

RESEARCH

Open Access

Feature amplified voting algorithm for functional analysis of protein superfamily

Che-Lun Hung¹, Chihan Lee², Chun-Yuan Lin³, Chih-Hung Chang¹, Yeh-Ching Chung^{1*}, Chuan Yi Tang¹

From The ISIBM International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing (IJCBS)

Shanghai, China. 3-8 August 2009

Abstract

Background: Identifying the regions associated with protein function is a singularly important task in the post-genomic era. Biological studies often identify functional enzyme residues by amino acid sequences, particularly when related structural information is unavailable. In some cases of protein superfamilies, functional residues are difficult to detect by current alignment tools or evolutionary strategies when phylogenetic relationships do not parallel their protein functions. The solution proposed in this study is Feature Amplified Voting Algorithm with Three-profile alignment (FAVAT). The core concept of FAVAT is to reveal the desired features of a target enzyme or protein by voting on three different property groups aligned by three-profile alignment method. Functional residues of a target protein can then be retrieved by FAVAT analysis. In this study, the amidohydrolase superfamily was an interesting case for verifying the proposed approach because it contains divergent enzymes and proteins.

Results: The FAVAT was used to identify critical residues of mammalian imidase, a member of the amidohydrolase superfamily. Members of this superfamily were first classified by their functional properties and sources of original organisms. After FAVAT analysis, candidate residues were identified and compared to a bacterial hydantoinase in which the crystal structure (1GKQ) has been fully elucidated. One modified lysine, three histidines and one aspartate were found to participate in the coordination of metal ions in the active site. The FAVAT analysis also redressed the misrecognition of metal coordinator Asp57 by the multiple sequence alignment (MSA) method. Several other amino acid residues known to be related to the function or structure of mammalian imidase were also identified.

Conclusions: The FAVAT is shown to predict functionally important amino acids in amidohydrolase superfamily. This strategy effectively identifies functionally important residues by analyzing the discrepancy between the sequence and functional properties of related proteins in a superfamily, and it should be applicable to other protein families.

Background

(The software is freely available for download from reference [1]).

Retrieving useful functional/structural information from a set of amino acid sequences is essential in experimental biological studies. Desired information is often obtainable by analyzing the sequence

conservations, functional correlations and related structures that belong to a protein/enzyme family or superfamily. An enzyme superfamily is defined as a group of proteins that share the same structural scaffold and that undergo fundamentally similar chemical reactions [2]. Earlier studies [3-5] adopted various pair-wise alignment and multiple sequence alignment (MSA) methods to detect the conserved residues that reveal functional roles in a set of sequences. Classical sequence comparison tools such as FASTA [6], BLAST [3], CLUSTALW [7], T-COFFEE [8] and

* Correspondence: ychung@cs.nthu.edu.tw

¹Department of Computer Science, National Tsing Hua University, 101, Section 2 Kuang Fu Road, Hsinchu, Taiwan 300

Full list of author information is available at the end of the article

MUSCLE [9] can detect similarities in aligned sequences and identify the conserved positions. These positions are essential for further functional analysis. Hierarchical analysis [10-12] is often used to select the most desirable pattern of an alignment. Some protein groups with dissimilar sequences but substantial structural fold similarity (hereinafter referred to as remote homologues) have similar or related biochemical functions [13]. These proteins can be classified into the same superfamily according to their biological properties. Due to their low overall similarity, using alignment methods alone may not reveal the amino acid residues that reflect their physicochemical properties.

In addition to alignment methods, the most common strategy for predicting functional residues from sequences is motif-based sequence analysis [14-17]. However, the motif-based approach often obtains excessive false positives, which limits its use for analyzing a protein superfamily. Phylogenomic techniques such as the evolutionary trace method of identifying functionally important residues [18] use evolutionary information to improve accuracy and are particularly useful for large-scale analyses. This method automatically relates the results back to a given structure and identifies key features structurally clustered around substrate and dimer interfaces [19-22]. This tool is useful for analyzing protein or enzyme superfamilies and for extracting functional information from enzyme families or superfamilies when the phylogenetic tree or dendrogram approximates a functional distribution.

This study employs a voting concept to search for functional key residues in an enzyme superfamily. Voting or voting-like concepts are widely used in computing algorithms for various purposes. In computational biological applications, voting concepts are often integrated with neural networks for protein clustering and structure prediction [23]. Some theoretical analyses [24-29] indicate that comparing three sequences is better than comparing two sequences because it increases the alignment power needed to distinguish significant matches. Likewise, aligning three groups provides more information than aligning two groups does. Therefore, we developed a Feature Amplified Voting Algorithm with Three-profile alignment (FAVAT) according to the observed sequence similarity and biochemical properties of proteins in the amidohydrolase superfamily. The FAVAT identifies the key residues by calculating a score for each residue in a rat imidase. The functional residues of a rat imidase were identified and further confirmed by experimental references and available structural information.

Results and discussion

Case study: Imidase-related proteins in amidohydrolase superfamily

In this study, rat imidase was the target sequence, and DRPs (Group II proteins) were classified into ~A proteins. Bacterial hydantoinases (Group III enzymes) were classified as A proteins. Although Dihydroorotase, allantoinase and other amidohydrolases (Group IV) were also classified as A proteins, they differ from the Group II enzymes in their functional correlation to target sequence (rat imidase). Following the above classification, the clustered sequences were subjected to FAVAT analysis, and two sets of scores were obtained for each residue of the target sequence. In the experiments, Groups II, III and IV were aligned using the MUSCLE tool adopted by the National Center for Biotechnology Information (NCBI) for protein database alignment.

Voting scores of imidase by FAVAT analysis

Figure 1 shows the FAVAT analysis results. The T-score suggests the importance of each residue of the target sequence (rat imidase) after accumulating the total voting scores (V-scores), each V-score is calculated from the target sequence, an A protein sequence and a ~A protein sequence. For each rat imidase residue, two sets of T-scores were obtained from the Group II-Group III votes and the Group II-Group IV votes. Ten amino acids (Ala34, His67, His69, Ala134, Lys159, His248, Met297, Arg302, Asp326 and His459) were further analyzed after merging the two sets of the higher T-scores over 60. Table 1 summarizes the amino acid residues selected by FAVAT and MSA analyses and their corresponding locations in two proteins with known crystal structures [PDB:1GKQ, PDB:1KCX].

Comparison of FAVAT and MSA results

Biologists often use MSA to conjecture important residues of proteins or enzymes of interest among their related sequences. Figure 2 shows the MSA fragments of rat imidase, hamster dihydroorotase domain, yeast allantoinase and *Bacillus* sp. D-hydantoinase. The MSA analyses revealed that one aspartate and four histidines were highly conserved. The results in Fig. 2 are consistent with data published previously [30-32].

