An Entropy Based Heuristic Model for Predicting Functional Sub-type Divisions of Protein Families

Deniz Yorukoglu
Sabanci University
Computational Biology Lab
FENS, Tuzla 34956 Istanbul, Turkey
(+90)-216-5770059
denizy@su.sabanciuniv.edu

Yasin Bakis
Abant Izzet Baysal University
Department of Biology
Golkoy Campus, Bolu, Turkey
(+90)-374-2541000-(2746)
bakis_y@ibu.edu.tr

Ugur Sezerman
Sabanci University
Computational Biology Lab
FENS, Tuzla 34956 Istanbul, Turkey
(+90)-216-4839513
ugur@sabanciuniv.edu

ABSTRACT
Multiple sequence alignments of protein families are often used for locating residues that are widely apart in the sequence, which are considered as influential for determining functional specificity of proteins towards various substrates, ligands, DNA and other proteins.

In this paper, we propose an entropy-score based heuristic algorithm model for predicting functional sub-family divisions of protein families, given the multiple sequence alignment of the protein family as input without any functional sub-type or key site information given for any protein sequence.

Two of the experimented test-cases are reported in this paper. First test-case is Nucleotidyl Cyclase protein family consisting of guanalyate and adenylate cyclases. And the second test-case is a dataset of proteins taken from six superfamilies in Structure-Function Linkage Database (SFLD). Results from these test-cases are reported in terms of confirmed sub-type divisions with phylogeny relations from former studies in the literature.

Categories and Subject Descriptors
I.5.2 [**Computing Methodologies**]: Pattern Recognition – Design Methodology – Classifier Design and Evaluation

J.3 [**Computer Application**]: Life and Medical Sciences – Biology and Genetics

General Terms
Algorithms, Design, Experimentation, Verification.

Keywords
Protein Function, Classification, Multiple Sequence Alignment

1. INTRODUCTION
Determining functionality of proteins and classification of protein sub-types due to their functional specificity is one of the major objectives of current researches in molecular biology. Such a classification of proteins would trigger experiments upon protein redesign and functional analysis of proteins, providing a more extensive understanding of the nature of protein's functional specificity [1].

Functional specificity of proteins represents their selective behavior, for being very specific to the reaction they catalyze and their corresponding substrates. The factors that are responsible for this specificity are generally the complementary shape of the protein and the substrate, their charges, and hydrophobic-hydrophilic properties. Most of the information in these factors can be reached by a detailed analysis of their sequential and structural properties. For this reason, in the literature, there is a variety of algorithms that exploit these types of sequential and structural information; in order to find motifs that correspond to functional specificity determining positions in protein families. Some of these methods will be described in more detail in the Background section.

In this study, the main aim of the devised algorithm is to exploit sequential information of proteins, the multiple sequence alignment of the protein family; in order to find a suitable division of the family proteins into functional sub-groups, together with the key positions in the sequence alignment that are highly correlated to this sub-group division profile. A detailed explanation of this algorithm and its supplementary methods, such as amino acid group labeling and entropy score calculations, will be provided in the Methodology section.

2. BACKGROUND
In the literature, there are various algorithms that exploit different types of structural and sequence motifs. Some of these methods use the information from the functional sites originating from the protein structure and/or acquire information from structural alignment of proteins [2-3]. Whereas, some algorithms only process the sequence data without needing any additional structural information: SDPpred [1] [4] and QuasiMotiFinder [5].

In recent years, there were further novel approaches for finding specificity determining positions and consequently determining functional sub-types of a protein family. One of these studies is carried out by Wallace & Higgins in 2007 [6]. In this work they proposed a statistical approach using Between Group Analysis method and further utilizing principle component analysis and correspondence analysis on Lactate/Malate dehydrogenase, Nucleotidyl Cyclase and Serine Protease data, in order to identify specificity determining positions. Nucleotidyl

GECCO'09, July 8–12, 2009, Montreal, QC, Canada.

Cyclase protein family results of the heuristic method described in this paper are shown in comparison to the results in [6] as well.

Furthermore, an efficient and powerful entropy method is proposed by Hannenhalli S. & Russell R. in 2000 [10]. In this study, they have utilized a relative entropy calculation method in order to determine significant positions in a protein family sequence alignment, and a similarity method for determining functional sub-types in the protein family using this identified positions. The entropy method, as in both relative entropy and cumulative relative entropy score calculations [22-23], is also adopted in the algorithm presented in this paper. Details will be further explained in Supplementary Methods sub-section.

