Computer Methods and Programs in Biomedicine (2005) **78**, 209—222

Computer Methods and
Programs in Biomedicine

ELSEVIER

www.intl.elsevierhealth.com/journals/cmpb

Prediction of orthologous relationship by functionally important sites

Hsuan-Chao Chiu[a], **C. Allen Chang**[b], **Yuh-Jyh Hu**[a,*]

[a] *Department of Computer and Information Science, National Chiao Tung University,*
1001 Ta Shueh Rd., Hsinchu, Taiwan
[b] *Department of Biological Science and Technology, National Chiao Tung University,*
1001 Ta Shueh Rd., Hsinchu, Taiwan

Received 19 May 2004; received in revised form 1 March 2005; accepted 3 March 2005

KEYWORDS
Orthologs;
Paralogs;
Protein functions;
Important sites

Summary Making accurate functional predictions plays an important role in the era of proteomics. Reliable functional information can be extracted from orthologs in other species when annotating an unknown gene. Here a site-based approach called PORFIS is proposed to predict orthologous relationship. When applied to the bacterial transcription factor PurR/LacI family and the protein kinase AGC family, our method was able to identify, with few false positives, the important sites that agree with those verified by biological experiments. We also tested it on the α-proteasome family, the glycoprotein hormone family and the growth hormone family to demonstrate its ability to predict orthologous relationship. Compared with other prediction methods based on phylogenetic analysis or hidden Markov models, PORFIS not only has competitive prediction accuracy, but also provides valuable biological information of functionally important sites associated with orthologs which can be further studied in biological experiments.

1. Introduction

Advanced sequencing technology and various genome projects have produced an enormous amount of data waiting to be annotated. A common strategy for annotation is first to search the sequence databases for the best-fit homolog, and then to assign this homolog's functions to the novel gene (or protein) of interest. In order to find appropriate homologs, different sequence alignment methods, such as BLAST [1] and hidden Markov models [2], have been developed and widely used. Despite that significant homology usually reflects significant similarity in biological functions, homologs can be further divided into orthologs and paralogs. Orthologous sequences diverged because of speciation. They are under similar regulation and have identical biochemical functions. Unlike orthologs, paralogs arose from duplication events. They do not have the same biological functions. Therefore, it is not guaranteed to annotate sequences correctly simply according

[*] Corresponding author. Tel.: +886 3 573 1795.
E-mail address: yhu@cis.nctu.edu.tw (Y.-J. Hu).

0169-2607/$ — see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2005.03.002

to their homologous relations. Incorrect prediction may result in the wrong judgment of cellular functions or the erroneous reconstruction of metabolic pathways. As the advent of functional genomics, how to distinguish between orthologs and paralogs has drawn tremendous attention recently [3—5].

clusters of orthologous groups (COGs) are the first database that stores orthologous proteins in bacteria and archaea on a genomic scale [6]. Besides COGs, INPARANOID also shows the orthologous relationship among eukaryotes, including *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, and *Drosophila melanogaster* [7]. Both systems were constructed by applying pairwise BLAST search for homologs. However, since BLAST is an algorithm based on heuristics rather than phylogeny, the best hit found by BLAST may not be a true ortholog but instead a paralog that shares only partial evolutionary features [8,9].

To mitigate the limitations of standard sequence alignment approaches, several methods have been developed to enhance the detection of orthologous sequences. For example, Cotter et al. [10] used closely related sequences as a sequence outgroup to refine the BLAST search results. Yuan et al. [9] built the reconciled trees to verify the reliability of phylogenetic trees by applying statistical resampling techniques to multiple related sequences. Storm and Sonnhammer [11], on the other hand, introduced the support value for evaluating sequence orthology. Though they showed some positive results, yet several drawbacks still limit their applicability. Firstly, the performance of these methods highly depends on the correct phylogenetic trees that may not be verified in advance. Secondly, they often require substantial domain knowledge, e.g., to select a proper sequence outgroup, which is not always available. Thirdly, none of their output results can be easily justified or further analyzed by biological experiments. To avoid the limitations, we develop a novel orthology prediction method based on the functionally important sites associated with orthologs. The motivation is that active protein residues are under evolutionary pressure to maintain their functional integrity. They undergo fewer mutations than less functionally important amino acids. As a result, functionally important sites may be used to better characterize orthologous relationships.

2. Background

There is a great deal of work on functionally important sites. One possible way to extract functional information is through the comparison of evolutionarily related sequences [12,13]. The evolutionary trace method combined the knowledge of protein structures with sets of homologous sequences to infer functional interface [14]. Casari et al. [15] used a principle component analysis of a vector representation of sequences in space to identify functional residues. Hannenhalli and Russell [16] analyzed functional sub-types using relative entropy. Mirny and Gelfand [17] used orthologous and paralogous proteins to identify specificity-determining residues. Bickel et al. [18] found the important residues in phycobiliproteins and vertebrate globin sequences based on the well-conserved motifs in homologous families. In spite of their capability of finding functionally important sites, most of the methods are too computation-intensive to be applicable in the prediction of orthologous relations.

Here we propose a method called PORFIS to predict orthologous relationship based on functionally important sites that confer the specificity and conservation of different orthologous subgroups within a set of homologous sequences. The orthologous relationship of a novel protein sequence to these subgroups is inferred from the important sites found. We hypothesize that the important residues should be conserved in orthologous proteins to maintain their identical function while divergent in paralogous proteins to reflect their specificity. We explore functionally important sites in the multiple sequence alignment of orthologous and paralogous proteins and use these sites to build a model to classify orthologous relations of novel proteins. Unlike previous works, PORFIS provides substantial information for further biological experiments, e.g. site-directed mutagenesis to verify orthologous relationship. We first applied PORFIS to the bacterial PurR/LacI family and the protein kinase AGC group family to demonstrate its ability to identify functionally important sites. To further evaluate its accuracy of orthology prediction, we compared PORFIS with other current approaches on four protein families, including the AGC family, the glycoprotein hormone family, the α-proteasome family and the growth hormone family.

3. Design considerations

We refer the functionally important sites of an orthologous family to those residues which are both: (1) well conserved within orthologs and (2) divergent among paralogs. Sites/residues with both properties in a multiple sequence alignment of homologs (orthologs and paralogs) are considered important and used to construct the classification model for orthologous subfamilies.

For an alignment of homologous proteins that have been properly partitioned into orthologous subfamilies, we evaluate the degree of inter-paralog divergence and intra-ortholog conservation of each site by calculating the adjusted Rand index (ARI) [19] and the entropy.

Given a set of n objects $O = \{o_1, \ldots, o_n\}$, suppose $P = \{p_1, p_2, \ldots, p_R\}$ and $Q = \{q_1, q_2, \ldots, q_S\}$ represent two different partitions of the objects in O such that $\cup_{i=1}^{R} p = \cup_{j=1}^{S} q = O$ and $p_a \cap p_b = \emptyset$, $q_c \cap q_d = \emptyset$ for $1 \leq a \neq b \leq R$, $1 \leq c \neq d \leq S$. For each object pair $\{O_i, O_j\}$ there are four possible outcomes:

Type1: O_i and O_j are in the same partition in P and in the same partition in Q
Type2: O_i and O_j are in different partition in P but in the same partition in Q
Type3: O_i and O_j are in the same partition in P but in different partition in Q
Type4: O_i and O_j are in different partition in P and in different partition in Q

Let a, b, c, d be the number of object pairs of Type 1 to Type 4, respectively, and $n = a + b + c + d$, Rand index is defined as the fraction of agreement, i.e. $(a + d)/(a + b + c + d)$ [20]. Rand index lies between 0 and 1. When the two partitions P and Q are identical, Rand index is 1. A problem with Rand index is that its expected value may not be constant. Hubert and Arabie then proposed the adjusted Rand index to solve this problem [19]. The adjusted Rand index is defined as follows:

$$\mathrm{ARI} = \frac{\text{Rand indexed} - E(\text{Rand indexed})}{\max(\text{Rand indexed}) - E(\text{Rand indexed})} = \frac{n(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{n^2 - [(a+b)(a+c) + (c+d)(b+d)]} \quad (1)$$

Adjusted Rand index is 1 when the two partitions P and Q are identical and its expected value is 0. The higher the adjusted Rand index, the higher the agreement between the two partitions. In our case, P may refer to the given orthologous partition of the homologous proteins, and Q may refer to the partition according to the amino acids in a particular column of the protein sequence alignment. The column (i.e. the site) with higher adjusted Rand index suggests that it matches the orthologous partition better, and hence can be an important site used to distinguish paralogs.

