BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 24 2010, pages 3075–3082
doi:10.1093/bioinformatics/btq595

Structural bioinformatics

Advance Access publication October 26, 2010

Identification of subfamily-specific sites based on active sites modeling and clustering

Raquel C. de Melo-Minardi[1,*,†], Karine Bastard[2,3,4,†] and François Artiguenave[2,3,4,]

[1]Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil, [2]Genoscope, Institut de Génomique, Comissariat à l'énergie atomique et aux énergies alternatives, Evry cedex, [3]UMR 8030, Cenre National de la Recherche Scientifique, Evry cedex and [4]Université Evry Val d'Essonne, Evry F-91057, France

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Current computational approaches to function prediction are mostly based on protein sequence classification and transfer of annotation from known proteins to their closest homologous sequences relying on the orthology concept of function conservation. This approach suffers a major weakness: annotation reliability depends on global sequence similarity to known proteins and is poorly efficient for enzyme superfamilies that catalyze different reactions. Structural biology offers a different strategy to overcome the problem of annotation by adding information about protein 3D structures. This information can be used to identify amino acids located in active sites, focusing on detection of functional polymorphisms residues in an enzyme superfamily. Structural genomics programs are providing more and more novel protein structures at a high-throughput rate. However, there is still a huge gap between the number of sequences and available structures. Computational methods, such as homology modeling provides reliable approaches to bridge this gap and could be a new precise tool to annotate protein functions.

Results: Here, we present Active Sites Modeling and Clustering (ASMC) method, a novel unsupervised method to classify sequences using structural information of protein pockets. ASMC combines homology modeling of family members, structural alignment of modeled active sites and a subsequent hierarchical conceptual classification. Comparison of profiles obtained from computed clusters allows the identification of residues correlated to subfamily function divergence, called specificity determining positions. ASMC method has been validated on a benchmark of 42 Pfam families for which previous resolved holo-structures were available. ASMC was also applied to several families containing known protein structures and comprehensive functional annotations. We will discuss how ASMC improves annotation and understanding of protein families functions by giving some specific illustrative examples on nucleotidyl cyclases, protein kinases and serine proteases.

Availability: http://www.genoscope.fr/ASMC/.

Contact: raquelcm@dcc.ufmg.br; kbastard@genoscope.cns.fr; artigue@genoscope.cns.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Received on July 29, 2010; revised on October 1, 2010; accepted on October 17, 2010

1 INTRODUCTION

With the increasing number of genomes and meta-genomes sequenced, a critical challenge concerns the functional prediction of proteins encoded by novel predicted genes. Genes and proteins are commonly classified in terms of families, subfamilies or superfamilies according to different molecular taxonomy (Finn *et al.*, 2008; Orengo *et al.*, 1997; Sonnhammer *et al.*, 1997).

Clustering methods are aimed at identifying functionally related proteins and defining thresholds to distinguish truly related proteins from homologous proteins with different functions. All these methods rely on the Darwinian evolution concept; protein sequences are subjected to random mutations and selective pressure adds (or causes) functional modifications. The sequence is hence composed of well conserved positions and others that tolerate mutations, insertions and deletions. Methods to detect these positions are based on multiple sequence alignment (MSeqA) analysis using different parameters metrics: chemical properties of residues, evolution (Kalinina *et al.*, 2004), quantitative information analysis (Hannenhalli and Russell, 2000; Sol *et al.*, 2003), evolutionary-based analysis (Donaldo and Shakhnovich, 2005, 2009; Pei *et al.*, 2006), phylogeny-independent methods (Pazos *et al.*, 2006) or combination of properties (Chakrabarti *et al.*, 2007). For a few years, new methods have been proposed to include structural information for enhancing the quality of amino acids function prediction (Gong and Blundell, 2008; Goldenberg *et al.*, 2009; Henschel *et al.*, 2007; Kalinina *et al.*, 2009; Langraf *et al.*, 2001; Najmanovich *et al.*, 2008; Pupko *et al.*, 2002). Capra and Singh (2008); Kristensen *et al.* (2008); Lichtarge *et al.* (1996); Madabushi *et al.* (2002); Redfern *et al.* (2009); Ward *et al.* (2009) proposed an evolutionary trace (ET) procedure to predict active sites and functional interfaces in proteins with known structure.

In this work, we propose ASMC, a novel methodology for unsupervised classification of protein subfamilies allowing functional and specificity determining positions (SDPs) detection. According to Rausell *et al.* (2010), SDPs are related to fundamental regions that correspond to ligand binding and protein interaction sites. Methods developed to detect SDPs are based on multiple sequence alignments and we propose with ASMC to improve the detection by using structural alignments of active site

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

residues. While 3D structural information gives valuable insights to understand the molecular function of proteins, there is still a small fraction of experimentally solved structures in comparison with the amount of sequences. In this context, homology modeling provides structural information for the members of a protein family, using one or more templates (Moult, 2005; Tramontano and Morea, 2003; Yu *et al.*, 2005). Different methodologies have been developed to improve protein annotation of protein specificity using 3D information (Chakrabarti and Panchenko, 2009; Halabi *et al.*, 2009; Nagao *et al.*, 2010; Rottig *et al.*, 2010; Tseng *et al.*, 2009).

ASMC is phylogeny independent and relies on (i) an improved alignment of specific active site cavities based on 3D protein structure alignment and (ii) a classification of these alignments based on information theory. Each classified cluster is defined by a unique profile. Comparison of these profiles pinpoints SDPs correlated with functional intra-family diversification.