Even though the methods described till this point utilize structural and sequential motif information in order to find functional sub-groups; there are also a variety of algorithms in the literature that tries to achieve a direct sub-family identification without any initial sub-type and functionality definition. These types of de novo functional group identification algorithms in the literature mainly fall into two categories: methods that build-up functional groups by comparing pair wise similarities [12-16]; and methods that define clusters by cutting an initially calculated hierarchical or phylogenetic tree into functional subgroups [17-19] including SCI-PHY method developed by Sjölander [20].

SCI-PHY algorithm initially constructs a hierarchical tree using Dirichlet mixture density [21] profiles of each sequences and employing a bottom-up pair-wise joining method using relative entropy scores [7] as the distance measure. The conversion of original amino acid sequences to Dirichlet mixture density profiles in SCI-PHY are done in order to increase sensitivity without causing any decrease in the specificity. A similar but a more discrete approach is taken in our algorithm as well, not in terms of creating statistical profiles, but determining amino acid grouping labels for entropy calculations. A more detailed explanation of this method is provided in the Supplementary Methods sub-section.

3. METHODOLOGY

3.1 Algorithm

The flow of the designed heuristic algorithm is described as follows:

a) The input is read consisting of initially unlabelled protein names and their aligned sequences.
b) A random division is applied to these protein families representing the first sub-family division. Note that; in this way, the method requires no specific initial division to be provided from input.
c) Division fitness score is calculated for this initial division
d) The division fitness score change is observed when each of the proteins is shifted to the other sub-family. The move with the highest division fitness score is selected.
e) This selected move is performed if the move fitness check yields a positive result, and the procedure is repeated back from (d) with this new sub-family division as the initial division.
f) If the highest score is not successful enough for performing a new move (move fitness check fails), then this division is accepted as the best 2-way sub-family division.
g) A sub-family label is selected to be divided according to Diverse Sub-type Selection method. In this selected subfamily, again some of the proteins are randomly selected to be the third subfamily, which constitutes the initial three-way division.
h) Continue to the algorithm with N=3.

Until here a two-way division heuristic was employed. From this point on, a multiple-way division scheme is explained, in which the symbol N represents the number of sub-families to be divided.

i) The protein sub-type division score is calculated for N-way sub-family division.
j) The fitness score changes are observed for each of the proteins, when their sub-type codes are changed to other sub-types. The move with the highest score is selected.
k) This selected move is performed if the move fitness check yields a positive result, and the procedure is repeated back from (j) with this new sub-family division as the initial division.
l) If the highest score is not successful enough for performing a move, then this division is accepted as the best N-way sub-family division.
m) A check is done in order to determine if this N-way division is better than the division with N-1 sub-families. If this check shows that new division is less successful than the original division with less number of sub-types, the algorithm is stopped. The output is reported as the original sub-family division with N-1 subtypes. Otherwise the algorithm continues with (n).
n) A sub-family label is selected to be divided according to Diverse Sub-type Selection method. Some of the proteins in this sub-family are randomly selected to be the $N+1^{th}$ subfamily, which constitutes the initial N+1 way division.
o) The algorithm is continued back from (i) with new N = N + 1.

The division fitness score may be variable from case to case. The types used until now in this project are either the average entropy score for all positions or the average entropy score of a small percentage or a specific number of positions. With more experimental results from different protein family datasets, this amount of significant positions can be optimized for better score representation. Additionally the N-way division fitness scores are calculated in the same way as 2-way divisions, due to the applicability of cumulative relative entropy score calculation over multiple sub-type definitions. Experimental results showed that selecting %10 of the positions as significant positions yielded successful division frequency, eliminating most of the position that are unrepresentative for the division and including more than minimum number of positions that are necessary for representing the division in a combinatory way.

Furthermore, the move fitness check is another part of this algorithm that may be changeable in order to get optimum convergence in the solution space. The method used in this study is quite simplistic: the move fitness check returns a positive result if the highest scoring move have a higher fitness score than the current sub-family division. However, incorporating a heat

function into this fitness check could allow searching further if a local optimum is reached. Of course, in this case there might be some modification to the algorithm scheme, since storing the highest fitness score reached and restoring it when getting stuck at a higher local minimum might be necessary.