Table 1 An example of the limitation of the adjusted Rand index (ARI)

Ortholog Classes	Sequences	Site 1	Site 2
Ortholog Class 1	Sequence 1	P	P
	Sequence 2	Q	Q
	Sequence 3	Q	Q
	Sequence 4	Q	Q
	Sequence 5	Q	Q
	Sequence 6	Q	Q
Ortholog Class 2	Sequence 7	C	C
	Sequence 8	Y	C
Ortholog Class 3	Sequence 9	L	Q
	Sequence 10	M	Q
Ortholog Class 4	Sequence 11	K	N
	Sequence 12	I	N
ARI		0.64	0.47
a (Type 1)		10	15
d (Type 4)		48	38

Compared with Site 1, Site 2 is apparently more conserved and discriminative in the ortholog classes. However, based on ARI, Site 1 will be incorrectly favored in the example.

A site may be mistaken for being important because its adjusted Rand index is relatively high due to the domination of Type 4 outcome especially when the homologous proteins are partitioned into many subfamilies. Suppose we have four ortholog classes as shown in Table 1. According to the adjusted Rand index, Site 1 is preferable to Site 2. Nevertheless, it is obvious to note that Site 2 is more conserved and distinctive in the classes. This example suggests that ARI alone does not provide sufficient information for selecting correct important sites. To alleviate the problem, we also use entropy as a complementary criterion to measure the conservation within orthologs. If the weighted average entropy of a site is high, the conservation is low.

The weighted average entropy of site k, E_k, is defined as follows:

$$E_k = \frac{1}{T}\Big[-\sum_{i=1}^{N}\sum_{w} W_i f_i(x)\log f_i(x)\Big] \quad (2)$$

$\forall 1 \leq k \leq L$ where N is the number of orthologous subfamilies that are paralogous to each other, L stands for the length of multiple sequence alignment, x is the amino acid at site k, $f_i(x)$ represents the frequency of x in orthologous subfamily i, $W_i$is the number of sequences in orthologous subfamily i, and T is the total number of homologs where $T = \sum_{i=1}^{N} W_i$.

4. System description

4.1. Evaluation of functional sites

We evaluate sites by the Z-scores of ARI and weighted average entropy. Z_{ARI} and Z_{E} are defined as the following:

$$Z_{\mathrm{ARI}_k} = \frac{\mathrm{ARI}_k - \mu_{\mathrm{ARI}}}{\sigma_{\mathrm{ARI}}} \tag{3}$$

$$Z_{\mathrm{E}_k} = \frac{\mu_{\mathrm{E}} - E_k}{\sigma_{\mathrm{E}}} \tag{4}$$

$\forall k 1 \le k \le L$ where L is the length of multiple sequence alignment. Since ARI and entropy do not necessarily follow normal distribution, normalization of ARI and entropy is required [21]. A site with high Z_{ARI} and Z_{E} is considered a functionally important site. To favor the sites that have high Z_{ARI} as well as high Z_{E}, after normalizing their values to the range between zero and one, we combine both Z_{ARI} and Z_{E} into an F-score [22] to measure the unified functional importance. The F-score is defined as follows:

$$F\text{-score}_k = \frac{2}{(1/Z_{\mathrm{ARI}_k})/(1/Z_{\mathrm{E}_k})} \tag{5}$$

$\forall k 1 \le k \le L$ where L stands for the length of multiple sequence alignment.

4.2. Prediction of orthologous relations

Given an unknown protein x and a set of homologs already divided into I orthologous subfamilies that are paralogous to each other, our goal is to assign x to the most appropriate subfamily based on the important sites found. The procedure of classification is as follows:

(1) Calculate $S_{ij}(x)$, the similarity of x to sequence j in subfamily i, for all i, j:

$$S_{ij}(x) = \frac{1}{|F|} \sum_{k \in F} \frac{M(x_k, C_{ijk})}{\max(M(x_k, x_k), M(C_{ijk}, C_{ijk}))} \tag{6}$$

where x_k represents the amino acid of x at site k, C_{ij} is the jth sequence in subfamily i of training data, C_{ijk} is the amino acid of C_{ij} at site k, M is the substitution matrix (e.g. Blosum62), F represents the set of predicted important sites, $|F|$ is the total number of predicted important sites.

(2) Calculate $S_i(x)$, the similarity of x to subfamily i, for all i:

$$S_i(x) = \frac{1}{J} \sum_{j=1}^{J} S_{ij}(x) \tag{7}$$

where J is the size of subfamily i.

(3) Assign x to subfamily C_{final}

$$x \in C_{\mathrm{final}} \text{ if } S_{\mathrm{final}}(x) \text{ is maximal} \tag{8}$$

The system flow is presented in Fig. 1.

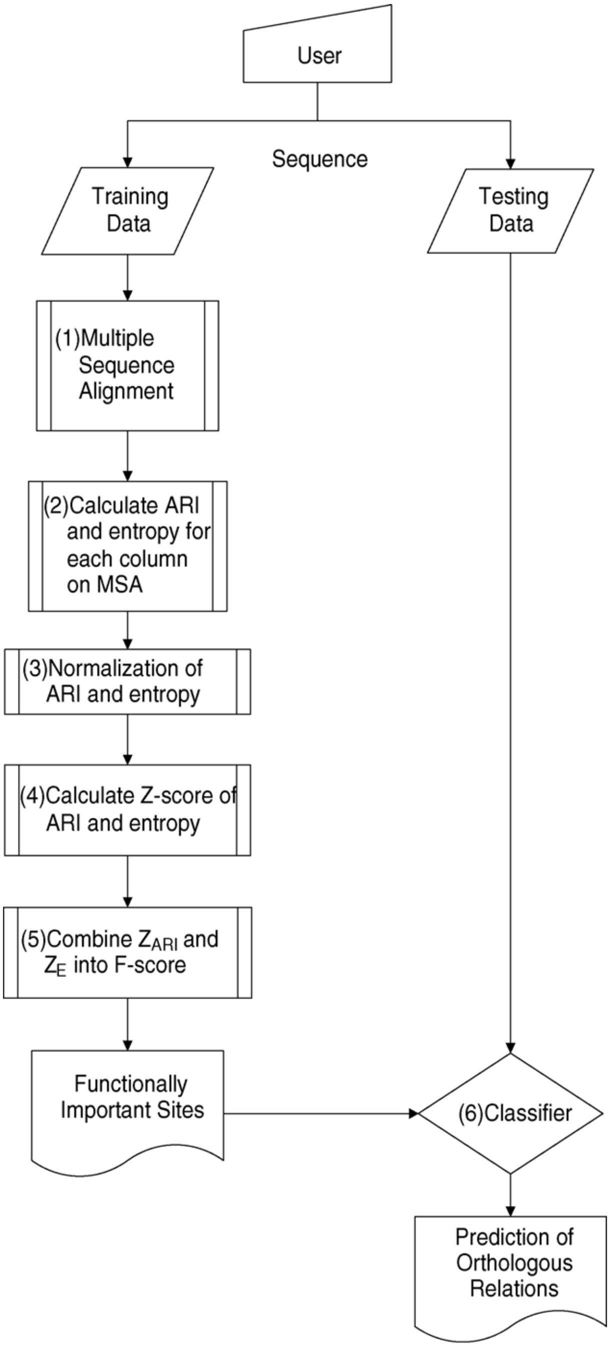
User
Sequence
Training Data
Testing Data
(1)Multiple Sequence Alignment
(2)Calculate ARI and entropy for each column on MSA
(3)Normalization of ARI and entropy
(4)Calculate Z-score of ARI and entropy
(5)Combine Z_{ARI} and Z_{E} into F-score
Functionally Important Sites
(6)Classifier
Prediction of Orthologous Relations

Fig. 1 System control flow.