In their studies [30-32], residues Asp57, His59, His61, His183 and His239 were hypothesized to be critical amino acids for the metal coordinators and function of D-hydantoinase of *Thermus* sp. Notably, the study [33] reported that a crystal structure of hydantoinase (1GKQ), in which a carboxyl-lysine is responsible for metal binding and is important for enzyme activity, was not revealed by MSA analysis. Another residue, Asp57, which was incorrectly identified as a metal coordinator

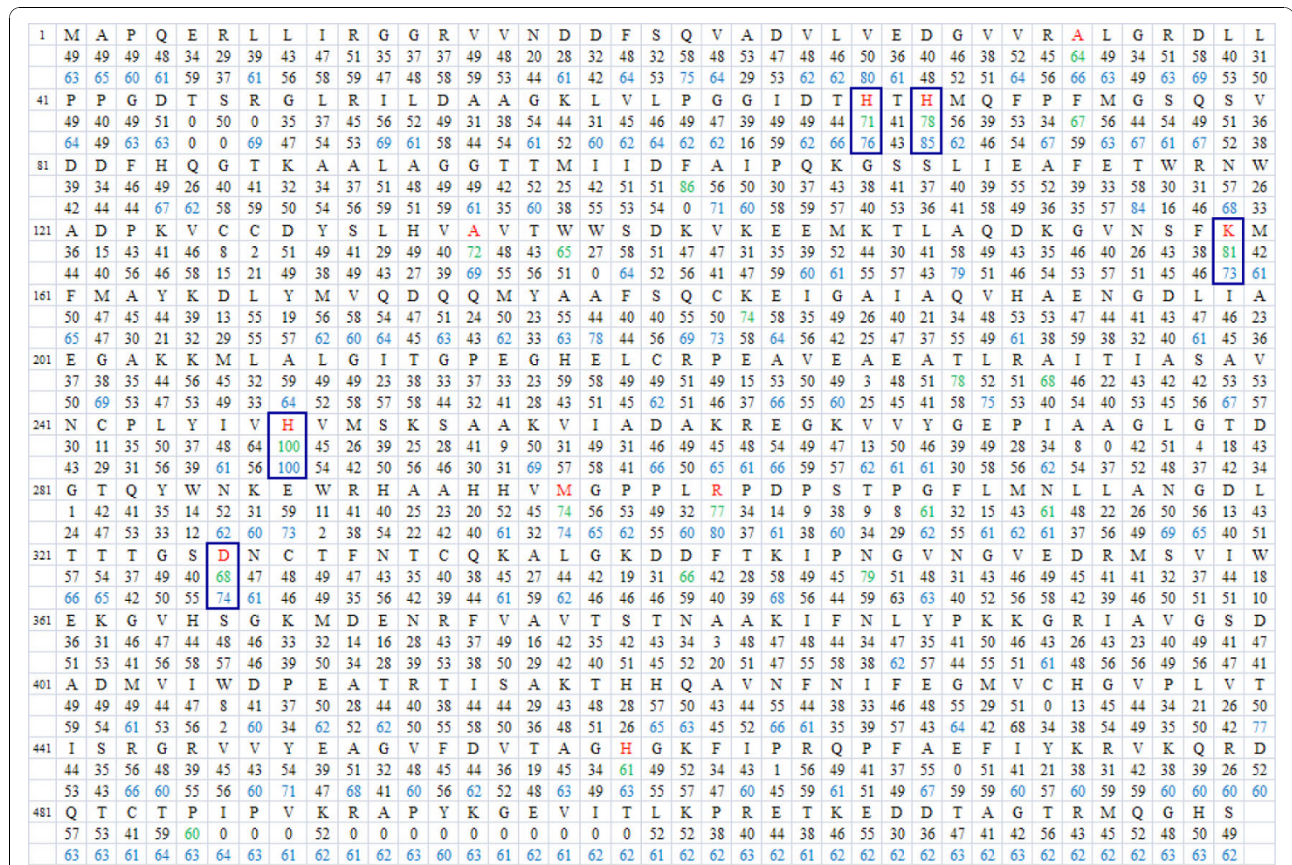


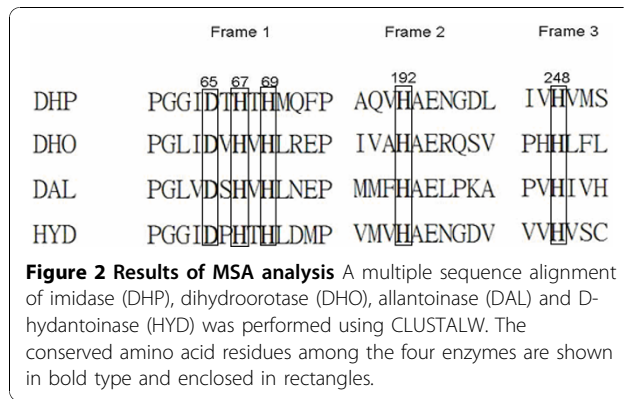
Figure 1 Total voting scores (T-scores) of rat imidase by FAVAT analysis The scores were obtained by voting algorithm as described in Methods. The first row is the rat imidase sequence. The scores in the second and third rows are accumulated V-scores (T-score) for the Group II-Group III and Group II-Group IV votes, respectively. Amino acid residues are marked in red if both votes resulted in T-scores over 60. The T-scores over 60 for Group II-Group III and Group II-Group IV votes appear in green and blue, respectively. The boxed residues indicate the residues corresponding to metal coordination in bacterial hydantoinase (1GKQ).

Table 1 Functional annotations of the residues in rat imidase selected by FAVAT

FAVAT score ranking	FAVAT selected residues	Corresponding residues in 1GKQ ¹	Corresponding residues in 1KCX ¹	MSA predicted residues ²	Functional annotation base on 1GKQ
1	His248	His239	Lys254	His239	Metal coordinate
2	His69	His61	Tyr75	His61	Metal coordinate
3	Arg302	Arg292	Ser308		Conformation
4	Lys159	Lys150 (Kcx)	Gln165		Metal coordinate
5	Met297	Met287	Thr303		Conformation
6	His67	His59	Asn73	His59	Metal coordinate
7	Asp326	Asp315	Gly332		Metal coordinate
8	Ala134	Ala126	Asp139		Secondary structure core residue
9	Ala34	Arg30	Gln44		Quaternary structure
10	His459	Trp448	Met465		Quaternary structure

¹GKQ and 1KCX are the structure IDs of D-hydantoinase and collapsin response mediator proteins (CRMPs), respectively.

²MSA result according to the manuscript [66].



in a previous study that applied MSA method as the corresponding residue in 1GKQ, revealed no involvement in metal coordination. Table 1 shows that, in the current study, FAVAT successfully identified all known important residues in rat imidase. The Lys159, His67, His69, His248 and Asp326, which correspond to Lys150, His59, His61, His239 and Asp315 in hydantoinase

(1GKQ), are metal ion coordinators (boxed amino acids in Fig. 3).

The corresponding locations of FAVAT-selected residues in 1GKQ and 1KCX

The possible functions of imidase amino acids selected by FAVAT were further analyzed using 1GKQ and 1KCX, which are known structures of imidase related proteins. The former is the crystal structure of a D-hydantoinase that represents an A protein (Group III) in FAVAT analysis. The latter is the crystal structure of a dihydropyrimidinase-related protein (CRMP1) that represents a ~A protein (Group II) in FAVAT analysis. Figure 3 shows their corresponding sequences and secondary structures. The similar β/α core structures were observed in the wiring diagrams of 1GKQ and 1KCX. The significant difference in these structures is that 1GKQ forms a typical $(\beta/\alpha)_8$ domain, but 1KCX does not. The FAVAT-selected amino acids may reflect both the structure feature and metal requirement that are responsible for the different functions of the A and