An important aspect of the multiple subtype division search is that, the borders defined within the former best division method with less sub-families can be updated. Since through the search, the algorithm looks at all possible sub-type switches for any protein to any subtype, the sub-families that are not subjected to random division may be changed as well.

Finally the diverse sub-family selection after an optimum division is reached for a number of sub-families, is also defined as a parametric function; since different branching selection may infer further separate division schemes. The diverse sub-family selection used in our algorithm is a measure that is combined by size of the subtype and average relative entropy score of positions within the subtype.

However, it should be noted that, even though a sub-family is selected into two separate groups for the next iteration for n+1 sub-types; since the borders are changeable as explained earlier, the sub-family branching may converge to a different division state as if the $n+1^{th}$ sub-type has branched out of a completely another sub-family. This process was observed in several cases through the convergence of Nucleotidyl Cyclases dataset runs.

3.2 Supplementary Methods

3.2.1 Amino Acid Group Labels

In the algorithm, when calculating cumulative relative entropy scores of positions, entropy calculations are not using 20 different amino acid codes, but calculated according to their conserved groups.

These amino acid groups are determined due to physicochemical properties of amino acids. Amino acid group definitions are as follows:

Label 1: I (Isoleuc.), V (Valine), L (Leucine), M (Methion.)
Label 2: Q (Glutamine), N (Asparegine)
Label 3: S (Serine), T (Threonine)
Label 4: R (Arginine), K (Lysine)
Label 5: F (Phenylalanine), Y (Tyrosine)
Label 6: D (Aspartic Acid), E (Glutamic Acid)
Label 7: A (Alanine)
Label 8: H (Histidine)
Label 9: W (Tryptophan)
Label 10: G (Glycine)
Label 11: C (Cysteine)
Label 12: P (Proline)

3.2.2 Relative Entropy Score

Calculating relative entropy scores given the sub-family division of a protein family is a method for analyzing and detecting positions that are highly correlated to the division. A single pass of cumulative relative entropy score calculations over the positions in the alignment, finds positions that are similar within a sub-type and different between separate sub-types [7].

a) $\sum_{x \in \text{a.a. labels}} p_{i,x} = 1$ b) $RE_i^A = \sum_{x \in \text{a.a. labels}} p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^{A^*}}$

c) $CRE^i = \sum_{\text{all subtypes}} RE^i$

Figure 1: Formulas and equations for entropy score calculations.

In the figure above, (a) shows the equation for the probability sum of amino acid group labels, such that; *i* representing a position and *x* representing the a.a. group labels, the sum of probability of all labels for a position is 1. In order to satisfy this, each appearance of group labels are counted, background frequencies are defined as 0.1 if a label doesn't appear in the position, and finally the sum is normalized to 1 for the position. (b) shows the formula of the relative entropy score calculation for position *i* and subtype *A*. In the right hand side of the formula, *A** represents all subtypes other than *A*. Finally, (c) defines the cumulative relative entropy calculation for a position, which is the sum of all relative entropy scores for different subtypes in a position.

Entropy score calculations used in the algorithm and formulas described in Figure 1 are adopted from their use in the paper by Hannenhalli & Russell [7]. A more detailed explanation about the idea behind and application of entropy score calculations in functional sub-family prediction can be found in their paper and deeper theoretical information behind entropy methods can be found in [22-23].

4. RESULTS

A computer application was developed in order to test the sub-type division of different protein family test-sets. Results shown in this paper are for Clustal W [8] multiple sequence alignments of Nucleotidyl Cyclase protein family and a dataset of protein sequences taken from six different protein families in Structure-Function Linkage Database (SFLD) [9].

4.1 Nucleotidyl Cyclases

Nucleotidyl cyclases are a membrane attached cytosolic protein family which catalyses the reaction of nucleotide triphosphate forming into a cyclic nucleotide monophosphate. These cyclases act on either guanalyate cyclases (GTP) or adenylate cyclases (ATP). In 1998, Tucker et al. has shown that residue substitutions in two different positions are sufficient to convert a guanylyl cyclase into an adenylyl cyclase, changing the enzyme specificity from GTPs towards ATPs [10].

Multiple sequence alignment of Nucleotidyl Cyclases used in this study is the same data-set used by Hannenhalli and Russell in [7] and Wallace and Higgins in [6] with 41 adenylate and 29 guanylate sequences.