5. Status report

5.1. Data sets

We tested PORFIS on the PurR/LacI family and the protein kinase AGC group family to demonstrate its ability to identify functionally important sites. We later applied PORFIS to the AGC family, the glycoprotein hormone family, the α-proteasome family and the growth hormone family to evaluate its performance in the prediction of orthologous relations.

The PurR/LacI family is a large family of bacterial transcription factors divided into 15 orthologous subfamilies [17]. Some of the subfamilies are relatively small. To avoid the bias incurred by the skewed subfamily size, we applied random shuffling techniques to fill in random sequences for balancing the size among all the subfamilies. The AGC family is related to phosphorylation in the process of signal transduction in living organism and is divided into six orthologous subfamilies [23]. We thank Mirny and Lee for providing these two datasets. The α-proteasome family has seven orthologous subfamilies and was downloaded from NCBI according to Bouzat et al. [24]. The subfamilies in glycoprotein hormone family we chose in our study are FSHβ, TSHβ and LHβ. The members of the glycoprotein hormone family are crucial to the complex endocrine system that regulates normal growth, sexual development, and reproductive function. The growth hormone family, which plays an important role in growth control, is divided into three major subfamilies, PL, GH, and PRL. Both the glycoprotein hormone family and the growth hormone family can be downloaded from NCBI. The datasets are summarized in Table 2.

5.2. Performance evaluation

Sensitivity and positive predictive value (PPV) are two commonly used performance measures. They are defined as follows:

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \tag{9}$$

$$\text{PPV} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \tag{10}$$

Higher sensitivity of a prediction algorithm reflects its ability to identify more true positives. On the other hand, higher positive predictive value indicates it is more capable of avoiding false positives. However, for most prediction algorithms, it is difficult to obtain a higher score for one of the measures without sacrificing the other because these two measures generally contradict each other. To consider both measures at the same time, we further combine them into an *F*-score [22] to evaluate the overall performance. The definition of *F*-score on prediction is defined as:

$$\text{Prediction } F\text{-score} = \frac{2}{((1/\text{Senxitivity}) + (1 + \text{PPV}))} \tag{11}$$

6. Lessons learned

6.1. Identification of important sites

We compared PORFIS with Mirny and Gelfand's [17] in the identification of functionally important sites within two families, the PurR/LacI family and the AGC family. The results are presented in Tables 3 and 4. The reason for this particular comparative study with Mirny and Galfand's is that their method is one of the most recent studies of important sites, and they were kind enough to provide the same data used in their work so we can keep the consistency in experiments.

In the PurR/LacI family there are twelve important sites published in literature, nine of which are binding sites (DNA or ligand) and the rest interact with other residues or contribute to the protein

Table 2 Datasets used in our study

Family (number of sequences)	Subfamily (number of sequences)
PurR/LacI(54)	araR(2) kdgR(2) ccpA(12) degA(4) yjmH(2) rbsR(4) purR(4) cytR(3) galSR(4) ascG(4) LacI(2) treR(3) gntR(3) idnR(2) fruR(3)
AGC(380)	PKA(69) RAC(23) GRK(58) S6PK(41) PVPK1(50) PKC(139)
α-Proteasome(54)	A1(7) A2(6) A3(7) A4(8) A5(8) A6(10) A7(8)
GPH(60)	FSHb(20) TSHb(20) LHb(20)
GH(35)	GH(12) PL(11) PRL(12)

Thanks to Mirny et al. for providing the PurR/LacI and AGC datasets. The α-proteasome, GPH (glycoprotein hormone) and GH (growth hormone) families were downloaded from NCBI.

Table 3 Important sites identified in the PurR/LacI family

Site	Residue	Mirny and Gelfand	PORFIS	Description
15	Thr	*	+	DNA binding site
16	Thr	*	+	DNA binding site
50	Val	*	+	Contact to other residue
53	Ser		+	Diminish repression
55	Lys	*	+	DNA binding site
85	Cys		+	Putative important site
91	Thr		+	Putative important site
98	Trp	*	+	Putative important site
107	Tyr		+	Putative important site
114	Lys	*		Putative important site
122	Met		+	Ligand binding site
124	Ser		+	Avoid steric clash with the side chain of Arg190
145	Met		+	Putative important site
146	Asp	*	+	Ligand binding site
147	Trp		+	Ligand binding site
148	Gly		+	Putative important site
160	Asp	*	+	Ligand binding site
221	Phe	*	+	Ligand binding site
249	Ile		+	Ligand binding site

Sites 15, 16, 50, 53, 55, 122, 124, 146, 147, 160, 221 and 249 have been proved to be important [25—27]. Note that the site numbering conforms to the PurR, PDB code 1WET [26].

conformation. We underscore these twelve important sites in Table 3. The sites found by Mirny and Gelfand are labeled with a '*'; the sites identified by our method, with a '+'. As shown in Fig. 2, PORFIS successfully identified all the twelve important sites plus six putative sites that are in the proximity of the published binding sites. According to previous studies, Thr15, Thr16 and Lys55 are deeply buried in the DNA groves forming a dense network of interactions with the bases. Met122, Asp146, Asp160, Phe221 and Ile249 are within 8 Å from the ligand in PurR. Val50 forms a hydrophobic contact with its counterpart. Ser53 diminishes (but not abrogates) repression, and Ser124 is located directly above the corepressor-binding pocket and takes this conformation to avoid steric clash with the side chain of Arg190, which contributes to corepressor binding affinity. Trp147 is the key switch residue in the corepressor binding pocket. In the holorepressor, Trp147 is far from the corepressor binding pocket and stacks against Tyr126. However, in the open state, it rotates into the ligand binding pocket, resulting in a 10.7 Å translation of its Nε atom. In this position, Trp147 hydrogen binds to the side chain of Tyr73 by its Nε and stacks against Phe74. Thus, Trp147 may play a role as the structural surrogate of corepressor to stabilize the open conformation, which controls the operator DNA binding affinity [17,25—27]. All the residues identified by PORFIS are either directly or indirectly related to protein functions.

Table 4 presents the result of the AGC family. The sites labeled with a '*' were identified by Li et al. [23], and those with a '+' were predicted by PORFIS. As shown in Table 4, there are 36 published important sites which are underlined, including the substrate binding sites, the Mg_2ATP binding sites, and some residues that are close to or interact with these binding sites. PORFIS identified 25 important sites, seven (Ser53, Leu82, Gln84, Phe129, Thr183, Thr197 and Pro202) of which are substrate binding sites or ATP binding sites, two (His87 and Pro243) of which make direct contact with other residues, one (Arg133) pack to Arg side chain at P-3, two (Thr48 and Arg56) of which belong to the nucleotide positioning motif, and four (Lys83, Pro169, Ala188 and Trp195) of which are next to particular binding sites. Thr48 and Arg56 located in the glycine-rich loop (residue 48—57) and are two of the residues that constitute a nucleotide positioning motif, which spans the entire length of the wedge shaped nucleotide binding pocket and forms the ceiling of this pocket with the nucleotide fitting snugly against this motif. His87 binds Thr197 under some condition. Phe129 generates the specificity of Pro or Met at P-3. Pro169 lie in the catalytic loop. Our method identified four residues (Ala188, Thr195, Thr197 and Pro202) situated in the activation loop of cAMP-dependent protein kinase (see Fig. 3). Thr197 coordinates the activation loop and contributes to the correct configuration of residues at the active site cleft. The hydrophobic residue Pro202 is one of the residues forming the binding pocket. In mitogen-activated protein kinases, the corresponding residue of position 202 (Leu) is diagnostic of the obligatory Pro specificity at P + 1 [28—30]. Sixteen residues identified by our method have been confirmed with biological meanings.