To test our method, we used a benchmark dataset defined by Kalinina *et al.* (2009) composed of 42 Pfam families. The dataset is divided into two classes of enzymes depending on their mono or multi-functional character [more than one enzyme classification (EC) number]. Sensitivity (Se) and specificity (Sp) obtained for multi-functional (Se = 47%, Sp = 68%) and mono-activity (Se = 72%, Sp = 37%) families demonstrate the reliability of ASMC predictions.

Detailed analysis of three well-known families confirmed the accuracy of SDP prediction by providing molecular explanation for identified SDPs residues.

2 MATERIALS AND METHODS

2.1 Data selection

The test set benchmark was extracted as described in Kalinina *et al.* (2009) (Pfam families are listed in Supplementary Material). The set is composed of enzymes families with well-characterized functions and at least one structure with bound natural ligands available in the Protein Data Bank (PDB). One part of the dataset is composed of protein families acting on a variety of substrates (diverse dataset), whereas the second set is composed of mono-activity protein families. With respect to the original set, two families were deleted: (i) PF00896 is a 'dead' Pfam family and has been removed from the database, (ii) in PF03061, no pocket corresponding to active sites was detected.

Protein sequences were selected from the Pfam database (Finn *et al.*, 2008) and length and identity filters were applied to remove sequences with lengths differing by more than one SD from the family average length or sequences with less than 30% similarity to one of the template structures. PDB identifiers for the templates were extracted from the same database.

Nucleotidyl cyclases: we selected 2201 sequences from PF00211. This set presented sequences of 199.80 ± 70.13 residues and after selection by size we retained 1646 sequences of 187.25 ± 14.44 residues. We used one template for guanylate cyclases (PDB id:chain, 3ET6:A), one for adenylate cyclases (1AB8:A) and we analyzed 536 sequences with more than 30% identity to these templates. These 536 sequences have 41.18 ± 7.61% sequence identity with their template sequences.

Protein kinases: we selected 33 665 sequences from Pfam families PF00069 and PF07714 (protein lengths: 219.36 ± 81.09 residues). A total of 3403 sequences were selected by the length and identity different filters (templates used 2CPK:E for serine/threonine kinase and 1U46:A for tyrosine kinase). The average sequence identity with the templates was 35.96 ± 7.49%.

Serine proteases: we selected 7256 sequences from PF00089 (193.82 ± 57.23 residues) and kept 6016 after filtering by size (202.85 ± 30.60 residues). The template structures are 1EST:A for elastase, 5PTP:A for trypsin and 1AB9: (we joined chains A, B, C and D to form one chain) for chymotrypsin. The retained 1686 serine proteases with more than 30% identity with the templates have an average identity with them of 39.04 ± 10.08%.

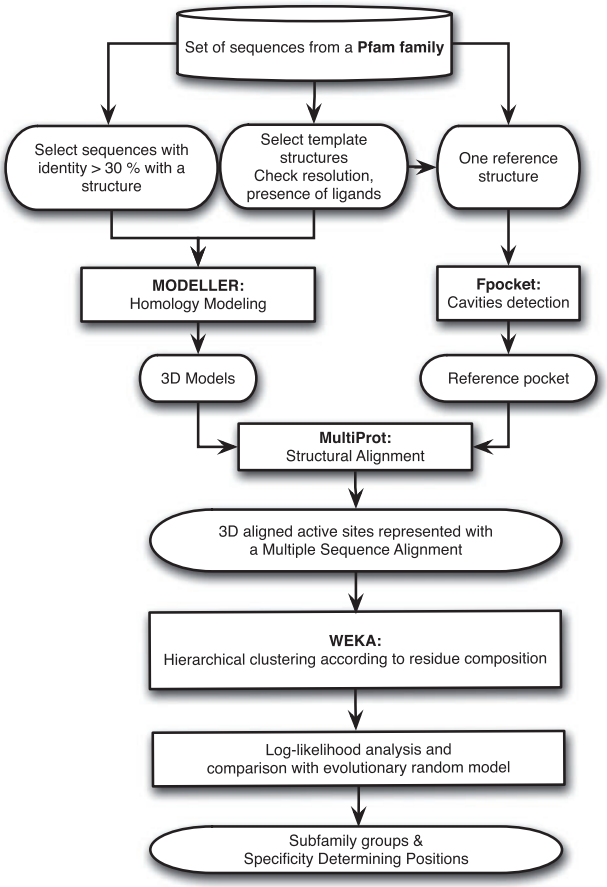

Fig. 1. ASMC method diagram. Software and database are written in bold.

2.2 ASMC methodology

The different steps of the ASMC method are summarized in Figure 1.

Modeling: homology models using multiple templates are constructed using Modeller version 9v6 (Eswar *et al.*, 2008). Prior, for each family, a 'structure profile' was built with a multiple structure alignment using the SALIGN module (Madhusudhan *et al.*, 2009).

For each sequence, we built 50 models and the best model was selected as the one with the lowest objective function. This function corresponds to the total energy of the system, based on CHARMM all atom potential, and indicates a model with the best fit with the input data (inter-residue distances and residue dihedral angles, i.e. stereochemical constraints).

Pocket prediction and selection: detection of pockets was performed on a reference structure, chosen as the holo form structure if exists or an apo form with crystallographic structure with the best resolution. Cavities are computed using Fpocket software (Le Guilloux *et al.*, 2009), based on the theory of alpha-shapes, Voronoi diagrams and Delaunay triangulation. Fpocket ranks cavities as the most probable active site. Nevertheless, we have chosen to perform our ASMC over all predicted pockets. We present results for the most conserved pockets, which turned out to be the enzyme active site. In majority of the cases, the first ranked cavity by Fpocket software was

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

the enzyme active site, except for families PF00278, PF01227, PF01583, PF02901 and PF03414.