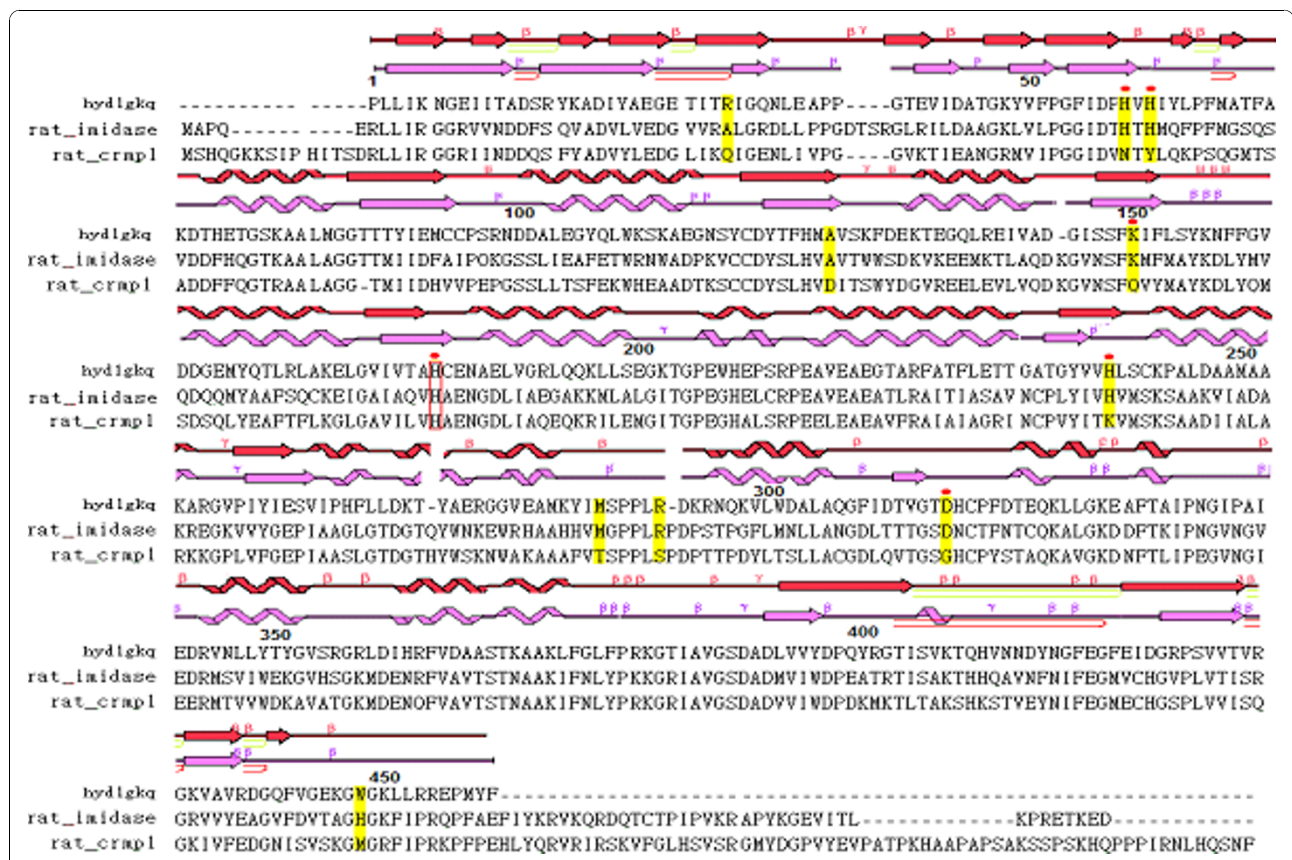
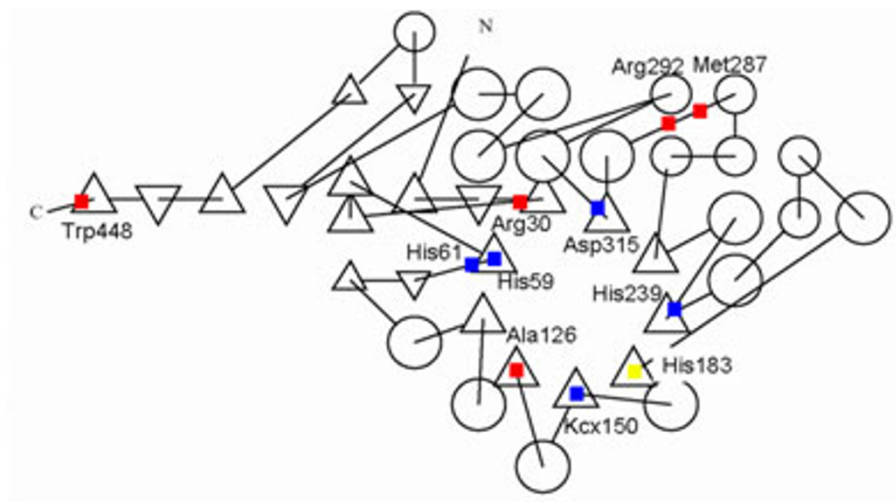
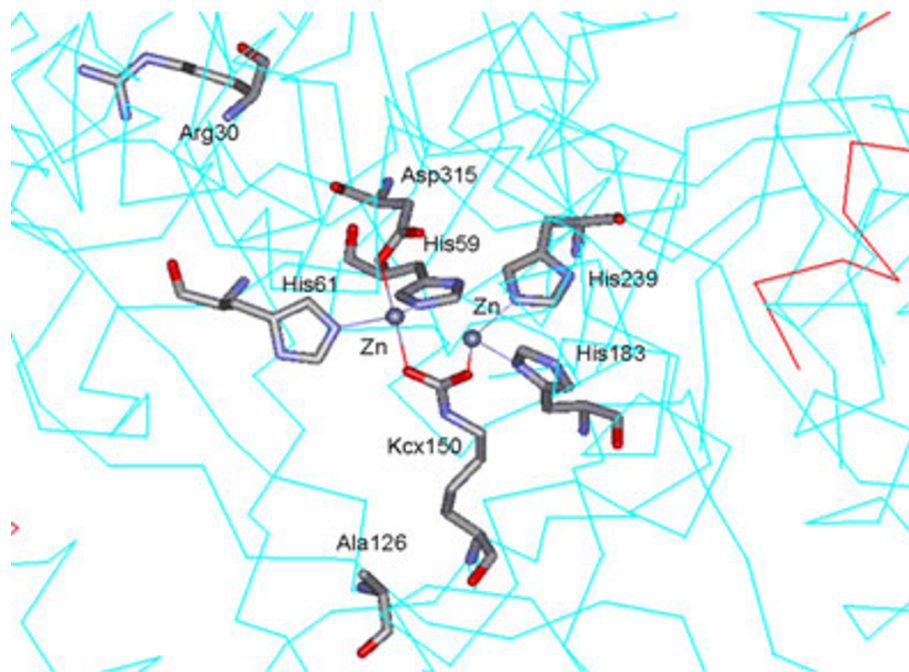


Figure 3 Corresponding locations of FAVAT-selected residues in the wiring diagrams of 1KCX and 1GKQ The secondary structures are those of CRMP-1 (1KCX, red font) and D-hydantoinase (1GKQ, blue font), respectively. The sequences are for D-hydantoinase (hyd1gkq), rat imidase (rat_imidase) and rat CRMP1 protein (rat_crmp1). The top ten residues obtained by FAVAT analysis are highlighted in yellow. Residue His183, which was not among the top ten residues selected by FAVAT, is boxed in red. Residues indicated by dotted lines were those reported for metal coordination of hydantoinase.



(a)



(b)

Figure 4 Corresponding locations of amino acid residues identified by FAVAT in 1GKQ (a) The top ten residues selected by FAVAT on the topology of 1GKQ. Carboxylated lysine residue (Kcx150), His59, His61, His239 and Asp315 that were involved in metal coordination are indicated by blue font. Residue His183 (yellow font) was not identified in this study. Other selected residues are indicated by red font. (b) The metal binding site of 1GKQ. The corresponding amino acids identified by FAVAT are highlighted.

~A proteins. The corresponding locations of Ala34 and His459 in 1GKQ (Arg30 and Trp448 in the N-terminal and C-terminal β -Sheet, respectively) and in 1KCX (Gln44 and Met465 in the N-terminal and C-terminal β -Sheet, respectively) were domains in which they

interact with another monomer to form a quaternary structure in both hydantoinase and CRMP1 [34,35]. Residues His67, His69, Ala134, Lys159, His248 and Asp326 (His59, His61, Ala126, Kcx150, His239 and Asp315 in 1GKQ; His73, Tyr75, Asp139, Gln165,

Lys254 and Gly332 in 1KCX) are located in the β/α core region.