The results of sensitivity and positive predictive value are summarized in Table 5. The sensitivity and positive predictive value of PORFIS are 1.000 and

Table 4 Important sites identified in the protein kinases AGC family

Site	Residue	Li et al.	PORFIS	Description
48	Thr	*	+	Nucleotide positioning Motif
52	Gly			Substrate binding site
53	Ser	*	+	Substrate binding site
54	Phe			Substrate binding site
55	Gly			ATP-binding site
56	Arg	*	+	Nucleotide positioning Motif
72	Lys			ATP-binding site
77	Gln		+	Putative important sites
82	Leu		+	Substrate binding site
83	Lys	*	+	Next to substrate binding site
84	Gln	*	+	Substrate binding site
87	His	*	+	Bind 197 under some condition
91	Glu			ATP-binding site
109	Ser		+	Putative important sites
115	Asn	*	+	Putative important sites
118	Met		+	Putative important sites
121	Glu			ATP-binding site
123	Val			ATP-binding site
127	Glu			Substrate binding site
129	Phe	*	+	Substrate binding site
130	Ser			Substrate binding site
133	Arg	*	+	Pack to Arg side chain at P-3
156	Tyr		+	Putative important sites
166	Asp			ATP-binding site
168	Lys			ATP-binding site
169	Pro		+	Next to substrate binding site
170	Glu			Substrate binding site
171	Asn			ATP-binding site
181	Gln		+	Putative important sites
183	Thr		+	ATP-binding site
187	Phe			Substrate binding site
188	Ala		+	Next to substrate binding site
195	Thr		+	Next to substrate binding site
196	Trp	*		Substrate binding site
197	Thr		+	Substrate binding site
198	Leu			Substrate binding site
201	Thr			Substrate binding site
202	Pro		+	Substrate binding site
203	Glu			Substrate binding site
204	Tyr			Substrate binding site
205	Leu			Substrate binding site
226	Val		+	Putative important sites
230	Glu			Substrate binding site
243	Pro	*	+	Bind 15 under some condition
247	Tyr	*	+	Putative important sites
249	Lys	*	+	Putative important sites

The published important sites are underlined. Note that Li et al. used Mirny and Gelfand's method [17] to find important sites. Residues underscored are those that contribute to substrate binding, protein structure or the residues next to these sites. PORFIS identified more important sites in the activation loop (as shown in boldface) of the cAMP-dependent protein kinase. The site numbering conforms to the cAMP-dependent protein kinase, PDB code 1ATP [35].

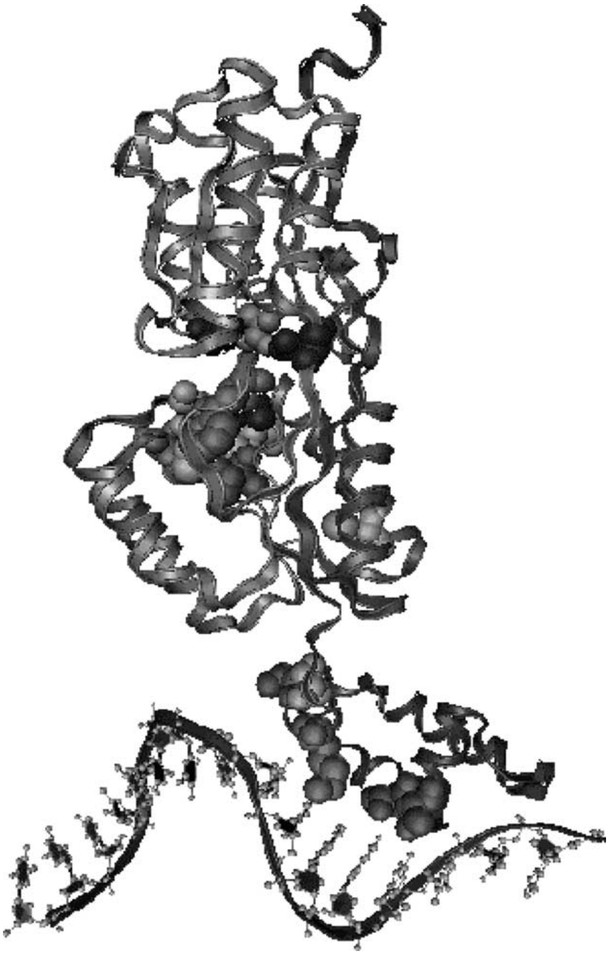

Fig. 2 Structure of PurR bound to DNA [26]. The important residues identified by our method are shown as spheres. The structure is obtained from PDB 1wet. The figure was generated by VMD [36].

0.667 in the PurR/LacI family, 0.444 and 0.640 in the AGC family, respectively. In both cases, PORFIS achieves higher *F*-scores than Mirny and Gelfand's [17,23]. Furthermore, PORFIS requires less CPU time than Mirny and Gelfand's, which is hindered by the complex resampling procedure. Simulated on an AMD Athlon 1.0 G machine with 512 MB RAM, our computational time was in the order of minutes compared with hours of Mirny and Gelfand's.

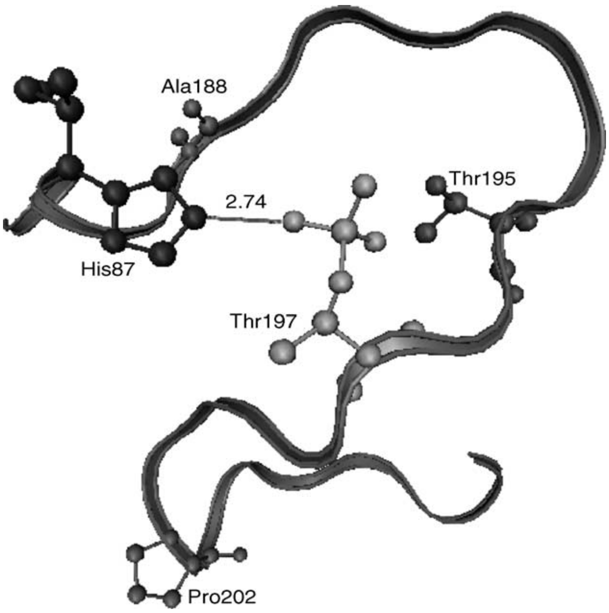

Fig. 3 The activation loop of cAMP-dependent protein kinase. Important residues identified by our method in the loop are shown as spheres. The conditional interaction of His87 and Thr197 is also illustrated as a link. The figure was generated by VMD [36].

Besides demonstrating PORFIS's performance of identifying functionally important sites within real protein families, we also applied the Monte Carlo simulation to verify the statistical significance of the sites found. By random shuffling of the amino acids in the given protein sequences to keep their distributions, we create a random protein family that is later used to generate a multiple sequence alignment background. For each position in the background alignment, we compute its *F*-score. The same procedure can be repeated, and the *F*-score of each position in the alignment is averaged over 10,000 times as shown in Figs. 4 and 5. By comparing the *F*-score of each position between the real protein family and the random background, we notice there exists significant difference between the *F*-score distributions, which suggests that the predicted sites are indeed statistical significant. Furthermore, we also conducted an receiver operating

Table 5 Prediction result for the two protein families

	Mirny and Gelfand		PORFIS	
	PurR/LacI	AGC	PurR/LacI	AGC
Sensitivity	0.583	0.250	1.000	0.444
Positive predictive value	0.778	0.563	0.667	0.640
F-score	0.667	0.346	0.800	0.525

The total number of published important sites of PurR/LacI and AGC is 12 and 36, respectively.