Structural 3D alignment: the alignment of pocket residues were obtained using MultiProt software (Shatsky *et al.*, 2004). Results of the multiple structure alignment of modeled pockets were compiled to build a multiple sequence alignment (MSA) where each residue of the modeled pockets is positioned relatively to the reference structure.

Clustering: selected residues of catalytic pockets identified by the 3D alignment are used as the support to separate subfamilies by a conceptual clustering approach (Fisher, 1987) . This approach creates a hierarchical categorization of the instances (protein pockets), generating a conceptual description for each class in which concepts are sets of residues responsible for functional differentiation in a protein family. The algorithm, COBWEB (Fisher, 1987) was run using WEKA software (Holmes *et al.*, 1994) and the following parameters: C = 0.25, A = 1.0.

Statistical significance: a log-likelihood analysis (Pei *et al.*, 2006) is computed for each MSA position. This procedure is a test based on the ratio between the probabilities of two different hypotheses. For each position i of the MSA, the frequency of the most frequent residue in a specific cluster is computed both for the individual cluster and all the others merged. The null hypothesis states that the most frequent residue is found with similar frequencies in the cluster and elsewhere so that it is not important for the specificity. The alternative hypothesis states the opposite. The log-likelihood analysis states on the specificity signification (the null hypothesis tends to be rejected with high values). The log-likehood $log(L_i)$ is approximated as:

$$\log(L_i) \approx 2\left[a \log\left(\frac{a}{\frac{c\ (a+b)}{(c+d)}}\right) + b \log\left(\frac{b}{\frac{d\ (a+b)}{(c+d)}}\right)\right] \quad (1)$$

where a corresponds to the number of times that we find the most frequent residue in position i in a specific cluster, b is the frequency of this residue in the other clusters, c is the number of sequences in the specific cluster and d is the number of sequences in the other clusters.

Residues are shuffled at all positions of the MSA and a new log-likelihood is computed. We perform 100 simulations and consider the average value for the log-likelihood computation. Thus,

$$\mathrm{LLR_i} = \frac{\log(L_i)}{<\log(L_i^{sh})>} \quad (2)$$

where L_i is the log-likelihood for the original residues of the position i and $<L_i^{sh}>$, the average log-likelihood for the 100 shuffle observations.

To obtain a significant scoring function tracking the relevant specific residues, ASMC constructs a random model based on the frequencies of amino acids at each position. This randomization uses the Whelan and Goldman (WAG) substitution matrix (Whelan and Goldman, 2001). For this second step, 1000 simulations are performed by shuffling the original residues at each position, allowing them to mutate according to substitution probabilities. The LLR_i^{sim} is computed for each simulation and the Z-score is obtained as follows (P-value is deduced from Z-score):

$$Z\text{-score} = \frac{\mathrm{LLR}_i - <\mathrm{LLR}_i^{sim}>}{\sigma(\mathrm{LLR}_i^{sim})} \quad (3)$$

SDPs/CPs/OPs categories: For each family, SDPs are identified as the residues with $P < 1e\text{-}4$ and conserved positions (CPs) as residues with conservation >75%. Remaining amino acids from the pocket list are labeled other positions (OPs).

2.3 ASMC validation on the benchmark

In order to evaluate the performances of ASMC, we set up different criteria:

Average distance: SDPs, CPs and OPs residues have been positioned onto the crystal structure bound to the ligand(s). For each category, the average distance between residues and the ligand(s) is computed as described in Kalinina *et al.* (2009). The average distance is defined as the sum of the minimal distances between residues and ligands divided by the number of residues. The minimal distance is the distance between the closest atoms of one residue and the ligand(s). As OPs residues are not supposed to be catalytic, functional or binding residues, it is expected to have longer distances for OPs than for CPS or SDPs residues.

Sensitivity (Se) and specificity (Sp): the set of residues in contact with ligand(s) has been extracted from the holo structure and are defined as positives residues (i.e. functional residues). The remaining residues are defined as negatives. We considered that a residue is in contact with the ligand when its minimal distance is smaller than 5 Å. Se is the ratio of true positives (CPs or SDPs) to true positives plus false negatives. Sp is the ratio of true negatives to true negatives plus false positives.

2.4 Comparison with multiple sequence alignments

In order to measure precisely the influence of homology modeling on SPDs detection, we performed a similar ASMC pipeline but changing the structural alignment step by a multiple sequence alignment for the 42 families of our benchmark. The multiple sequence alignment was performed using MAFFT (Katoh *et al.*, 2005) and the following steps (identification of residues of active site and classification) were applied as described in ASMC methodology.

3 RESULTS

3.1 ASMC performance

Sensitivity and specificity of ASMC: for each family, ASMC sensitivity and specificity was calculated and the average distances between SDPs + CPs and OPs with the ligands have been computed as described in Section 2 (Fig. 2). For some families, two results (family-1/-2) are presented when two ligands are bound to the enzyme in different pockets. Over the 47 studied pockets (42 families, 5 with two pockets), ASMC predicted 47% (Se > 47%) of the residues in contact with ligands. We observed significant difference between results obtained on uni-functional (Se = 72%, Sp = 37%) and multi-functional families (Se = 47%, Sp = 68%). As we compared our predictions to one holo structure bound to a unique ligand, we only used residues interacting with a single ligand. For the diverse dataset, we did not include potential residues binding different substrates. This bias introduced an underestimation of sensitivity/specificity and underlines that ASMC can perform better on multiactivities families than indicated in this study.

In families PF00108, PF01293 and PF02901, ASMC does not identify OP residues and seems to overpredict CP residues which are not in contact with the ligands. These residues may be important for the structural organization or stability of the pocket. However, a detailed analysis of these cases confirmed that all SDPs are in contact with the ligand. In other families (PF00293, PF00755, PF01135, PF01467 and PF03171), the observed weak sensitivity is explained by the weak sequence identity between members of the family (<30%). This reveals one limit of the method to analyze poorly conserved super-family not allowing reasonable structural alignments.