Figure 4a is a topological view of the corresponding FAVAT-selected amino acids in 1GKQ. *Thermus sp.* D-hydantoinase (1GKQ) contains two divalent metal ions in active site (Fig. 4b). The central binuclear zinc center is bridged by the carboxylated lysine residue (Kcx150) and a hydroxide ion. Residues His59, His61, His183, His239 and Asp315 correspond to the active site zinc ion (Fig. 4b). The corresponding residues were conserved in other members of the amidohydrolase superfamily [36,37]. However, in CRMP1, four of the five corresponding residues are diverse [38]. These residues were all identified by FAVAT analysis. Other candidates recognized by FAVAT analysis, Ala134, Arg302, Ala34, Met297 and His459, were also found to reside in positions critical to protein function. The corresponding residue of Ala134 in 1GKQ (Ala126), was located near the active site (Fig. 4b). Residues Arg302 and Met297 (Arg292 and Met287 in 1GKQ, respectively) were located at the helix-loop domain outside the $(\beta/\alpha)_8$ catalytic domain. This implies that they may be important for maintaining structure or stabilizing the protein conformation (Fig. 4a). These preliminary findings merit further detailed study.

Although the metal coordinators of imide-hydrolyzing enzymes in this case study were dispersed sequentially, almost all the known metal coordinators in 1GKQ were identified by FAVAT except His 183 (His 192 for rat imidase). This residue is conserved in CRMP1 but lacks metal and amidohydrolytic activity. The role of this histidine needs further study. The major difference between bacterial hydantoinase and mammalian imidase is their metal content. The former contains two metal ions while the later contains only one metal ion [39,40]. Fewer metal coordinators may be needed for mammalian imidase, and residue His 192 may not be required as a coordinator of metal ions in rat imidase. A mammalian imidase was crystallized recently [41]. The difference between mammalian imidase and non-mammalian imidase is expected to be clarified in the near future.

Conclusions

The FAVAT was developed to predict functionally important amino acids in mammalian imidase. A T-score was given to each residue of the target enzyme by analyzing imidase-related proteins in the amidohydrolase superfamily on the basis of their sequence-function relationships. Of the ten top T-score amino acids selected, six (His67, His69, Lys159, His192, His248 and Asp326) corresponded to metal coordination in D-hydantoinase. The other four amino acids corresponded with positions that were structurally important for forming quaternary structures and secondary structures in 1GKQ. Residue

Asp57, which was misrecognized as a metal coordinator in previous MSA analyses, was correctly recognized by FAVAT. This study showed that analyzing the discrepancy between the sequence and functional properties of related proteins in a superfamily is an effective method of identifying functionally important residues. This strategy should be applicable to other protein families, and the authors expect to employ this strategy for analyzing critical residues of viruses in future works.

Methods

Imidase and sequence clustering in the amidohydrolase superfamily

Hydantoinase activity was first reported in plants and animals [42,43] to hydrolyze hydantoin derivatives that are not known as physiological metabolites. This enzymatic activity is useful for preparing optically pure amino acids that are precursors for various antibiotics [44]. Due to its industrial application, several hydantoinases have been studied and purified from microorganisms [45,46]. A dihydropyrimidinase (5, 6-dihydropyrimidine amidohydrolase) partially purified from animal livers was shown to hydrolyze the physiological substrate dihydropyrimidine [47]. A detailed study of a homogenous imide-hydrolyzing enzyme, imidase, which was purified from rat, pig or fish livers [48-51], revealed that it catalyzes a wide spectrum of substrates, including dihydropyrimidines, hydantoins and other imides [52]. Despite the substrate spectra of hydantoinase highly similar to imidase, these imide-hydrolyzing enzymes from bacterial and mammalian sources reportedly have relatively low sequence similarity. Some mammals, flies and *C. elegans*, reveal proteins with high sequence similarity to dihydropyrimidinase (or imidase). These dihydropyrimidinase-related proteins (DRPs) may be involved in cancer and neuron cells development, but possess no imidase activity [53-55]. Additionally, other enzymes revealed by the studies in evolution of the metabolic pathway are also known to use mechanisms similar to those observed in imidase [56,57]. These enzymes include dihydroorotase, allantoinase, urease and amidohydrolases, which originate in mammals, plants and fungi [58]. All use distinct substrates that contain similar imide functional groups.

All of the above enzymes can be classified into the amidohydrolase superfamily according to their properties and structures [59]. In this superfamily, some proteins have similar sequences but divergent functions whereas others have similar functions but low sequence similarity. This phenomenon strongly suggests that only a few critical amino acid residues in this superfamily are needed for specific protein functions. Proteins in the amidohydrolase superfamily can be grouped according to their sequence similarity and biochemical properties,

and an effective strategy for analyzing these proteins may yield valuable information.

A string search containing the rat imidase sequence (Accession No.: NP_113893) yielded 156 protein sequences of amidohydrolase superfamily were obtained from the PIR database [60]. According to their sequence similarity and functional properties, sequences were further clustered into the following five groups (Table 2): I. imidase (imide-hydrolyzing enzyme from mammal); II. sequence-related proteins (dihydropyrimidinase-related proteins with 50% or higher sequence similarity to mammalian imidase but without imidase activity); III. functionally identical enzymes (hydantoinase, or the imide-hydrolyzing enzyme from bacteria with 30-40% sequence similarity to mammalian imidase); IV. functionally-related enzymes (dihydroorotase, allantoinase, urease and amidohydrolase with 25-48% sequence similarity to mammalian imidase); and V. putative sequences (gene products with unknown function) with 30% or higher sequence similarity to mammalian imidase.

Observation and assumption

Table 2 shows the significant findings of the comparisons of sequence identity and functional properties of the imidase-related proteins in the amidohydrolase superfamily. Sequence-related proteins (Group II) had no imidase activity even though the overall similarity of sequences in this group was higher than 50% [61]. This phenomenon implicated that key amino acid residues for imidase activity have been altered or removed from Group II proteins. In contrast, functionally identical and functionally related enzymes (Group III and Group IV, respectively) had lower sequence identity, but they basically catalyzed the same reaction. For these enzymes, few conserved residues should be needed to provide a similar imidase function. The above observation implies a principle for classifying the sequence or the functional divergent enzymes in a superfamily, which may help to develop a feature amplified voting

algorithm for identifying key residues in a target protein. Biologists generally select a sequence of interest as a target and perform BLAST analysis to recover related protein sequences. By definition, a specific function or property (e.g., substrate specificity) of a sequence of interest can be used to cluster the related sequences into different groups. For example, if property A is used to classify these sequences, sequences with property A should be classified into group A. Otherwise, the sequence should be classified into group $\sim A$. Based on the classification of these three groups, a critical assumption can be made. When the target sequence, the A sequence and the $\sim A$ sequence belong to a protein superfamily, the conserved residues of the target sequence and the A sequence should correlate to property A. However, the non-conserved residues of the target sequence and $\sim A$ sequence should also reflect the negative correlation in property A. The intersection of the residues provides important functional clues. Thus, important residues can be found by comparing these three property groups. Therefore, the three-profile alignment method was developed to optimize the alignment of three groups so that key residues are then revealed by voting algorithm. Figure 5 shows the FAVAT flowchart that was developed to retrieve useful information for these distinct groups. Analysis of the imidase-related sequences in Table 2 indicated that, within this superfamily, the degree of sequence similarity did not necessarily reflect the similarity in biochemical properties. This provides a good opportunity to develop a novel method for extracting functional residues of a target enzyme. In this work, FAVAT was used to examine each residue in mammalian imidase by comparing other sequences in the amidohydrolase superfamily. Given the limited structural information about mammalian imidase, the analytical results of this study should provide important clues for enzymologists to perform further in-depth biochemical analyses of these results.

