bl + B2 and subgroup B3 metallo-β-lactamases. J Mol Evol 2004; 59:133-41; PMID:15383916; http://dx.doi.org/10.1007/s00239-003-2572-9.
15. Pandit SB, Skolnick J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics 2008; 9:531; PMID:19077267; http://dx.doi.org/10.1186/1471-2105-9-531.
16. Holm L, Sander C. Mapping the protein universe. Science 1996; 273:595-603; PMID:8662544; http://dx.doi.org/10.1126/science.273.5275.595.
17. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 2003; 19(Suppl 2):ii246-55; PMID:14534198; http://dx.doi.org/10.1093/bioinformatics/btg1086.
18. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998; 11:739-47; PMID:9796821; http://dx.doi.org/10.1093/protein/11.9.739.
19. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002; 11:2606-21; PMID:12381844; http://dx.doi.org/10.1110/ps.0215902.
20. Sippl MJ, Wiederstein M. A note on difficult structure alignment problems. Bioinformatics 2008; 24:426-7; PMID:18174182; http://dx.doi.org/10.1093/bioinformatics/btm622.
21. Ochagavía ME, Wodak S. Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins. Proteins 2004; 55:436-54; PMID:15048834; http://dx.doi.org/10.1002/prot.10587.
22. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: a multiple structural alignment algorithm. Proteins 2006; 64:559-74; PMID:16736488; http://dx.doi.org/10.1002/prot.20921.
23. Dror O, Benyamini H, Nussinov R, Wolfson HJ. Multiple structural alignment by secondary structures: algorithm and applications. Protein Sci 2003; 12:2492-507; PMID:14573862; http://dx.doi.org/10.1110/ps.03200603.
24. Ye Y, Godzik A. Multiple flexible structure alignment using partial order graphs. Bioinformatics 2005; 21:2362-9; PMID:15746292; http://dx.doi.org/10.1093/bioinformatics/bti353.
25. Nguyen MN, Tan KP, Madhusudhan MS. CLICK-topology-independent comparison of biomolecular 3D structures. Nucleic Acids Res 2011; 39:W24-28.
26. Madhusudhan MS, Webb BM, Marti-Renom MA, Eswar N, Sali A. Alignment of multiple protein structures based on sequence and structure features. Protein Eng Des Sel 2009; 22:569-74; PMID:19587024; http://dx.doi.org/10.1093/protein/gzp040.
27. Shatsky M, Nussinov R, Wolfson HJ. A method for simultaneous alignment of multiple protein structures. Proteins 2004; 56:143-56; PMID:15162494; http://dx.doi.org/10.1002/prot.10628.
28. Lu G. TOP: a new method for protein structure comparisons and similarity searches. J Appl Cryst 2000; 33:176-83; http://dx.doi.org/10.1107/S0021889899012339.
29. Shealy P, Valafar H. Multiple structure alignment with msTALI. BMC Bioinformatics 2012; 13:105; PMID:22607234; http://dx.doi.org/10.1186/1471-2105-13-105.
30. Menke M, Berger B, Cowen L. Matt: local flexibility aids protein multiple structure alignment. PLoS Comput Biol 2008; 4:e10; PMID:18193941; http://dx.doi.org/10.1371/journal.pcbi.0040010.
31. Micheletti C, Orland H. MISTRAL: a tool for energy-based multiple structural alignment of proteins. Bioinformatics 2009; 25:2663-9; PMID:19692555; http://dx.doi.org/10.1093/bioinformatics/btp506.
32. Chakraborty S, Minda R, Salaye L, Bhattacharjee SK, Rao BJ. Active site detection by spatial conformity and electrostatic analysis--unravelling a proteolytic function in shrimp alkaline phosphatase. PLoS One 2011; 6:e28470; PMID:22174814; http://dx.doi.org/10.1371/journal.pone.0028470.
33. Helland R, Larsen RL, Asgeirsson B. The 1.4 A crystal structure of the large and cold-active *Vibrio sp*. alkaline phosphatase. Biochim Biophys Acta 2009; 1794:297-308; PMID:18977465; http://dx.doi.org/10.1016/j.bbapap.2008.09.020.
34. Chakraborty S, Asgeirsson B, Minda R, Salaye L, Frère JM, Rao BJ. Inhibition of a cold-active alkaline phosphatase by imipenem revealed by *in silico* modeling of metallo-β-lactamase active sites. FEBS Lett 2012; 586:3710-5; PMID:22982109; http://dx.doi.org/10.1016/j.febslet.2012.08.030.
35. Errami M, Geourjon C, Deléage G. Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. Bioinformatics 2003; 19:506-12; PMID:12611806; http://dx.doi.org/10.1093/bioinformatics/btg016.
36. Reetz MT, Carballeira JD. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. Nat Protoc 2007; 2:891-903; PMID:17446890; http://dx.doi.org/10.1038/nprot.2007.72.
37. Baurin S, Vercheval L, Bouillenne F, Falzone C, Brans A, Jacquamet L, et al. Critical role of tryptophan 154 for the activity and stability of class D β-lactamases. Biochemistry 2009; 48:11252-63; PMID:19860471; http://dx.doi.org/10.1021/bi901548c.