Comparison of clusters with EC classification: in order to test the relevance of ASMC clustering, we checked the coherence of clusters using experimentally proven EC numbers. Seventy percent of the clusters are in agreement with the EC number classification, each EC number segregating the clusters. In Supplementary Figure S1, we give an example of classification obtained with ASMC for PF02274 Pfam family. ASMC has separated two main clusters with very distinct activities (EC 2.1.4.; aminidotransferases and EC 3.5.3.6;

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

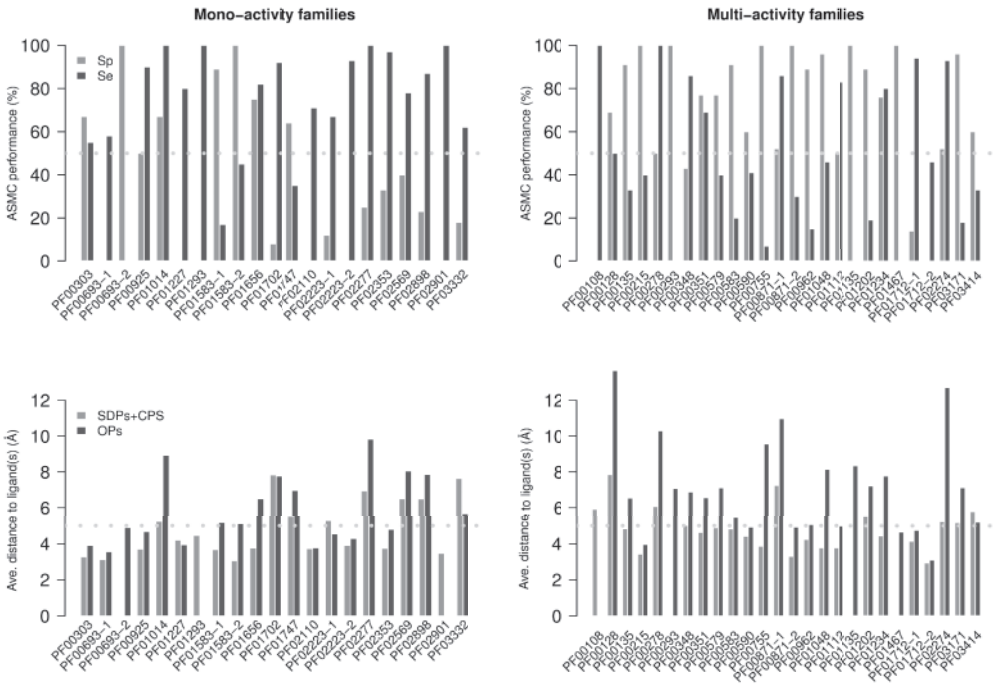

Fig. 2. ASMC results over mono- and multifunctional families. On top, ASMC performance is described in terms of Specificity and Sensitivity (see Section 2). Below, average distance to bound ligand(s) is compared between CPs + SDPs categories, predicted to be functional residues, and OPs categories, the remaining residues. For some families, two results (family-1/-2) are presented when two ligands are bound to the enzyme in different pockets.

arginine deiminase) while Pfam classification has grouped them. In the remaining 30%, enzymes can manifest infidelity of molecular recognition. Many enzymes can promiscuously catalyze reactions or act on substrates, other than those for which they evolved (Khersonsky and Tawfik, 2010). This could be the case for enzymes belonging to families PF00128, PF01112, PF01135 and PF01712. For PF00128, as underlined by Kalinina *et al.* (2009), the same enzymatic activity seems to have evolved independently on two separate branches of the phylogenetic tree. For PF01112, ASMC was not able to separate β-aspartyl-peptidase (EC 3.5.1.26) and aspartylglucosylaminase (EC 3.4.19.5) activities. For PF01135, ASMC has mixed *N*-acetylglucosamine-6-phosphate deacetylase (EC 3.5.1.25) and N4-(β-*N*-acetylglucosaminyl) - L - asparaginase (EC 3.5.1.26), which are very promiscuous reactions. In PF01712, ASMC has been able to separate pyrimidine and purine kinases, but was enable to discriminate guanosine to adenosine, and cytidine to thymidine specificities .

Influence of sequence similarity on ASMC: one of the factors that influences the quality of models predicted by homology modeling is the sequence identity between the target sequence to be modeled, and the reference one (associated to a structure). It has been shown that models obtained for a similarity greater that 30% have, on average, more than 60% of the backbone atoms correctly modeled and a root mean squared deviation (RMSD) less than 3.5 Å(Eswar *et al.*, 2008). The efficiency decreases below the 30% threshold, which defines the upper bond of the twilight zone. Supplementary Figure S2 presents the evolution of Se and Sp depending on sequence similarity obtained for the 42 benchmark families. It spotlights that detection of SDPs is more efficient with dataset composed of divergent sequences (increase of specificity). Also, this analysis underlines that ASMC is still efficient below the 30% threshold, a compromise to obtain good Sp and Se.

Average distance of CPs + SDPs to cognate ligand onto related family members crystal structure: for 85% of the cases, average distance between CPs + SDPs and ligands is smaller than distance between ODPs and ligands, showing that ASMC recovers most functional residues. These distances have been compared with results obtained by SDPsite (Kalinina *et al.*, 2009) for CPS + SDPs. For 80% of the cases, ASMC performed better (see Supplementary Material for detailed results on 42 families).