Table 2 Grouping of imidase related proteins¹

Group	Member number	Sequence identity	Imidase Activity
I. Imidase (target enzyme)	5	98-100%	Yes
II. Sequence related proteins ²	43	50-80%	No
III. Functionally identical enzymes ³	16	30-40%	Yes
IV. Functionally related enzymes ⁴	63	25-48%	Yes
V. Putative proteins ⁵	29	30-60%	Unknown

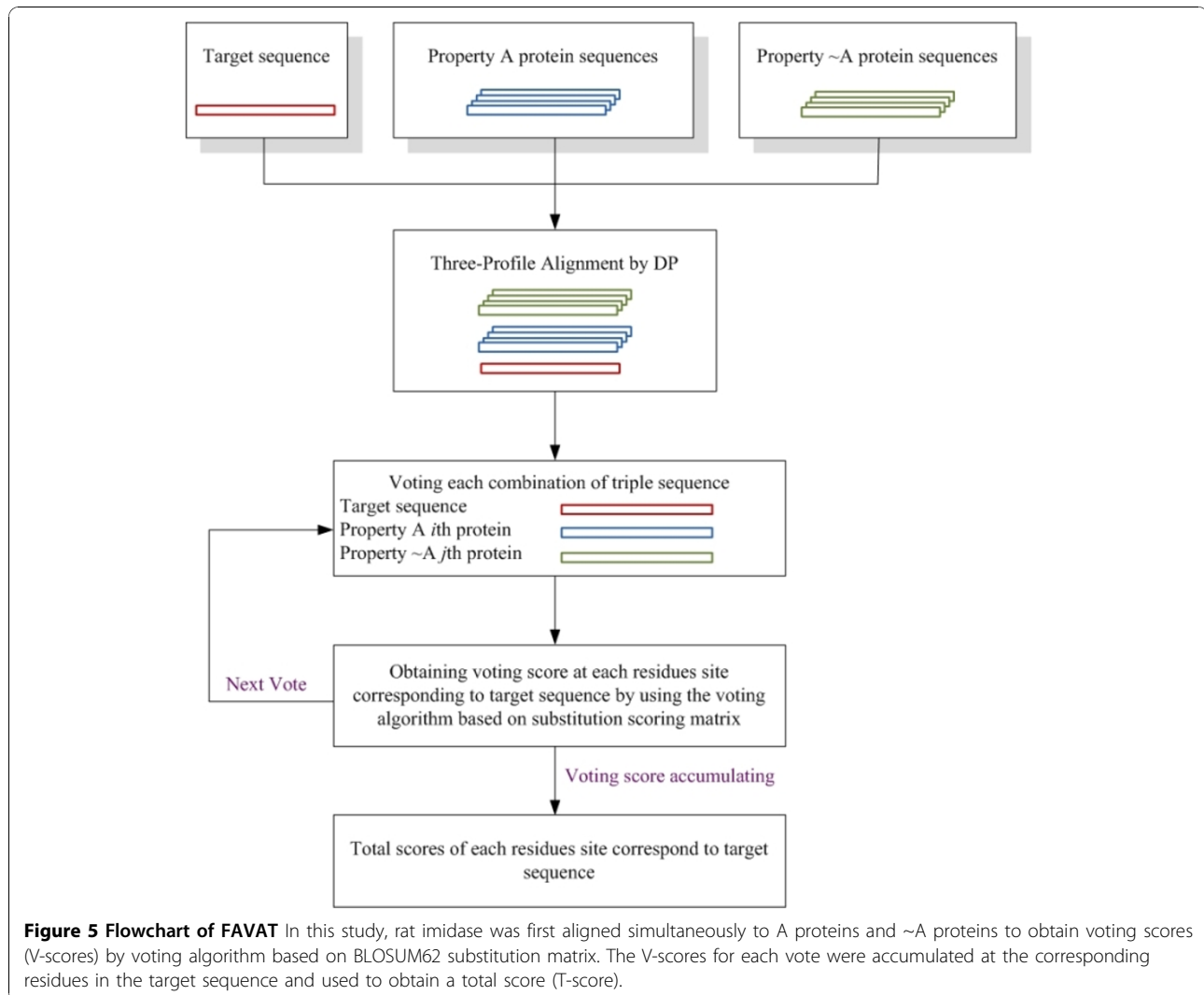
¹Proteins are grouped according to properties related to the target enzyme (rat imidase).

²Sequence-related proteins, referred to as dihydropyrimidinase related proteins (DRPs), with no imidase activity.

³Functionally identical enzymes, referred to as bacterial imide-hydrolyzing enzyme or hydantoinase, in which substrate spectrums were virtually identical to that of the target enzyme.

⁴Functionally-related enzymes, including dihydroorotase, allantoinase, and urease. Each enzyme catalyzes its own distinct substrate that contains an imide functional group.

⁵ Putative proteins are referred to as gene products with unknown function.



Algorithm

The FAVAT was performed in two steps. The first step was to align the target enzyme, functionally identical enzymes (A proteins) and sequence-related proteins (~A proteins) using three-profile alignment. The three-profile alignment algorithm, which is based on the dynamic programming three-way alignment approach [62,63], was designed to align three profiles in a space. As in the FAVAT pre-process, each profile can be generated by multiple sequence alignment tools such as T-COFFEE, HMMER [64] and MUSCLE.

Let P_1 , P_2 and P_3 be three profiles, and P_{1i} , P_{2j} and P_{3k} refer to the i th, j th and k th positions in P_1 , P_2 and P_3 , respectively, starting from 1. The symbol “-” denotes a “gap” in the alignment. Scores for the alignment of two columns are denoted by $Sp(\alpha, \beta)$. The scoring pair profiles P_1 - P_2 are defined as follows:

$$Sp_{12}(i, j) = \frac{\sum_{a=1}^m \sum_{b=1}^n (W_a \times W_b \times M[r_{1(a,i)}][r_{2(b,j)}])}{m \times n},$$

where Sp_{12} is the score at the i th and j th columns on P_1 and P_2 , respectively. The P_1 has m sequences, and P_2 has n sequences. The W_a and W_b are the sequence weights for sequence a in P_1 and sequence b in P_2 , respectively. The residue at i th column for sequence a in P_1 is denoted by $r_{1(a,i)}$. The M is the value of the substitute matrix for $r_{1(a,i)}$ and $r_{2(b,j)}$. Many substitution matrices, such as BLOSUM, have been proposed to improve alignment accuracy [65]. Similarly, the definitions of scoring pair profiles P_1 - P_3 and P_2 - P_3 are similar to those of pair profile P_1 - P_2 . Gap penalties are determined by gap opening (*GOP*) and gap extension (*GEP*).