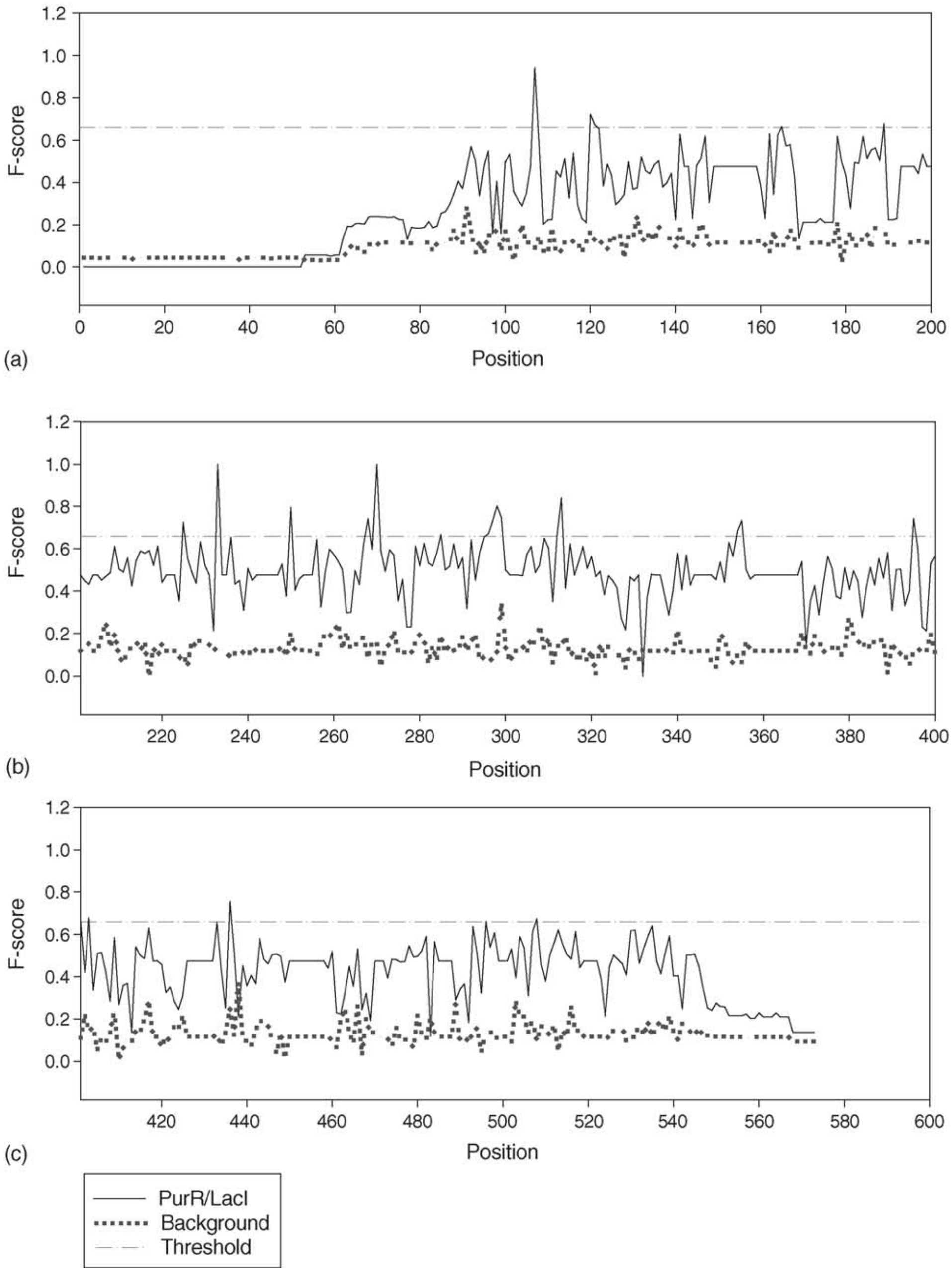

Fig. 4 Results of the Monte Carlo simulations for the *F*-score distribution in the PurR/LacI family and the background. The *F*-score of each site in the alignment is plotted in (a), (b) and (c), respectively along its positions.

characteristic (ROC) analysis, and the results are summarized in Fig. 6. It shows that PORFIS performs better on the PurR/LacI family. This may be due to that the subfamily size variation is smaller in the PurR/LacI family, so that PORFIS can more easily identify the correct important sites, compared with the AGC family as shown in Table 5. Though there is some difference between the ROC curves for the two protein families in our experiments, yet both curves indicate reasonable performance of PORFIS.

6.2. Prediction of orthologous relations

We tested PORFIS on the AGC family, the glycoprotein hormone family (GPH), the α-proteasome family and the growth hormone family (GH) to demonstrate its performance in the prediction of orthologous relations. For comparison, CLUSTALW [31], profile HMMs [2], PSI-BLAST [32] and Meta-MEME [33,34] were tested on the same data. Ten times of three-fold cross validation were used to evalu-