3.2 Alignment improvement

In this study, target sequences were aligned with template structures using the SALIGN module of Modeler (Madhusudhan *et al.*, 2009). Structural alignment differs from sequence alignment methods as it takes into account structural information from the template. This information is used to introduce a gap penalty function that tends to place gaps in exposed solvent and curved regions, outside secondary structure segments, and between stericaly inconsistent positions. The alignment error rate is reduced by 1/3 relatively to standard sequence alignment and this improvement is enhanced as the similarity between the sequences decreases (Eswar *et al.*, 2006). To optimize the alignment process (Chakravarty *et al.*, 2008), we computed 50 models, for each sequence, using all the structures available for a family (Eswar *et al.*, 2008).

The impact of alignment quality of ASMC over classical sequence global multi-alignment is illustrated in Figure 3. Assuming that the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

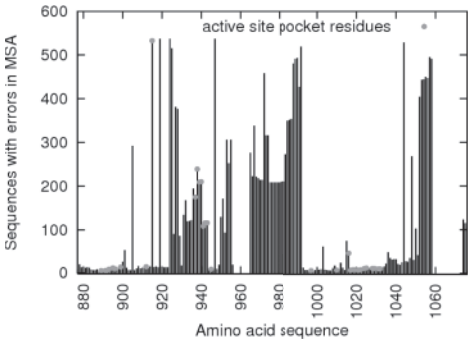

Fig. 3. Divergence in residue positions in sequence alignment (MSeqA) and structural alignment (MSA). A total of 2201 sequences from family Pfam PF00211 (nucleotidyl cyclases) were aligned using clustalW, and by the structural alignment used in ASMC. The curve presents the deviation average between the methods for all the proteins of the family. Dots show the positions of residues in the multiple sequence alignment that correspond to the active site residues according to Fpocket prediction (Le Guilloux *et al.*, 2009). For more clarity, they have been placed on top of the error bar. For instance, position 938 is mismatching in 200 sequences resulted of the MSeqA. This result highlights the impact of structural information on alignment quality, correcting cumulative sequence alignment errors due to insertion/deletion misplacements (*x*-axis: MSeqA positions; *y*-axis: number of sequences that present divergent alignment between the two alignments).

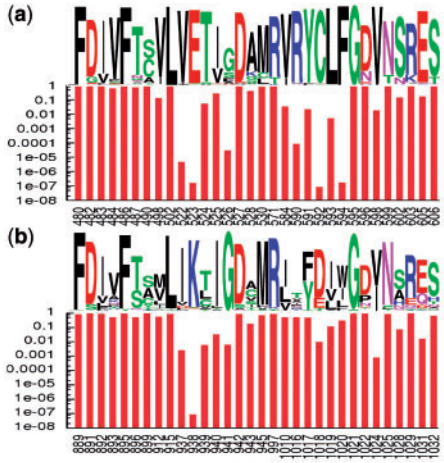

Fig. 4. Analysis of the active site pocket residues of (**a**) guanylate cyclases and (**b**) adenylate cyclases. Sequence logos were generated using WebLogo (Crooks *et al.*, 2004). For each position of the catalytic pocket 3D alignment (Section 2) supporting the classification method, representation of the observed frequency of residues is encoded by the height of the letter. The overall height of each stack is proportional to the sequence conservation at that position. The letters in each stack are ordered from the most to the least frequent so that one may read the consensus sequence from the tops of the stacks. *P*-values are computed as described in Section 2. Positions of amino acids in PDB IDs (a) 3ET6:A and (b) 1AB8:A.

structural alignment is ideal, we computed the deviation between residue positions in an MSeqA and the position obtained by structural alignment. We can note that the cumulative error rate in sequence-based comparison, along the sequence, is corrected by the structural alignment. The highest alignment quality for residues in the catalytic cavity illustrates the sensibility of sequence-based alignment to sequence conservation (residues in active site regions are usually more conserved). The curve presents the averages between all proteins of the family and shows errors in the MSeqA due to insertions and/or deletions. This result highlights the impact of structural information on sequence alignment quality.

Supplementary Table S4 compares Se and Sp obtained with ASMC and with a modified version procedure in which alignment has been replaced by multiple sequence alignment (MSeqA) (see Sections 2 and 2.4). ASMC performs better (Sp = 52.5%; Se = 59.5%) than the multiple sequence alignment (Sp = 41%; Se = 56%). Using MSeqA, sequences of families PF00693, PF01135 and PF01234 could not be clustered.

3.3 Nucleotidyl cyclases

Nucleotidyl cyclases are enzymes that catalyze the formation of cyclic nucleotide monophosphate from nucleotide triphosphate. The guanylate cyclase (GC) group catalyzes the formation of cGMP from GTP and the adenylate cyclase (AC) group converts ATP to cAMP. As shown by Tucker *et al.* (1998), the specificity of GC can be modified to AC by two amino acid substitutions of guanylate cyclase (PDB ID 3ET6:A): Glu523Lys and Cys592Asp. The Cys592Asp mutation abolishes guanidine binding by creating an electrostatic repulsion between the aspartate and the guanine O6. In the AC group, it stabilizes adenine binding by adding a hydrogen bond with N6. The Glu523Lys mutation creates a hydrogen bond with adenine. Other conserved amino acids have been described in these families: Arg590, Leu593 and Phe594 in GC and their counterparts in AC (Gln, Ile and Trp). Mutagenesis experiments could not identify the roles of these three amino acids in substrate specificity.

ASMC, applied to 536 sequences (Section 2), detects two clusters in the first tree ramifications: one is composed of 323 GC sequences and the other is composed of the 213 AC sequences (Fig. 4). It detects conserved residues (highest *P*-value) implied in activity or structure integrity: Phe480, Phe486, Leu502, Asp527 and Gly595. The positions known to be involved in specificity are detected as SDPs (lowest *P*-value): the Glu523/Lys938 position and the position Cys592 which is substituted by a negative residue at position 1018 in the AC group.