38. Hata M, Fujii Y, Ishii M, Hoshino T, Tsuda M. Catalytic mechanism of class A β-lactamase. I. The role of Glu166 and Serl30 in the deacylation reaction. Chem Pharm Bull (Tokyo) 2000; 48:447-53; PMID:10783059; http://dx.doi.org/10.1248/cpb.48.447.
39. Rawlings ND, Barrett AJ. Evolutionary families of peptidases. Biochem J 1993; 290:205-18; PMID:8439290.
40. Sárkány Z, Polgár L. The unusual catalytic triad of poliovirus protease 3C. Biochemistry 2003; 42:516-22; PMID:12525179; http://dx.doi.org/10.1021/bi027004w.
41. Gherardini PF, Wass MN, Helmer-Citterich M, Sternberg MJ. Convergent evolution of enzyme active sites is not a rare phenomenon. J Mol Biol 2007; 372:817-45; PMID:17681532; http://dx.doi.org/10.1016/j.jmb.2007.06.017.
42. Iengar P, Ramakrishnan C. Knowledge-based modeling of the serine protease triad into non-proteases. Protein Eng 1999; 12:649-56; PMID:10469825; http://dx.doi.org/10.1093/protein/12.8.649.
43. Ghuysen JM. Serine β-lactamases and penicillin-binding proteins. Annu Rev Microbiol 1991; 45:37-67; PMID:1741619; http://dx.doi.org/10.1146/annurev.mi.45.100191.000345.
44. Bebrone C. Metallo-β-lactamases (classification, activity, genetic organization, structure, zinc coordination) and their superfamily. Biochem Pharmacol 2007; 74:1686-701; PMID:17597585; http://dx.doi.org/10.1016/j.bcp.2007.05.021.
45. Tranier S, Bouthors AT, Maveyraud L, Guillet V, Sougakoff W, Samama JP. The high resolution crystal structure for class A β-lactamase PER-1 reveals the bases for its increase in breadth of activity. J Biol Chem 2000; 275:28075-82; PMID:10825176.
46. Garau G, Bebrone C, Anne C, Galleni M, Frère JM, Dideberg O. A metallo-β-lactamase enzyme in action: crystal structures of the monozinc carbapenemase CphA and its complex with biapenem. J Mol Biol 2005; 345:785-95; PMID:15588826; http://dx.doi.org/10.1016/j.jmb.2004.10.070.
47. Bebrone C, Delbrück H, Kupper MB, Schlömer P, Willmann C, Frère JM, et al. The structure of the dizinc subclass B2 metallo-β-lactamase CphA reveals that the second inhibitory zinc ion binds in the histidine site. Antimicrob Agents Chemother 2009; 53:4464-71; PMID:19651913; http://dx.doi.org/10.1128/AAC.00288-09.
48. Zhu YF, Curran IH, Joris B, Ghuysen JM, Lampen JO. Identification of BlaR, the signal transducer for β-lactamase production in Bacillus licheniformis, as a penicillin-binding protein with strong homology to the OXA-2 β-lactamase (class D) of Salmonella typhimurium. J Bacteriol 1990; 172:1137-41; PMID:2404938.
49. Nicholas RA, Krings S, Tomberg J, Nicola G, Davies C. Crystal structure of wild-type penicillin-binding protein 5 from *Escherichia coli*: implications for deacylation of the acyl-enzyme complex. J Biol Chem 2003; 278:52826-33; PMID:14555648; http://dx.doi.org/10.1074/jbc.M310177200.
50. Chakraborty S. An automated flow for directed evolution based on detection of promiscuous scaffolds using spatial and electrostatic properties of catalytic residues. PLoS One 2012; 7:e40408; PMID:22811760; http://dx.doi.org/10.1371/journal.pone.0040408.
51. Urbach C, Evrard C, Pudzaitis V, Fastrez J, Soumillion P, Declercq JP. Structure of PBP-A from *Thermosynechococcus elongatus*, a penicillin-binding protein closely related to class A β-lactamases. J Mol Biol 2009; 386:109-20; PMID:19100272; http://dx.doi.org/10.1016/j.jmb.2008.12.001.
52. Gowri VS, Pandit SB, Karthik PS, Srinivasan N, Balaji S. Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. Nucleic Acids Res 2003; 31:486-8; PMID:12520058; http://dx.doi.org/10.1093/nar/gkg063.
53. Holm L, Kääriäinen S, Rosenström P, Schenkel A. Searching protein structure databases with DaliLite v.3. Bioinformatics 2008; 24:2780-1; PMID:18818215; http://dx.doi.org/10.1093/bioinformatics/btn507.
54. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 2012; 40(Database issue):D343-50; PMID:22086950; http://dx.doi.org/10.1093/nar/gkr987.
55. Lodola A, Branduardi D, De Vivo M, Capoferri L, Mor M, Piomelli D, et al. A catalytic mechanism for cysteine N-terminal nucleophile hydrolases, as revealed by free energy simulations. PLoS One 2012; 7:e32397; PMID:22389698; http://dx.doi.org/10.1371/journal.pone.0032397.