$$S(i, j, k) = \begin{cases} 0 & \text{if } i = 0, j = 0 \text{ and } k = 0 \\ - \left(GOP_2(j, k) + \sum_{i=1}^{i-i} GEP_2(j, k) \right) & \text{if } i > 0, j = 0 \text{ and } k = 0 \\ - \left(GOP_2(i, k) + \sum_{i=1}^{i-j} GEP_2(i, k) \right) & \text{if } i = 0, j > 0 \text{ and } k = 0 \\ - \left(GOP_2(i, j) + \sum_{i=1}^{i-k} GEP_2(i, j) \right) & \text{if } i = 0, j = 0 \text{ and } k > 0 \\ \max [G(i, j, k), H(i, j, k), I(i, j, k)] & \text{if } i > 0, j > 0 \text{ and } k = 0 \\ \max [E(i, j, k), H(i, j, k), J(i, j, k)] & \text{if } i > 0, j = 0 \text{ and } k > 0 \\ \max [F(i, j, k), I(i, j, k), J(i, j, k)] & \text{if } i = 0, j > 0 \text{ and } k > 0 \\ \max \begin{bmatrix} E(i, j, k), F(i, j, k), G(i, j, k), \\ H(i, j, k), I(i, j, k), J(i, j, k), \\ S(i-1, j-1, k-1) \\ + Sp(i, j) + Sp(i, k) + Sp(j, k) \end{bmatrix} & \text{if } i > 0, j > 0 \text{ and } k > 0 \end{cases}$$

$$G(i, j, k) = \max \left\{ \begin{matrix} S_{(i-1, j-1, k)} - GOP_1(k) \\ G_{(i-1, j-1, k)} \end{matrix} \right\} + Sp(i, j) - GEP_1(k) \quad \text{if } i > 0, j > 0$$

$$E(i, j, k) = \max \left\{ \begin{matrix} S_{(i-1, j, k-1)} - GOP_1(j) \\ E_{(i-1, j, k-1)} \end{matrix} \right\} + Sp(i, k) - GEP_1(j) \quad \text{if } i > 0, k > 0$$

$$F(i, j, k) = \max \left\{ \begin{matrix} S_{(i, j-1, k-1)} - GOP_1(i) \\ F_{(i, j-1, k-1)} \end{matrix} \right\} + Sp(j, k) - GEP_1(i) \quad \text{if } j > 0, k > 0$$

$$H(i, j, k) = \max \left\{ \begin{matrix} S_{(i-1, j, k)} - GOP_2(j, k) \\ H_{(i-1, j, k)} \end{matrix} \right\} - GEP_2(j, k) \quad \text{if } i > 0$$

$$I(i, j, k) = \max \left\{ \begin{matrix} S_{(i, j-1, k)} - GOP_2(i, k) \\ H_{(i, j-1, k)} \end{matrix} \right\} - GEP_2(i, k) \quad \text{if } j > 0$$

$$J(i, j, k) = \max \left\{ \begin{matrix} S_{(i, j, k-1)} - GOP_2(i, j) \\ H_{(i, j, k-1)} \end{matrix} \right\} - GEP_2(i, j) \quad \text{if } k > 0$$

Figure 6 Dynamic programming recursions for three-profile alignment with affine gap costs Scores for the alignment of two residue positions are denoted by $Sp(\alpha, \beta)$. The GOP and GEP are gap opening and gap extension penalties, respectively. In affine gap costs, GOP and GEP are fixed values at any residue position.

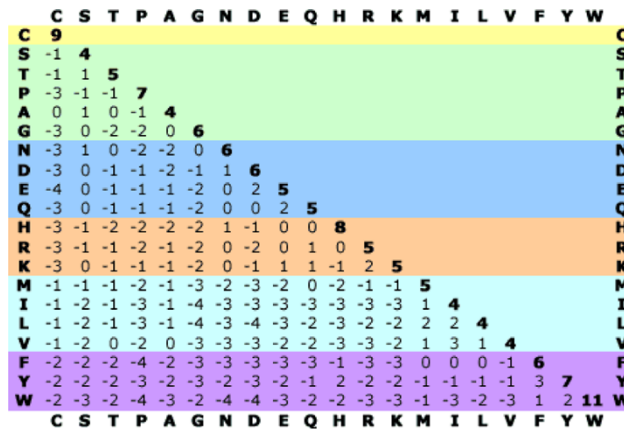
scores. The best score of the alignments with prefixes P_{1i} , P_{2j} and P_{3k} is denoted by $S(i, j, k)$ if the residues (P_{1i}, P_{2j}, P_{3k}) are aligned; $G(i, j, k)$ is the best score given that $(P_{1i}, P_{2j}, -)$ is the last column of the partial alignment, and $H(i, j, k)$ is the best score given that the last column is of the form $(P_{1i}, -, -)$. $E(i, j, k)$, $F(i, j, k)$, $I(i, j, k)$ and $J(i, j, k)$ are defined analogously. These quantities clearly satisfy the recursions summarized in Fig. 6.

The next step after the alignment is to determine whether amino acid residues critical for imidase activity exist in target and A proteins but are absent in ~A proteins. In the second step, a voting score (V-score) is

given based on the previous assumption, and the V-scores are then summed in each comparison. In this step, a substitution matrix (BLOSUM62) is used to give V-score when each sequence of property A and ~A is compared to the target sequence. The V-score is calculated as follows:

$$V_{k(a,b)} = M[t_k, A_{(a,k)}] - M[t_k, \sim A_{(b,k)}],$$

where $V_{k(a,b)}$ is the V-score at the k th residues on target sequence, sequence a in A proteins and sequence b in ~A proteins. The A and ~A proteins have m and n sequences, respectively. The t_k is the k th residue of the



(a)

Target sequence	A	L	E	D	N
Functionally Identical or Related Enzymes (property A proteins)	H	L	E	P	L
Sequence-related Proteins (~A proteins)	L	L	P	D	L
Score	-1	0	6	-7	0

(b)

Figure 7 Score setting and scoring example (a) Substitution matrix for BLOSUM62. In this study, the BLOSUM 62 substitution matrix was used to obtain V-scores for the voting algorithm. (b) Example of FAVAT scoring. When three-profile alignment was performed in the target sequence, the voting scores for A proteins and ~A proteins corresponded with the BLOSUM62 substitution matrix. For example, residues for target sequence, an A protein and a ~A protein were A, H and L, respectively. The V-score is -2 (A to H in BLOSUM62) - (-1) (A to L in BLOSUM62) = -1.

target sequence. The $A(a,k)$ and $\sim A(b,k)$ are the k th residues on sequence a in A proteins and on sequence b in ~A proteins, respectively. The $M[t_k, A(a,k)]$ is the value of the substitution matrix for t_k and $A(a,k)$. Figure 7 shows the V-score calculation. The BLOSUM62 substitution matrix is generally used for protein or nucleic acid sequence alignment. In the proposed algorithm, each V-score is given and accumulated to a total score

(T-Score) until all sequences of property A and ~A are compared. The T-Score is calculated as follows:

$$T_k = \sum_{a=1}^m \sum_{b=1}^n |V_{k(a,b)}|$$

where T_k is the T-Score at the k th residue on target sequence. The T-Score at each residue position is

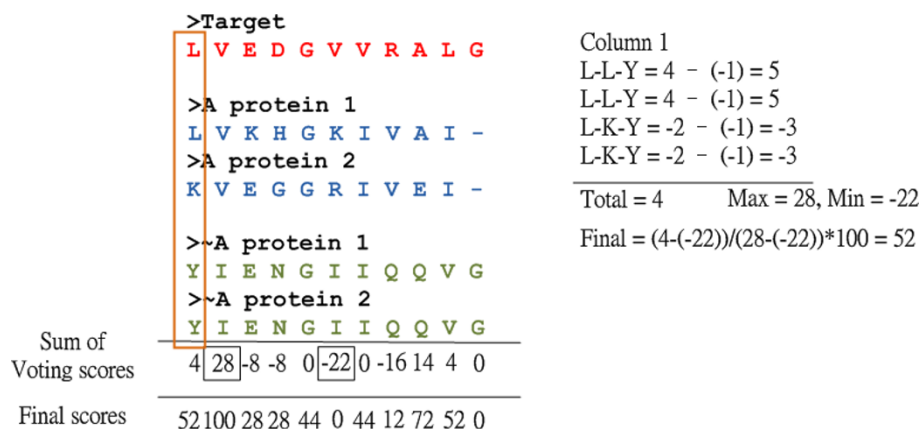


Figure 8 FAVAT scoring example For the first position of the target sequence, four V-scores were calculated for L-L-Y, L-L-Y, L-K-Y and L-K-Y, respectively. The T-score at each position is the sum of four V-scores, and the T-score (total score) for the first position of the target sequence is 4. After the normalization function, the final T-score is 52. When any gap existed at the position in one of three groups such as column 11, the T-score for this position is zero.

obtained by adding $m \times n$ V-scores. The normalization function is used to transform all T-scores into a range from 0 to 100 in FAVAT. The normalization function is as follows:

$$T_k = \frac{T_k - \text{Min}(T)}{\text{Max}(T) - \text{Min}(T)} \times 100,$$

where $\text{Max}(T)$ and $\text{Min}(T)$ are the maximum and minimum scores of all T-scores, respectively. The details of the FAVAT algorithm are shown below, and Fig. 8 presents an example of V-Score and T-Score calculations using FAVAT.