In Figure 1, results of our entropy based heuristic model for two way and three way sub-family divisions of Nucleotidyl Cyclases are shown juxtaposed to the phylogenetic tree presented in [6] obtained using Neighbor-Joining method for reconstructing phylogenetic trees by Saitou and Nei [11].

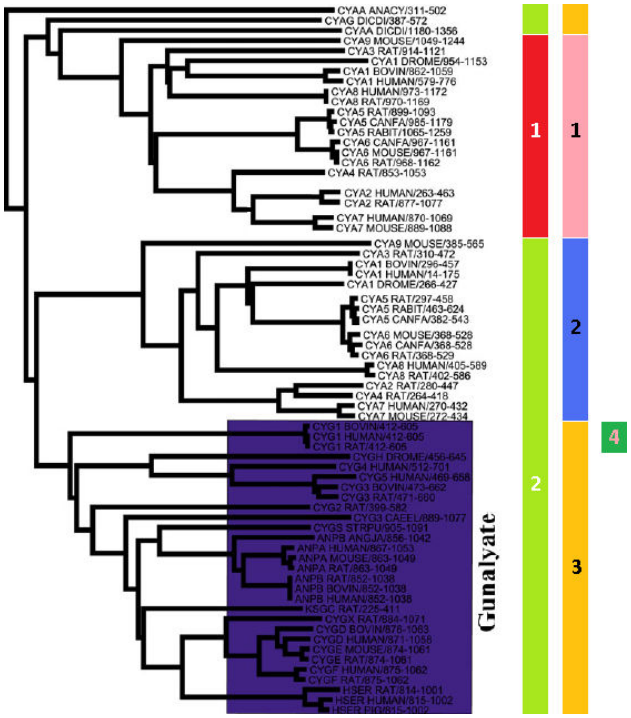

Figure 2 – Sub-type division results of the algorithm for Nucleotidyl Cyclases protein family shown in comparison to the phylogenetic tree of Nucleotidyl Cyclases shown by Wallace & Higgins [6].

In the figure above, the purple rectangle covering one of the sub-branches represents guanalyate cyclases, where as others represent adenylate cyclases. The red & green division column on the right corresponds to the two-way division of the family, where as pink, blue & orange division corresponds to the three-way division that is found by our heuristic method. The little dark green area represents the possible fourth sub-group if the algorithm would continue to 4-way division.

The division results shown in the figure above are maximum scoring results of 250 runs of our algorithm. This score level was reached in the %38.8 of the runs for 2-way division and %25 of the runs when 3-way division was employed. However, it should be noted that these highest frequency results, which are shown in the figure, are also the highest scoring division results found by this method.

As seen in the figure, the 2-way and 3-way divisions are highly parallel to the phylogenetic tree of the protein family. Other than some outcast protein sequences such as, Cyaa Anacy, Cyag Dicdi and Cyaa Dicdi at the top, the resulting divisions have one to one correspondence to the main 2-way and 3-way branching of the phylogenetic tree.

Since the algorithm requires n+1 way division as well in order to determine N-way as the ultimate functional division; when the application is run with unsupervised division sub-family quantity, it also calculates best scoring 4-way division for the Nucleotidyl Cyclases data. However, in this case the highest scoring 4-way division is a minor modification of 3-way division with the identification of Cyg1 Bovin, Cyg1 Human and Cyg1 Rat proteins as the fourth sub-family. As it can be seen from the figure, this selection is quite meaningful since these are the sequences are in the highest level branching within the Gunalyate protein sub-group. However, this result was found only in %7.2 of the results, even though it is both highest frequency result and highest scoring 4-way division found by the algorithm. For this reason, this division doesn't constitute a better answer than the previously calculated 3-way division, which means that the algorithm gives the 3-way division as the final result.

4.2 Structure-Function Linkage Database

Structure-function Linkage Database (SFLD) is a database linking evolutionarily related protein sequences and structures from six different super families to the functionality of these proteins, in other words the chemical reactions that these enzymes catalyze [9].

The superfamilies that are included in SFLD are amidohydrolase, crotonase, enolase, haloacid dehalogenase, terpene cyclase, and vicinal oxygen chelate protein families.