Results were compared with the TreeDet method that identifies six clusters. The main one is composed of 19% of the sequences (mainly but not only of GC), and the second one includes 12% of the sequences (mainly but not only AC). The remaining sequences (69%) are divided between four clusters. Eighteen determinant positions are predicted (Gln880, Tyr882, Ala890, Leu912, Ile919, Val934, Glu935, Ile937, Lys938, Val1009, Lys1014, Tyr1017, Asp1018, Ile1019, Trp1020, Val1024, Phe1074 and Val1075). Only one of the two experimentally validated residues (Asp1018) was detected with a poor conservation rate. In this example, ASMC performed better for subfamily division and SDP prediction. One should note that the TreeDet analysis combines both the MSeqA building (ClustalW slow/accurate alignment) and the method itself.

Results were also compared with the method proposed by Hannenhalli and Russell (2000). This method requires pre-definition of clusters and we use those predicted by ASMC as input to the program. We obtained seven SDPs covering the two known positions, the Ile937, Ile1019 and Trp1020 positions that are also involved with specificity. For the two additional predicted positions

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

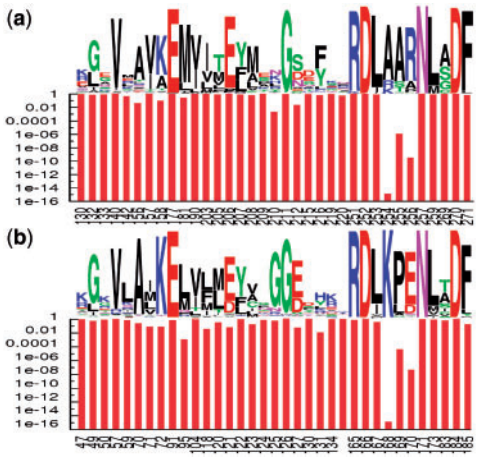

Fig. 5. Analysis of the active site pocket residues of (**a**) tyrosine kinases and (**b**) serine/threonine kinases. The ASMC analysis was performed on 3403 sequences from PFAM family PF00069 and PF007714 (see Section 2). Positions of residues refer to the sequence PDB 2CPK:E (serine/threonine kinase) and 1U46:A (tyrosine kinase).

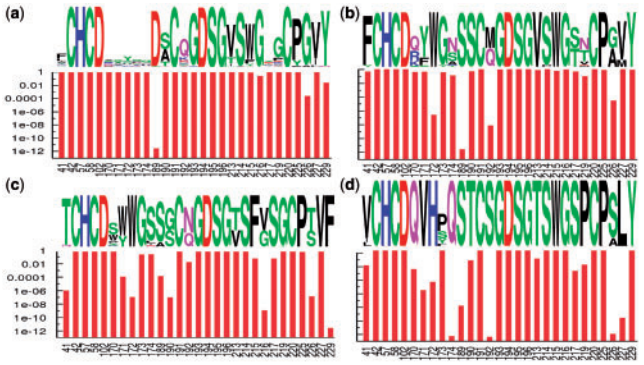

Fig. 6. Analysis of the active site pocket residues of (**a**) trypsins, (**b**) chymotrypsins, (**c**) elastases and (**d**) kallikreins.

(Ala890 and Lys1014), not belonging to the pocket cavity and hence not covered by ASMC, no experimental evidence or interpretation are available.

3.4 Protein kinases

Protein kinases are a family of enzymes that transfer phosphates from nucleotide triphosphates (usually ATP) to proteins (Hanks *et al.*, 1988). Two broad classes have been characterized with respect to substrate specificity: tyrosine kinases and serine/threonine kinases.

The two classes are characterized by two consensus sequences, RDLKPEN present in serine/threonine kinases and RDLAARN in tyrosine kinases. Figure 5 shows the active site pocket residue composition in each subfamily found by our method. The mentioned patterns are well identified and, according to the review presented in Hanks *et al.* (1988), predicted CPs (Lys72, Glu91, Arg165, Asp 166, Gln171, Asp184 and Phe185) are all involved with catalytic activities of both the two families. The three discriminating amino acids in patterns are the three residues scored with the lowest *P*-value: Ala254, Ala255 and Arg256 for the tyrosine kinases and Lys168, Pro169, Glu170 for the serine/threonine kinases. These residues are known to be involved in the differentiation of substrate specificity. Other residues have been described as determining specificity (Hannenhalli and Russell, 2000). The Thr201 and Tyr04 residues are close to the P+1 recognition loop identified by Hanks *et al.* (1988). In this study, we focused the analysis on residues of the active site pocket and we did not extend the analysis to the P+1 loop. A third group composed of 237 sequences, defined by a very well conserved pattern of residues, has also been identified by ASMC. Annotation of sequences identified them as epidermal growth factor receptors (EGFR). In our classification, it appears as subfamily of the tyrosine kinases.