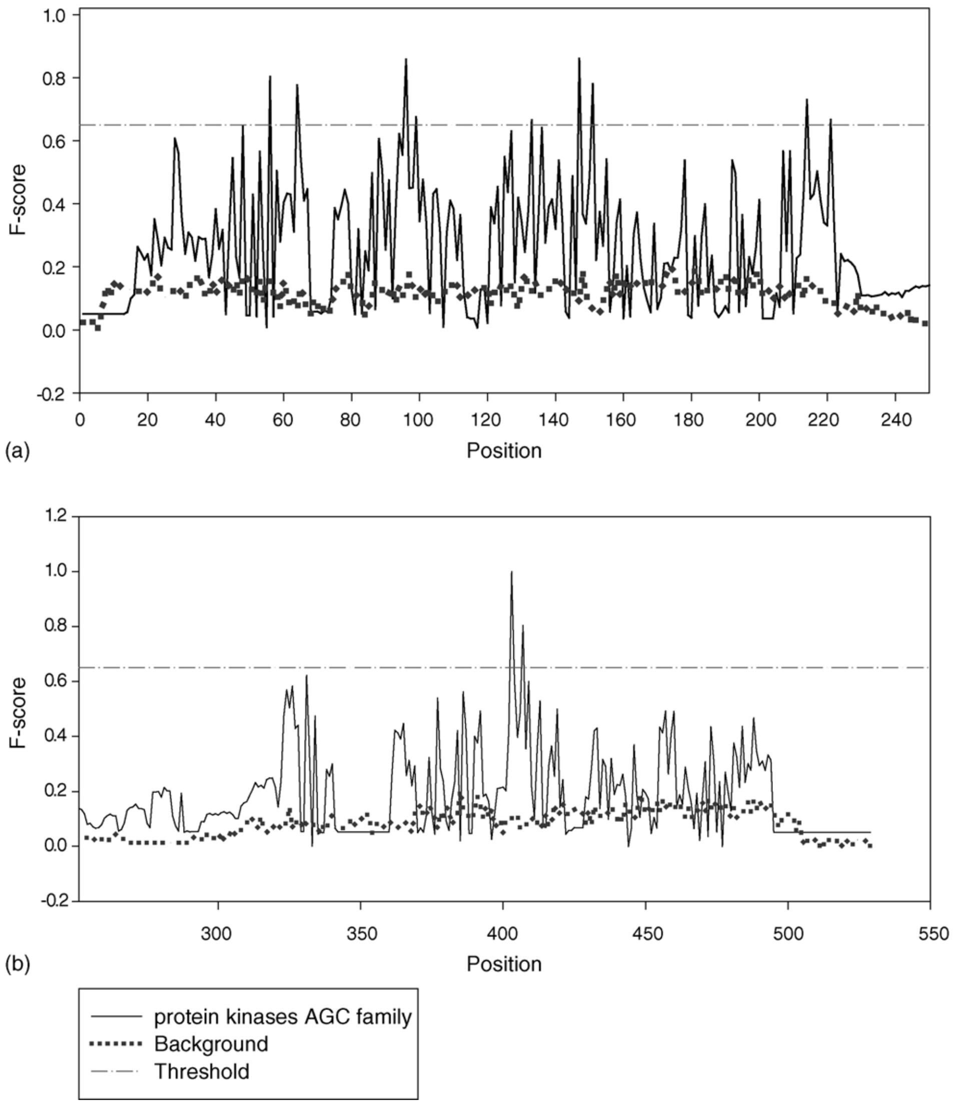

Fig. 5 Results of the Monte Carlo simulations for the *F*-score distribution in the protein kinase AGC family and the background. The *F*-score of each site in the alignment is plotted in (a) and (b), respectively along its positions.

Table 6 Orthology prediction accuracies of four families

Protein families	AGC group family ($\mu \pm \sigma$)	G-hormone family ($\mu \pm \sigma$)	α-Proteasome family ($\mu \pm \sigma$)	Growth hormone family ($\mu \pm \sigma$)
CLUSTALW	0.841 ± 0.036	0.873 ± 0.066	1.000 ± 0.000	0.800 ± 0.070
ProfileHMM	0.987 ± 0.008	1.000 ± 0.000	1.000 ± 0.000	0.913 ± 0.113
PSI-BLAST[a]	0.855 ± 0.056	0.731 ± 0.103	0.934 ± 0073	0.859 ± 0.099
PSI-BLAST[b]	0.856 ± 0.051	0.836 ± 0.133	0.934 ± 0.073	0.831 ± 0.071
Meta-MEME[c]	0.921 ± 0.050	0.950 ± 0.046	0.944 ± 0.078	0.916 ± 0.053
Meta-MEME[d]	0.988 ± 0.008	0.992 ± 0.020	1.000 ± 0.000	0.859 ± 0.066
Mirny's method[e]	0.828 ± 0.087	0.709 ± 0.326	0.956 ± 0.054	0.773 ± 0.092
PORFIS	0.976 ± 0.009	0.977 ± 0.029	1.000 ± 0.000	0.937 ± 0.026

[a] PSI-BLAST using the default iteration threshold 0.005.

[b] PSI-BLAST using iteration threshold 1e-10.

[c] The parameters were set as default: −nmotifs 1 and −maxw 50.

[d] The parameters were set as −nmotifs 5 and −maxw 20.

[e] We re-implemented Mirny's method to find important sites and then applied the same classification procedure as in our method to make predictions. Note we set cutoff MI = 0.8 and $P(I) = 1/L$, where L is the length of multiple sequence alignment, after personal contact with Mirny.

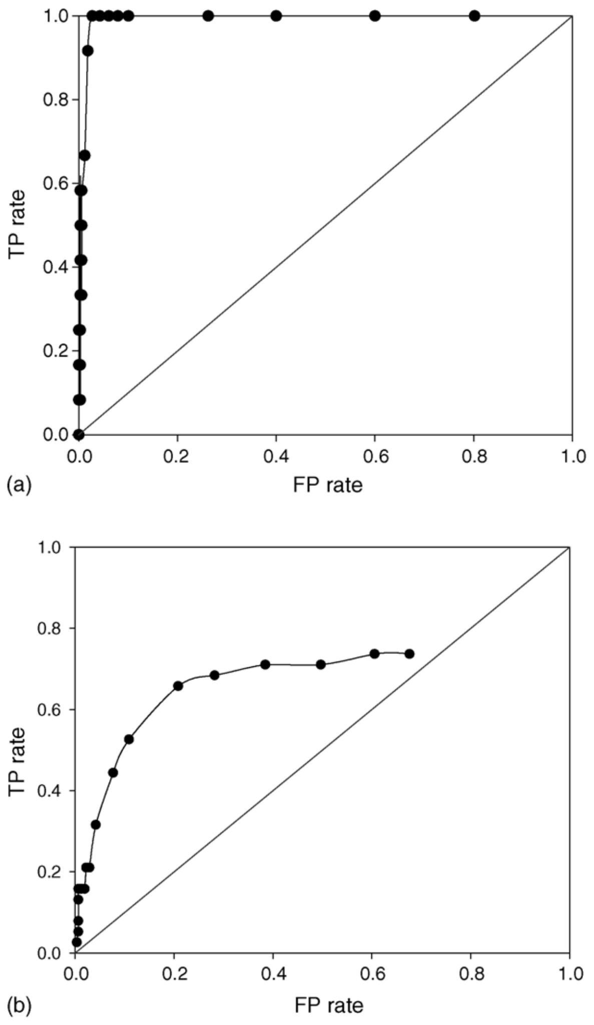

Fig. 6 Summary of the ROC analysis for important site prediction; (a) is the ROC curve for the PurR/LacI family; (b) for the protein kinase AGC family.

ate the prediction accuracy. In each run, we used one third of the data for testing, and the remaining data for training. The results were summarized in Table 6.

The results show that PORFIS is comparable with others. The likely reason why PORFIS outperforms CLUSTALW and PSI-BLAST is the following. PORFIS makes prediction based on functionally important sites that are associated with subfamilies. Thus, it avoids being misled by other irrelevant sites during classification. On the other hand, to make classifications based on sequence alignments through CLUSTAL and PSI-BLAST, we strongly rely on the sequence similarity; however, sequence alignments may inherently contain some irrelevant sites that can jeopardize the prediction. In addition to the given protein families, Meta-MEME requires a set of motif models found by MEME. Meta-MEME combines these models into a single motif-based hidden Markov model. As a consequence, the performance of Meta-MEME in classification highly depends on the motif models, such as the number of motifs used to form the single HMM and the width of the motifs, etc. The increase of motif number and width generally reduces the complexity of the final HMM produced, but it also incurs the loss of sequence information by over-generalizing sequence segments into motifs. Moreover, as the motif width

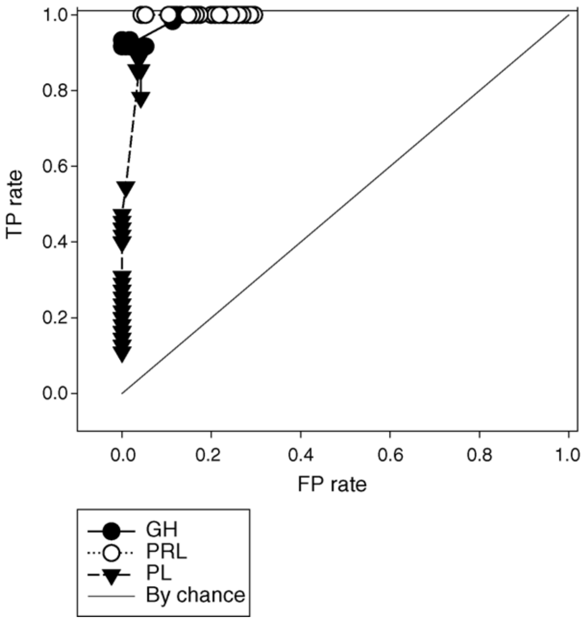

Fig. 7 The ROC curve for the classification of the GH family.