Initially we have tested our heuristic algorithm for three-way division between ten protein sequences taken from each of crotonase, enolase and amidohydrolase protein families. The algorithm was run a hundred times for a statistical analysis. Our method was able to differentiate between these two protein families with %100 success for %29 of the iterations.

Furthermore, we included protein sequences from all six protein superfamilies corresponding to an input with the multiple sequence alignment of 60 protein sequences. We ran our algorithm for a six-way division over these sequences. Out of a hundred iterations, the highest frequency result (%30) and also the highest scoring result was an almost exact 6-way division as the input families, apart from a single haloacid dehalogenase wrongly classified as a protein from the enolase super family. However this protein was misclassified as among enolase protein family for %85 of the iterations, which might infer that this protein has a strong common functionality with the proteins selected from enolase family.

5. REFERENCES

[1] Kalinina,O.V., Novichkov,P.S., Mironov,A.A., Gelfand,M.S., Rakhmaninova,A.B. (2004). SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Research,* 32, 424-428

[2] Johnson, J.M. & Church, G.M. (2000). Predicting ligand-binding function in families of bacterial receptors. *Proc. Natl Acad. Sci*. USA, 97, 3965–3970.

[3] Lichtarge,O., Bourne,H,R. and Cohen,F,E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257, 342–358.

[4] Kalinina,O.V., Mironov,A.A., Gelfand,M.S., Rakhmaninova,A.B. (2004). Automated selection of Positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, 13, 443-456.

[5] Gutman, R., Berezin, C., Wollman, R., Rosenberg, Y., Ben-Tal, N. (2005). QuasiMotiFinder: protein annotation by

searching for evolutionarily conserved motif-like patterns. *Nucleic Acids Research,* 33, 255-261.

[6] Wallace I., Higgins D. (2007). Supervised Multivariate analysis of sequence groups to identify specificity determining residues, *BMC Bioinformatics, 8:135*.

[7] Hannenhalli S., Russell R., (2000). Analysis and prediction of functional subtypes from protein sequence alignments, *Journal of Molecular Biology,* 303, 61-67.

[8] Thompson, J.D., Higgins, D.G., Gibson, T. J., (1994). 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(22):4673-4680

[9] Pegg, S.C., Brown, S.D., Ojha, S., et al. (2006). Leveraging enzyme structure-function relationships for functional inference and experimental design: The Structure-Function Linkage Database. *Biochemistry*, 45: 2545-2555.

[10] Tucker, C.L., Hurley, J.H., Miller, T.R. & Hurley, J.B. (1998). Two amino acid substitutions convert a guanylyl cyclase, RetGC-1, into an adenylyl cyclase. *Proc. Natl Acad. Sci. USA*, 11, 5994-5997.

[11] Saitou, N., Nei, M., (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.*, 4(4): 406-425.

[12] Remm, M., Storm, C.E., Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314: 1041-1052.

[13] Li, L., Stoeckert, C.J. Jr., Roos, D.S., (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13: 2178-2189.

[14] Abascal, F., Valencia, A., (2002). Clustering of proximal sequence space for the identification of protein families. *Bioinformatics*, 18: 908-921.

[15] Li, W., Jaroszewski, L., Godzik, A., (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, 18: 77-82.

[16] Li, W., Godzik, A., (2006). CD-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22: 1658-1659.

[17] Zmasek, C.M., Eddy, S.R., (2002) RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3:14.

[18] Storm, C.E., Sonnhammer, E.L., (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliablility. *Bioinformatics*, 18: 92-99.

[19] Wicker, N., Perrin, G.R., Thierry, J.C., Poch, O., (2001). Secator: A program for inferring protein subfamilies from phylogenetic trees. *Mol. Bİol. Evol.*, 18: 1435-1441.

[20] Sjölander, K., (1998). Phylogenetic inference in protein superfamilies: Analysis of SH2 domains. In: Proceedings of the Sixth International Conference on Intelligent Systems in Molecular Biology; 28 June - 1 July, 1998; Montreal, Quebec, Canada. Pp. 165-174.

[21] Sjölander, K., Karplus, K., Brown, M., et al. (1996) Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci..*, 12: 327-347.

[22] Shannon, C. & Weaver, W. (1963). *Mathematical Theory of Communication*, University of Illinois press, Champaign, IL.

[23] Durbin, R., Eddy, S., Krogh, A., Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids*, Cambridge University Press, Cambridge, UK.