3.5 Serine proteases

Serine proteases are proteolytic enzymes involving a catalytic triad composed of a nucleophile serine, an electrophile aspartate and an histidine as a base. Trypsins hydrolyze peptides harboring arginine or lysine at P1 position, chymotrypsins act on large hydrophobic residues at this position and elastases act on small aliphatic residues. This substrate specificity is associated with structural changes in the S1 binding pockets: The trypsin Asp189 residue accounts for the preference for positive residues. The hydrophobic property in chymotrypsin is due to a modification of the aspartate to a serine residue. In elastases, the pocket is occluded by Val216 and Thr226 (both positions are occupied by a conserved glycine residue in trypsins and in chymotrypsins), that accounts for the preference for small aliphatic residues. ASMC separates trypsin, chymotrypsin and elastase subfamilies (Fig. 6). The lowest *P*-value position in the trypsin subfamily is Asp189 that is involved in Arg/Lys tropism at P1 position. This single position fully explains the trypsin specificity when compared with the other two families. This position is changed to serine for chymotrypsins and elastases with low *P*-values. The other known catalytic positions 216 (glycine) and 226 (glycine) are coherently scored with high *P*-values in the trypsin subfamily as they are determinant for chymotrypsin and elastase activity. Val216 and Thr216 residues may be substituted by glycine or serine in the elastase group. All sequences of this group are annotated as elastase by SwissProt / UniProt and our method did not identify a subgroup specific for these variations; these observations suggest that these changes do not modify the elastase specificity.

The Tyr172Trp position, which presents a low *P*-values in chymotrypsins and also elastases, has been described as determinant in the trypsin/chimotrypsin conversion (Hedstrom *et al.*, 1994).The TreeDet (ClustalW) method predicts the following discriminating residues: Trp51, His91 and Leu155 which do not include the four found to be determinant (172, 189, 216 and 226). The method of Hannenhalli and Russell (2000) detects the positions 189 and 226 but not 216 and 172, discriminating between trypsins and chymotrypsins. Positions 121, 137 and 164 are also detected and we have no experimental data on these residues.

In addition to the predicted trypsin, chymotrypsin and elastase subfamilies, ASMC identifies a 13 sequence cluster associated with a conserved SDP pattern (Fig. 7). This cluster is composed of sequences annotated as kallikreins, for which we did not use 3D structure information for protein modeling. This result illustrates

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

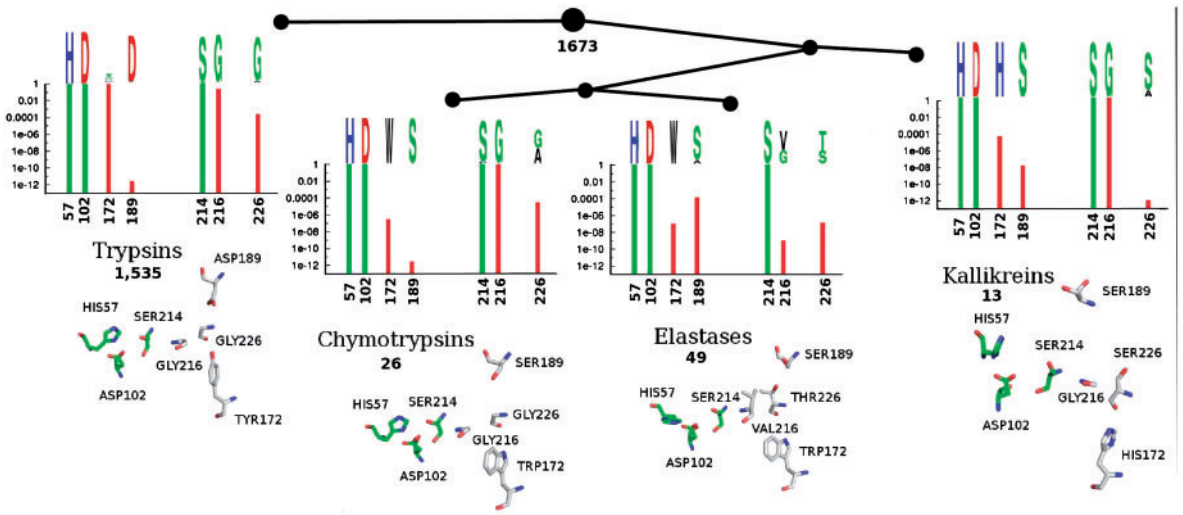

Fig. 7. Active site pattern tree for serine proteases. The 1673 initial sequences are split into four major clusters. Number of proteins belonging to each cluster is indicated. Under each pattern is represented the 3D active site of one representative where the conserved catalytic triad (His, Asp and Ser) is colored in green.

that ASMC is able to predict new subfamilies without the support of a 3D structure belonging to the subfamily.

4 CONCLUSION

While quantities of sequence and structural data continue to grow, only a few methods are using these two kinds of information to annotate proteins at a high-throughput rate. Our method, ASMC, proposes a new approach to the annotation of enzymes with a high level of precision. Indeed, prediction of SDPs in active site provides information for understanding protein functions and evolution. The main advantage of ASMC over existing methods is the structural alignment of pockets that provides a high-quality comparison of residues located in the active site. Comparison of the method with a multiple sequence alignment-based methods demonstrated the limitation of using only the sequence information, due to the low quality of sequence alignments. The conceptual clustering step separates the subfamilies and determines the residues responsible for the specificity. The methodology has been validated with families of known functions and predicted SDPs are well confirmed by experimental data.

We identified several applications of patterns proposed by ASMC. CPs and SDPs positions outline residues of active site pockets that are under selective constraint, conserved or correlated to subfamily differentiation. These residues, used as a geometric pattern, can be used for screening structure databases to identify new candidates for specific functions. ASMC can also be used to classify new sequences, belonging to families with at least one structure available, in order to annotate them into functional subfamilies.

Although we presented results for known protein families, the method can provide valuable insights in the study of families with unknown functions and can serve as an input to elucidate new enzymatic activities. This is a real challenge as 3000 above 11 912 families of the database of sequences families (Pfam-A) are referred to domains of unknown functions (DUFs). Among them, about 250 families have at least one structure deposited in PDB. ASMC can be used by screening and searching for homologous patterns in an active site profile database associated to an enzymatic activity. Another interesting perspective of ASMC patterns is the improvement of *de novo* function prediction through virtual screening approaches. SDPs and CPs can be used to improve selection of plausible configurations of substrates in the active sites by pinpointing important substrate/enzyme interactions.