Algorithm FAVAT (t, P, Q);

Input: Target sequence t , a set of proteins P without property A , a set of proteins Q with property A . P has p sequences and Q has q sequences.

Output: The scores correspond to the residues of t (high T-scores indicate potentially critical residues)

Begin

1 /***Step1:** Do three-profile alignment by dynamic programming method among t , the proteins P , and the proteins Q . The length of the resulting alignments is m_{max} . $t[k]$ indicates the k -th residue of t .*/

2 **fork** <- 1 **to** m_{max} **do**

3 **ift**[k] <> '-' **then**

4 /* T -score [k] indicates the potential importance of the k -th residue of t .*/

5 T -score[k] <- 0

6 max <- $-\infty$

7 min <- ∞

8 **fori** <- 1 **topdo**

9 **forj** <- 1 **toqdo**

10 /* X and Y are used to store the k -th residue of this i -th protein in P and this j -th protein in Q , respectively.*/

11 (X, Y) = ($P[i], Q[j]$), k -th

12 /***Step 2:** Find V -score[k] based on the BLOSUM62 substitution matrix.*/

13 V -score[k] <- BLOSUM62($t[k], Y$)

14 V -score[k] <- V -score[k] + (-1) \times BLOSUM62($t[k], X$)

15 T -score[k] <- T -score[k] + V -score[k]

16 max <- $\text{MAX}(max, T\text{-score}[k])$

17 min <- $\text{MIN}(min, T\text{-score}[k])$

18 **end if**

19 **fork** <- 1 **to** m_{max} **do**

20 T -score[k] = (T -score[k] - min/max - min) \times 100
/*normalization*/

End

The novel feature of the FAVAT algorithm is its use of the sequence and functional properties among target sequence, $\sim A$ proteins and A proteins. When voting for reliable critical residue candidates, three relations are considered: the relation between target sequence and A

proteins, the relation between target sequence and $\sim A$ proteins and the relation between A proteins and $\sim A$ proteins. To accurately identify the key residues, some useful alignment tools with physicochemical properties, such as T-COFFEE and HMMER, can be employed in the FAVAT pre-process to align A and $\sim A$ proteins separately (profiles). The appropriate alignments of A and $\sim A$ proteins can enhance the accuracy of the resulting alignment to the target sequence, A and $\sim A$ proteins by three-profile alignment. The most important residues can then be found accurately from the resulting alignment using FAVAT. The FAVAT algorithm was designed to account for the importance of alignment-based voting skill by V-score function. The time complexity of FAVAT is $O(m_{max}^3)$, and m_{max} is the length of the resulting alignment by three-profile alignment. To reduce the time complexity of three-profile alignment method, this study developed a parallel version implemented by the MPICH library. The time complexity for the parallel version is $O(m_{max}^3/p)$, where p is the number of processors. After the voting process, the residue candidates obtain high T-scores. The uncritical candidates can be eliminated by advanced research.

Acknowledgements

We would like to thank the anonymous referees for many constructive comments during the revision. We also would like to thank Ted Knoy for editorial assistance. Part of this work was supported by the National Science Council (NSC) under grant NSC98-2218-E-007-005 and NSC97-2221 -E-182-03 3 -MY3. Publication of this supplement was made possible with support from the International Society of Intelligent Biological Medicine (ISIBM). This article has been published as part of *BMC Genomics* Volume 11 Supplement 3, 2010: The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2164/11?issue=53>.

Author details

¹Department of Computer Science, National Tsing Hua University, 101, Section 2 Kuang Fu Road, Hsinchu, Taiwan 300. ²Macronix International Co., Ltd., 16, Li-Hsin Road, Science Park, Hsinchu, Taiwan 300. ³Department of Computer Science and Information Engineering, Chang Gung University, 259 Wen Hwa 1st Road, Kwei Shan, Taoyuan 333, Taiwan.

Authors' contributions

CLH, CL and CYT conceived the research. CLH and CL implemented the program and performed the experiments. CLH, CL and CYL arranged the test data and analyzed the experimental results. CLH, CL, CYL and YCC wrote the article. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 1 December 2010

References

1. The FAVAT software:[<http://sslabs.cs.nthu.edu.tw/~allen/favat/index.html>].
2. Babbitt PC, Gerlt JA: **Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities.** *J Biol Chem* 1997, **272**:30591-30594.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
4. Altschul SF, Boguski MS, Gish W, Wootton JC: **Issues in searching molecular sequence databases.** *Nat Genet* 1994, **6**:119-129.