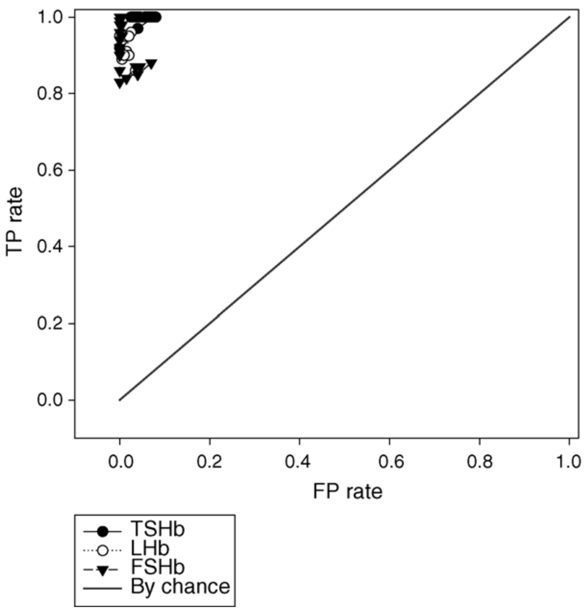

Fig. 8 The ROC curve for the classification of the GPH family.

Table 7 Some important residues identified from the four families

	Protein families							
	AGC group family		G-hormone family		α-Proteasome family		Growth hormone family	
	PKA[a]	53	FSH[a]	7	α2[a]	3	GH[a]	12
		82		24		7		18
		84		39		8		62
		87		59		9		63
		129		96		15		
		133	LH	71	α3[a]	4	PL[a]	60
		183		94		8		66
		197		95		9		107
		202				10		108
		243				16		
					α4[a]	2		
						6		
						7		
						8		
						14		
Reference	[28,29]		[37—39]		[40,41]		[42]	

[a] The important site numbering is as follows: PKA conforms to pdb 1ATP, FSH conforms to gi120552, LH conforms to gi1170834, α2 conforms to gi130880, α3 conforms to gi130861, α4 conforms to gi730374, GH conforms to gi1070555, and PL conforms to gi130300.

increases, it is more likely to include non-conserved positions in a motif. Similar to those irrelevant sites in alignments produced by CLUSTALW or PSI-BLAST, these degenerate positions can mislead classification. Profile HMMs had an almost perfect prediction of the AGC family, the glycoprotein hormone family, and the α-proteasome family, but it performs worse than PORFIS on the growth hormone family. As HMM methods usually require more training data to tune the probability parameters, it is no surprise that

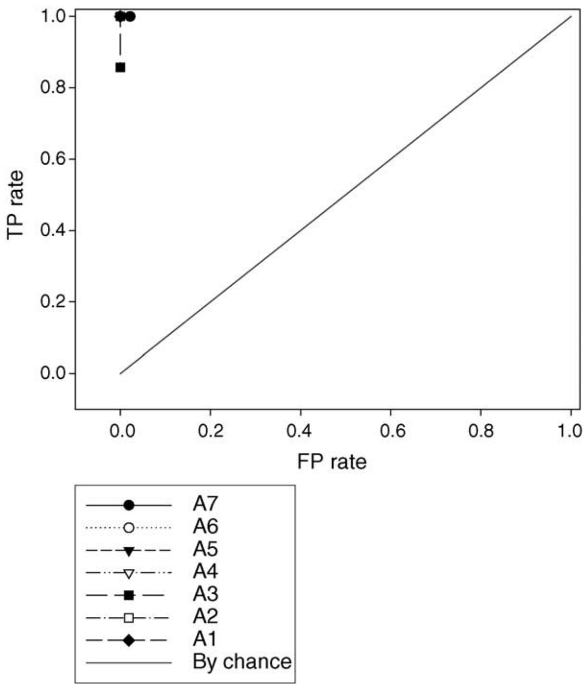

Fig. 9 The ROC curve for the classification of the α-proteasome family.

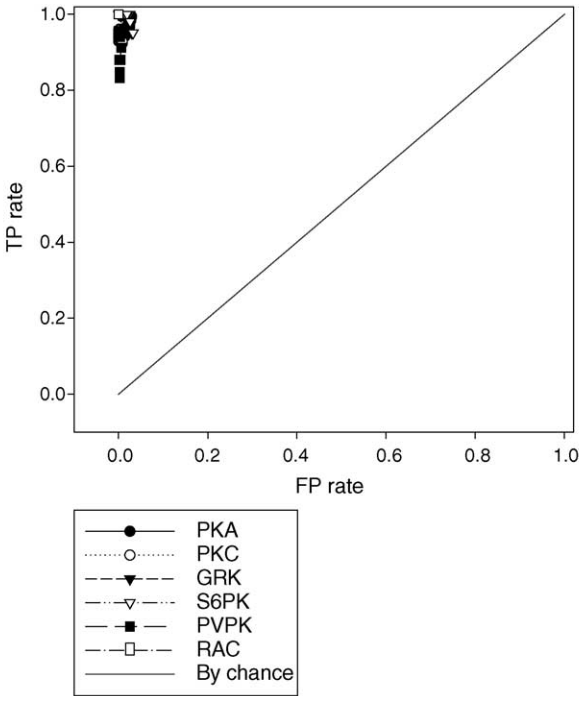

Fig. 10 The ROC curve for the classification of the AGC family.

Profile HMMs did the worst on the smallest family (i.e. the growth hormone family). Like most other methods, profile HMMs provides only classification results, but lacks comprehensible interpretations of the orthologous relationship. Unlike others, PORFIS makes a prediction based on the functionally important sites carrying biological meanings. Associations between functionally important residues and evolutionary relations can be established by our method. The orthologous relations based on the functional sites predicted by PORFIS can be further analyzed by site-directed mutagenesis. By biological verification, the functionally important sites prove to characterize orthologs more effectively and efficiently. Some of the important sites found by PORFIS are presented in Table 7.

We also did the ROC analysis for the four families by varying the *F*-score threshold in PORFIS. Since each family contains multiple subfamilies, we conducted ROC analysis for each subfamily separately and put the curves together in a single figure. The results are presented in Fig. 7 through 10. These figures show that most of the curves are clustered on the upper left corner, which suggests that the PORFIS's classification is not only reasonably accurate but also stable, even under varying thresholds.

7. Conclusion and future plans

We have proposed a method capable of not only identifying functionally important sites in a set of homologous proteins, but also predicting orthologous relationship for new protein sequences. It first identifies the functionally important residues related to the specificity and conservation among paralogous and orthologous proteins, and then uses these residues to construct a model to classify unknown protein sequences.

For the PurR/LacI family, our method successfully identified the important binding sites together with some residues contribute to the structural conformation of the protein. As for the AGC family, we found 16 residues that are located in the binding domains or interact with other important sites to form particular conformation related to the kinase function. We identified several active sites in the active cleft between the two lobes with the adenine ring of ATP deeply buried at the base of the cleft. Many of the important sites we identified interact with other residues to form the interaction network.

In addition to demonstrating the ability of our method to detect functionally important sites, we also systematically evaluated its performance in the prediction of orthologous relations with four families. Compared with other approaches, our method is more accurate and efficient in general. Unlike most previous works, besides the prediction of orthologous relationship, our method also suggests useful associations between functionally important sites and orthologous families. This type of information may provide biologists with new research topics and eventually become useful domain knowledge.

The current version of PORFIS can be further improved in two directions. Firstly, as multiple sequence alignment is essential to the identification of important sites, we can improve the quality of sequence alignment by incorporating more background knowledge to ensure the correctness of the alignment. Secondly, associations between important sites and their physicochemical properties can be further exploited to refine the prediction accuracy.