ACKNOWLEDGEMENTS

We would like to thank Marcel Salanoubat, Alain Perret and Anne Zaparucha for comments and insightful discussions. We also thank Jean Weissenbach for his continuous support.

Conflict of Interest: none declared.

REFERENCES

Capra,J.A. and Singh,M. (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, **24**, 1473–1480.

Chakrabarti,S. and Panchenko,A.R. (2009) Coevolution in defining the functional specificity. *Proteins*, **75**, 231–240.

Chakrabarti,S. *et al.* (2007) Functional specificity lies within the properties and evolutionary changes of amino acids. *J. Mol. Biol.*, **373**, 801–810.

Chakravarty,S. *et al.* (2008) Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. *BMC Struct. Biol.*, **8**, 31.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Donaldo,J.E. and Shakhnovich,E.I. (2005) Determining functional specificity from protein sequences. *Bioinformatics*, **21**, 2629–2635.

Donald,J.E. and Shakhnovich,E.I. (2009) SDR: a database of predicted specificity-determining residues in proteins. *Nucleic Acids Res.*, **37**, D191–D194.

Eswar,N. *et al.* (2006) Comparative protein structure modeling using modeller. *Curr. Protoc. Bioinformatics*, **Chapter 5**, Unit 5.6.

Eswar,N. *et al.* (2008) Protein structure modelling with Modeller. *Methods Mol. Biol.*, **426**, 145–159.

Finn,R.D. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

Fisher,D. (1987) Knowledge acquisition via incremental conceptual clustering. *Mach. Learn.*, **2**, 139–172.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

Goldenberg,O. *et al.* (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.*, **37**, D323–D327.

Gong,S. and Blundell,T.L. (2008) Discarding functional residues from the substitution table improves prediction of active sites within three-dimensional structures. *PLoS Comput. Biol.*, **4**, e1000179.

Halabi,N. *et al.* (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell*, **138**, 774–786.

Hanks,S.K. *et al.* (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science*, **241**, 42–52.

Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.

Hedstrom,L. *et al.* (1994) Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. *Biochemistry*, **33**, 8757–8763.

Henschel,A. *et al.* (2007) Using structural motif descriptors for sequence-based binding site prediction. *BMC Bioinformatics*, **8** (Suppl. 4), S5.

Holmes,G. *et al.* (1994) Weka: a machine learning workbench. In *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*. Brisbane, Australia, pp. 357–361.

Kalinina,O.V. (2004) SDPred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleid Acids Res.*, **32**, W424–W428.

Kalinina,O.V. *et al.* (2009) Combining specificity determining and conserved residues improves functional site prediction.*BMC Bioinformatics*, **10**, 174.

Katoh,K. *et al.* (2005) Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

Khersonsky,O. and Tawfik,D. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.*, **79**, 471–505.

Kristensen,D.M. *et al.* (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics*, **9**, 17.

Langraf,R. *et al.* (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.

Le Guilloux,V. *et al*. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, **10**, 168.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Madabushi,S. *et al.* (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.

Madhusudhan,M. *et al.* (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Sel.*, **22**, 569–574.

Moult,J. (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.*, **15**, 285–289.

Nagao,C. *et al.* (2010) Relationship between functional subclasses and information contained in active-site and ligand-binding residues in diverse superfamilies. *Proteins*, **78**, 2369–2384.

Najmanovich,R. *et al.* (2008) Detection of 3d atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, **26**, i105–i111.

Orengo,C. *et al.* (1997) CATH: a hierarchic database of protein domain structures. *Structure*, **5**, 1093–1108.

Pazos,F. *et al.* (2006) Phylogeny-independent detection of functional residues. *Bioinformatics*, **22**, 1440–1448.

Pei,J. *et al.* (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, **22**, 164–171.

Pupko,T. *et al.* (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants with their homologues. *Bioinformatics*, **18** (Suppl. 1), S71–S77.

Rausell,A. *et al.* (2010) Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl Acad. Sci. USA*, **107**, 1995–2000.

Redfern,O.C. *et al.* (2009) FLORA: a novel method to predict protein function from structure diverse superfamilies. *PLoS Comput. Biol.*, **5**, e1000485.

Rottig,M. *et al.* (2010) Combining structure and sequence information allows automated prediction of substrate specificities within enzyme families. *PLoS Comput. Biol.*, **6**, e1000636.

Shatsky,M. *et al.* (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–156.

Sol,A.D. *et al.* (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.

Sonnhammer,E. *et al.* (1997) Pfam: a comprehensive database of protein families based on seed alignments. *Proteins*, **28**, 405–420.

Tramontano,A. and Morea,V. (2003) Assessment of homology-based predictions in CASP5. *Proteins*, **53** (Suppl. 6), 652–368.

Tseng,Y.Y. *et al.* (2009) Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *J. Mol. Biol.*, **387**, 451–464.

Tucker,C.L. *et al.* (1998) Two amino acid substitutions convert a guanylyl cyclase, RetGC-1 into and adenylyl cyclase. *Proc. Natl Acad. Sci. USA*, **98**, 5993–5997.

Ward,R.M. *et al.* (2009) Evolutionary trace annotation server: automated enzyme function prediction in protein structures with 3D templates. *Bioinformatics*, **25**, 1426–1427.

Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.

Yu,G-X. *et al.* (2005) In silico discovery of enzyme-substrate specificity-determining residue clusters. *J. Mol. Biol.*, **352**, 1105–1117.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016