5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
6. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85**:2444-2448.
7. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
8. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**:205-217.
9. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792-1797.
10. Livingstone CD, Barton GJ: **Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation.** *Comput Appl Biosci* 1993, **9**:745-756.
11. Feng DF, Doolittle RF: **Progressive sequence alignment as a prerequisite to correct phylogenetic trees.** *J Mol Evol* 1987, **25**:351-360.
12. Barton GJ, Sternberg MJ: **A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons.** *J Mol Biol* 1987, **198**:327-337.
13. Grigoriev IV, Kim SH: **Detection of protein fold similarity based on correlation of amino acid properties.** *Proc Natl Acad Sci USA* 1999, **96**:14318-14323.
14. Henikoff S, Henikoff JG: **Automated assembly of protein blocks for database searching.** *Nucleic Acids Res* 1991, **19**:6565-6572.
15. Sonnhammer EL, Kahn D: **Modular arrangement of proteins as inferred from analysis of homology.** *Protein Sci* 1994, **3**:482-492.
16. Ogiwara A, Uchiyama I, Takagi T, Kanehisa M: **Construction and analysis of a profile library characterizing groups of structurally known proteins.** *Protein Sci* 1996, **5**:1991-1999.
17. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ: **ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins.** *Nucleic Acids Res* 2003, **31**:3625-3630.
18. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257**:342-358.
19. Lichtarge O, Sowa ME, Philippi A: **Evolutionary traces of functional surfaces along G protein signaling pathway.** *Methods Enzymol* 2002, **344**:536-556.
20. Lichtarge O, Yamamoto KR, Cohen FE: **Identification of functional surfaces of the zinc binding domains of intracellular receptors.** *J Mol Biol* 1997, **274**:325-337.
21. Innis CA, Shi J, Blundell TL: **Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis.** *Protein Eng* 2000, **13**:839-847.
22. Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O: **Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions.** *J Biol Chem* 2004, **279**:8126-8132.
23. King RD, Karwath A, Clare A, Dehaspe L: **The utility of different representations of protein sequence for predicting functional class.** *Bioinformatics* 2001, **17**:445-454.
24. Brudno M, Poliakov A, Salamov A, Cooper GM, Sidow A, Rubin EM, Solovyev V, Batzoglu S, Dubchak I: **Automated whole-genome multiple alignment of rat, mouse, and human.** *Genome Res* 2004, **14**:685-692.
25. Kolbe D, Taylor J, Elnitski L, Eswara P, Li J, Miller W, Hardison R, Chiaromonte F: **Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat.** *Genome Res* 2004, **14**:700-707.
26. Margulies EH, Chen CW, Green ED: **Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons.** *Trends Genet* 2006, **22**:187-193.
27. Reyes A, Gissi C, Catzeflis F, Nevo E, Pesole G, Saccone C: **Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods.** *Mol Biol Evol* 2004, **21**:397-403.
28. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ: **Molecular phylogenetics and the origins of placental mammals.** *Nature* 2001, **409**:614-618.
29. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS: **Resolution of the early placental mammal radiation using Bayesian phylogenetics.** *Science* 2001, **294**:2348-2351.
30. Thoden JB, Phillips GN Jr, Neal TM, Raushel FM, Holden HM: **Molecular structure of dihydroorotase: a paradigm for catalysis through the use of a binuclear metal center.** *Biochemistry* 2001, **40**:6989-6997.
31. Kim GJ, Kim HS: **C-terminal regions of D-hydantoinases are nonessential for catalysis, but affect the oligomeric structure.** *Biochem Biophys Res Commun* 1998, **243**:96-100.
32. Williams NK, Manthey MK, Hambley TW, O'Donoghue SI, Keegan M, Chapman BE, Christopher son RI: **Catalysis by hamster dihydroorotase: zinc binding, site-directed mutagenesis, and interaction with inhibitors.** *Biochemistry* 1995, **34**:11344-11352.
33. Abendroth J, Niefind K, Schomburg D: **X-ray structure of a dihydropyrimidinase from *Thermus* sp. at 1.3 Å resolution.** *J Mol Biol* 2002, **320**:143-156.
34. Deo RC, Schmidt EF, Elhabazi A, Togashi H, Burley SK, Strittmatter SM: **Structural bases for CRMP function in plexin-dependent semaphorin3A signaling.** *EMBO J* 2004, **23**:9-22.
35. Abendroth J, Niefind K, Schomburg D: **X-ray structure of a dihydropyrimidinase from *Thermus* sp. at 1.3 Å resolution.** *J Mol Biol* 2002, **320**:143-156.
36. Copley RR, Bork P: **Homology among (beta/alpha)(8) barrels: implications for the evolution of metabolic pathways.** *J Mol Biol* 2000, **303**:627-641.
37. Gerlt JA, Raushel FM: **Evolution of function in (beta/alpha)(8)-barrel enzymes.** *Curr Opin Chem Biol* 2003, **7**:252-264.
38. Deo RC, Schmidt EF, Elhabazi A, Togashi H, Burley SK, Strittmatter SM: **Structural bases for CRMP function in plexin-dependent semaphorin3A signaling.** *EMBO J* 2004, **23**:9-22.
39. Yang YS, Ramaswamy S, Jakoby WB: **Rat liver imidase.** *J Biol Chem* 1993, **268**:10870-10875.
40. Sylđatk C, May O, Altenbuchner J, Mattes R, Siemann M: **Microbial hydantoinases—industrial enzymes from the origin of life?** *Appl Microbiol Biotechnol* 1999, **51**:293-309.
41. Huang CY, Chiang SK, Yang YS, Sun YJ: **Crystallization and preliminary X-ray diffraction analysis of thermophilic imidase from pig liver.** *Acta Crystallogr D Biol Crystallogr* 2003, **59**:943-945.
42. Eadie GS, Bernheim F, Bernheim ML: **The partial purification and properties of animal and plant hydantoinases.** *J Biol Chem* 1949, **181**:449-458.
43. Bernheim F, Bernheim MLC: **The hydrolysis of hydantoin by various tissues.** *J Biol Chem* 1946, **163**:683-685.
44. Altenbuchner J, Siemann-Herzberg M, Sylđatk C: **Hydantoinases and related enzymes as biocatalysts for the synthesis of unnatural chiral amino acids.** *Curr Opin Biotechnol* 2001, **12**:559-563.
45. Altenbuchner J, Siemann-Herzberg M, Sylđatk C: **Hydantoinases and related enzymes as biocatalysts for the synthesis of unnatural chiral amino acids.** *Curr Opin Biotechnol* 2001, **12**:559-563.
46. Sonnhammer EL, Kahn D: **Modular arrangement of proteins as inferred from analysis of homology.** *Protein Sci* 1994, **3**:482-492.
47. Wallach DP, Grisolia S: **The purification and properties of hydroxyproline hydrazinase.** *J Biol Chem* 1957, **226**:277-288.
48. Huang CY, Yang YS: **Discovery of a novel N-iminylamidase activity: substrate specificity, chemoselectivity and catalytic mechanism.** *Protein Expr Purif* 2005, **40**:203-211.
49. Yang YS, Ramaswamy S, Jakoby WB: **Rat liver imidase.** *J Biol Chem* 1993, **268**:10870-10875.
50. Su TM, Yang YS: **Identification, purification, and characterization of a thermophilic imidase from pig liver.** *Protein Expr Purif* 2000, **19**:289-297.
51. Huang CY, Yang YS: **The role of metal on imide hydrolysis: metal content and pH profiles of metal ion-replaced mammalian imidase.** *Biochem Biophys Res Commun* 2002, **297**:1027-1032.
52. Sylđatk C, May O, Altenbuchner J, Mattes R, Siemann M: **Microbial hydantoinases— industrial enzymes from the origin of life?** *Appl Microbiol Biotechnol* 1999, **51**:293-309.

53. Hamajima N, Matsuda K, Sakata S, Tamaki N, Sasaki M, Nonaka M: **A novel gene family defined by human dihydropyrimidinase and three related proteins with differential tissue distribution.** *Gene* 1996, **180**:157-163.
54. Goshima Y, Nakamura F, Strittmatter P, Strittmatter SM: **Collapsin-induced growth cone collapse mediated by an intracellular protein related to UNC-33.** *Nature* 1995, **376**:509-514.
55. Wang LH, Strittmatter SM: **Brain CRMP forms heterotetramers similar to liver dihydropyrimidinase.** *J Neurochem* 1997, **69**:2261-2269.
56. Holm L, Sander C: **An evolutionary treasure: unification of a broad set of amidohydrolases related to urease.** *Proteins* 1997, **28**:72-82.
57. Copley RR, Bork P: **Homology among (beta/alpha)₈ barrels: implications for the evolution of metabolic pathways.** *J Mol Biol* 2000, **303**:627-641.
58. Gerlt JA, Raushel FM: **Evolution of function in (beta/alpha)₈-barrel enzymes.** *Curr Opin Chem Biol* 2003, **7**:252-264.
59. Holm L, Sander C: **An evolutionary treasure: unification of a broad set of amidohydrolases related to urease.** *Proteins* 1997, **28**:72-82.
60. Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KC, Mewes HW, Orcutt BC, Suzek BE, Tsugita A, Vinayaka CR, Yeh LS, Zhang J, Barker WC: **The Protein Information Resource: an integrated public resource of functional annotation of proteins.** *Nucleic Acids Res* 2002, **30**:35-37.
61. Hamajima N, Matsuda K, Sakata S, Tamaki N, Sasaki M, Nonaka M: **A novel gene family defined by human dihydropyrimidinase and three related proteins with differential tissue distribution.** *Gene* 1996, **180**:157-163.
62. Sankoff D: **Simultaneous solution of the RNA folding, alignment, and protosequence problems.** *SIAM Journal on Applied Mathematics* 1985, **45**:810-825.
63. Huang X: **Alignment of three sequences in quadratic space.** *ACM SIGAPP Applied Computing Review* 1993, **1**:7-12.
64. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
65. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89**:10915-10919.
66. Kim GJ, Kim HS: **C-terminal regions of D-hydantoinases are nonessential for catalysis, but affect the oligomeric structure.** *Biochem Biophys Res Commun* 1998, **243**:96-100.

doi:10.1186/1471-2164-11-S3-S14

Cite this article as: Hung *et al.*: Feature amplified voting algorithm for functional analysis of protein superfamily. *BMC Genomics* 2010 **11**(Suppl 3):S14.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