Acknowledgements

We thank Dr. Mirny and Dr. Gelfand for providing two of the datasets used in our experiments. Thanks also go to the anonymous reviewers and the editor for their valuable suggestions and comments. This work is partially supported by National Science Council and MediaTek research center at EECS of NCTU in Taiwan.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Molec. Biol. 215 (1990) 403—410.

[2] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755—763.

[3] W. Fitch, Distinguishing homologous from analogous proteins, Syst. Zool. 19 (1970) 99—113.

[4] R.F. Doolittle, D.F. Feng, S. Tsang, G. Cho, E. Little, Determining divergence times of the major kingdoms of living organisms with a protein clock, Science 271 (1996) 470—477.

[5] S. Henikoff, E.A. Greene, S. Pietrokovski, P. Bork, T.K. Attwood, L. Hood, Gene families: the taxonomy of protein paralogs and chimeras, Science 278 (1997) 609—614.

[6] R.L. Tatusov, D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, E.V. Koonin, The COG database: new developments in phylogenetic classification of proteins from complete genomes, Nucleic Acids Res. 29 (2001) 22—28.

[7] M. Remm, C.E.V. Storm, E.L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, J. Molec. Biol. 314 (2001) 1041—1052.

[8] T. Xie, D. Ding, Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale, Gene 261 (2000) 305—310.

[9] Y.P. Yuan, O. Eulenstein, M. Vingron, P. Bork, Towards detection of orthologues in sequence database, Bioinformatics 14 (1998) 285—289.

[10] R.J. Cotter, D.R. Caffrey, D.C. Shields, Improved database searches for orthologous sequences by conditioning on outgroup sequences, Bioinformatics 18 (2002) 87—91.

[11] C.E.V. Storm, E.L.L. Sonnhammer, Automated ortholog inference from phylogenetic trees and calculation of orthology reliability, Bioinformatics 18 (2002) 92—99.

[12] C.D. Livingstone, G.J. Barton, Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation, Comput. Appl. Biosci.: CABIOS 9 (1993) 745—756.

[13] M.J. Sternberg, F.E. Cohen, Interferon: a tertiary structure predicted from amino acid sequences, Philos. Trans. R. Soc. Lond. Ser. B, Biol. Sci. 299 (1982) 125—127.

[14] O. Lichtarge, H.R. Bourne, F.E. Cohen, An evolutionary trace method defines binding surfaces common to protein families, J. Molec. Biol. 257 (1996) 342—358.

[15] G. Casari, C. Sander, A. Valencia, A method to predict functional residues in proteins, Nat. Struct. Biol. 2 (1995) 171—178.

[16] R.S. Hannenhalli, R.R. Russell, Analysis and prediction of functional sub-types from protein sequence alignments, J. Molec. Biol. 303 (2000) 61—76.

[17] L.A. Mirny, M.S. Gelfand, Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors, J. Molec. Biol. 321 (2002) 7—20.

[18] P.J. Bickel, K.J. Kechris, P.C. Spector, G.J. Wedemayer, A.N. Glazer, Finding important sites in protein sequences, Proc. Nat. Acad. Sci. U.S.A. 99 (2002) 14764—14771.

[19] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193—218.

[20] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (1971) 846—850.

[21] L. Kuo, J. Lee, P. Cheng, J. Pai, Bayes inference for technological substitution data with data-based transformation, J. Forecast. 16 (1977) 65—82.

[22] D. Lewis, W.A. Gale, A sequential algorithm for training text classifier, in: Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 3—12.

[23] L. Li, E.I. Shakhnovich, L.A. Mirny, Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases, Proc. Nat. Acad. Sci. U.S.A. 100 (2003) 4463—4468.

[24] J.L. Bouzat, L.K. McNeil, H.M. Robertson, L.F. Solter, J.E. Nixon, J.E. Beever, H.R. Gaskins, G. Olsen, S. Subramaniam, M.L. Sogin, H.A. Lewin, Phylogenomic analysis of the alpha proteasome gene family from early-diverging eukaryotes, J. Molec. Evol. 51 (2000) 532—543.

[25] J. Suckow, P. Markiewicz, L.G. Kleina, J. Miller, B. Kisters-Woike, B. Muller-Hill, Genetic studies of the Lac repressor. XV. 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure, J. Molec. Biol. 261 (1996) 509—523.

[26] M.A. Schumacher, A. Glasfeld, H. Zalkin, R.G. Brennan, The X-ray structure of the PurR-guanine-purF operator complex reveals the contributions of complementary electrostatic surfaces and a water-mediated hydrogen bond to corepressor specificity and binding affinity, J. Biol. Chem. 272 (1997) 22648—22653.

[27] F. Lu, R.G. Brennan, H. Zalkin, *Escherichia coli* purine repressor: key residues for the allosteric transition between active and inactive conformations and for interdomain signaling, Biochemistry 37 (1998) 15680—15690.

[28] C.M. Smith, E. Radzio-Andzelm, P. Madhusudan, Akamine, S.S. Taylor, The catalytic subunit of cAMP-dependent protein kinase: prototype for an extended network of communication, Progress Biophys. Molec. Biol. 71 (1999) 313—341.

[29] R.I. Brinkworth, R.A. Breinl, B. Kobe, Structural basis and prediction of substrate specificity in protein serine/threonine kinases, Proc. Nat. Acad. Sci. U.S.A. 100 (2002) 74—79.

[30] L.A. Pinna, M. Ruzzene, How do protein kinases recognize their substrates? Biochim. Biophys. Acta 1314 (1996) 191—225.

[31] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673—4680.

[32] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389—3402.

[33] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, in: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, vol. 2, 1994, pp. 28—36.

[34] W.N. Grundy, T.L. Bailey, C. Elkan, M.E. Baker, Meta-MEME: Motif-based hidden Markov models of protein families, Comput. Appl. Biosci. 13 (1997) 397—406.

[35] J. Zheng, D.R. Knighton, L.F. ten Eyck, R. Karlsson, N. Xuong, S.S. Taylor, J.M. Sowadski, Crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MgATP and peptide inhibitor, Biochemistry 32 (1993) 2154—2161.

[36] W. Humphrey, A. Dalke, K. Schulten, VMD—visual molecular dynamics, J. Molec. Graph. 14 (1996) 33—38.

[37] K.M. Fox, J.A. Dias, P. Van Roey, Three-dimensional structure of human follicle-stimulating hormone, Molec. Endocrinol. 15 (2001) 378—389.

[38] J.A. Dias, Y. Zhang, X. Liu, Receptor binding and functional properties of chimeric human follitropin prepared by an exchange between a small hydrophilic intercysteine loop of human follitropin and human lutropin, J. Biol. Chem. 269 (1994) 25289—25294.

[39] W.R. Moyle, M.M. Matzuk, R.K. Campbell, E. Cogliani, D.M. Dean-Emig, A. Krichevsky, R.W. Barnett, I. Boime, Localization of residues that confer antibody binding specificity using human chorionic gonadotropin/luteinizing hormone beta subunit chimeras and mutants, J. Biol. Chem. 265 (1990) 8511—8518.

[40] A. Forster, F.G. Whitby, C.P. Hill, The pore of activated 20S proteasomes has an ordered 7-fold symmetric conformation, Eur. Molec. Biol. Organiz. J. 22 (2003) 4356—4364.

[41] M. Groll, M. Bajorek, A. Kohler, L. Moroder, D.M. Rubin, R. Huber, M.H. Glickman, D. Finley, A gated channel into the proteasome core particle, Nat. Struct. Biol. 7 (2000) 1062—1067.

[42] P.A. Elkins, H.W. Christinger, Y. Sandowski, E. Sakal, A. Gertler, A.M. de Vos, A.A. Kossiakoff, Ternary complex between placental lactogen and the extracellular domain of the prolactin receptor, Nat. Struct. Biol. 7 (2000) 808—815.