56. Ekici OD, Paetzel M, Dalbey RE. Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration. Protein Sci 2008; 17:2023-37; PMID:18824507; http://dx.doi.org/10.1110/ps.035436.108.
57. Rosenblum G, Van den Steen PE, Cohen SR, Bitler A, Brand DD, Opdenakker G, et al. Direct visualization of protease action on collagen triple helical structure. PLoS One 2010; 5:e11043; PMID:20585385; http://dx.doi.org/10.1371/journal.pone.0011043.
58. Rawlings ND, Barrett AJ, Bateman A. Asparagine peptide lyases: a seventh catalytic type of proteolytic enzymes. J Biol Chem 2011; 286:38321-8; PMID:21832066; http://dx.doi.org/10.1074/jbc.M111.260026.
59. Bhullar K, Waglechner N, Pawlowski A, Koteva K, Banks ED, Johnston MD, et al. Antibiotic resistance is prevalent in an isolated cave microbiome. PLoS One 2012; 7:e34953; PMID:22509370; http://dx.doi.org/10.1371/journal.pone.0034953.
60. Bush K, Jacoby GA. Updated functional classification of beta-lactamases. Antimicrob Agents Chemother 2010; 54:969-76; PMID:19995920; http://dx.doi.org/10.1128/AAC.01009-09.
61. Basu G, Sivanesan D, Kawabata T, Go N. Electrostatic potential of nucleotide-free protein is sufficient for discrimination between adenine and guanine-specific binding sites. J Mol Biol 2004; 342:1053-66; PMID:15342256; http://dx.doi.org/10.1016/j.jmb.2004.07.047.
62. Hodgkin EE, Richards WG. Molecular Similarity Based on Electrostatic Potential and Electric-Field. Int J Quantum Chem 1987; •••:105-10; http://dx.doi.org/10.1002/qua.560320814.
63. Blomberg N, Gabdoulline RR, Nilges M, Wade RC. Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. Proteins 1999; 37:379-87; PMID:10591098; http://dx.doi.org/10.1002/(SICI)1097-0134(19991115)37:3<379::AID-PROT6>3.0.CO;2-K.
64. Livesay DR, Jambeck P, Rojnuckarin A, Subramaniam S. Conservation of electrostatic properties within enzyme families and superfamilies. Biochemistry 2003; 42:3464-73; PMID:12653550; http://dx.doi.org/10.1021/bi026918f.
65. Henrich S, Richter S, Wade RC. On the use of PIPSA to guide target-selective drug design. ChemMedChem 2008; 3:413-7; PMID:18061917; http://dx.doi.org/10.1002/cmdc.200700154.
66. Zhang X, Bajaj C, Baker NA Affine Invariant Comparison of Molecular Shapes with Properties. 2004; Computer Science Technical Report.
67. Zhang X, Bajaj CL, Kwon B, Dolinsky TJ, Nielsen JE, Baker NA. Application of new multi-resolution methods for the comparison of biomolecular electrostatic properties in the absence of global structural similarity. Multiscale Model Simul 2006; 5:1196-213; PMID:18841247; http://dx.doi.org/10.1137/050647670.
68. Combet C, Jambon M, Deléage G, Geourjon C. Geno3D: automatic comparative molecular modelling of protein. Bioinformatics 2002; 18:213-4; PMID:11836238; http://dx.doi.org/10.1093/bioinformatics/18.1.213.
69. Liu W, Srivastava A, Zhang J. A mathematical framework for protein structure comparison. PLoS Comput Biol 2011; 7:e1001075; PMID:21304929; http://dx.doi.org/10.1371/journal.pcbi.1001075.
70. Edgar RC. Quality measures for protein alignment benchmarks. Nucleic Acids Res 2010; 38:2145-53; PMID:20047958; http://dx.doi.org/10.1093/nar/gkp1196.
71. Dickson RJ, Wahl LM, Fernandes AD, Gloor GB. Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein covariation. PLoS One 2010; 5:e11082; PMID:20596526; http://dx.doi.org/10.1371/journal.pone.0011082.
72. Aniba MR, Poch O, Thompson JD. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Res 2010; 38:7353-63; PMID:20639539; http://dx.doi.org/10.1093/nar/gkq625.
73. Kleywegt GJ. Recognition of spatial motifs in protein structures. J Mol Biol 1999; 285:1887-97; PMID:9917419; http://dx.doi.org/10.1006/jmbi.1998.2393.
74. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A 2001; 98:10037-41; PMID:11517324; http://dx.doi.org/10.1073/pnas.181342398.
75. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res 2004; 32(Web Server issue):W665-7; PMID:15215472; http://dx.doi.org/10.1093/nar/gkh381.
76. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 2010; 27:221-4; PMID:19854763; http://dx.doi.org/10.1093/molbev/msp259.
77. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol 2008; 25:1307-20; PMID:18367465; http://dx.doi.org/10.1093/molbev/msn067.
78. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 2011; 28:2731-9; PMID:21546353; http://dx.doi.org/10.1093/molbev/msr121.