A Robust Method to Detect Structural and Functional Remote Homologues

Ori Shachar[1] and Michal Linial[2]*

[1]*School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel*
[2]*Department of Biological Chemistry, Institute of Life Sciences, Hebrew University, Jerusalem, Israel*

ABSTRACT With currently available sequence data, it is feasible to conduct extensive comparisons among large sets of protein sequences. It is still a much more challenging task to partition the protein space into structurally and functionally related families solely based on sequence comparisons. The ProtoNet system automatically generates a treelike classification of the whole protein space. It stands to reason that this classification reflects evolutionary relationships, both close and remote. In this article, we examine this hypothesis. We present a semiautomatic procedure that singles out certain inner nodes in the ProtoNet tree that should ideally correspond to structurally and functionally defined protein families. We compare the performance of this method against several expert systems. Some of the competing methods incorporate additional extraneous information on protein structure or on enzymatic activities. The ProtoNet-based method performs at least as well as any of the methods with which it was compared. This article illustrates the ProtoNet-based method on several evolutionarily diverse families. Using this new method, an evolutionary divergence scheme can be proposed for a large number of structural and functional related superfamilies. Proteins 2004;57:531–538. © 2004 Wiley-Liss, Inc.

Key words: protein family; hierarchical classification; protein space; database

INTRODUCTION

The gap between the amount of sequences available and the number of known homologous families is still growing.[1,2] This article addresses the problem of finding remote homologues in the sequence space and suggesting for any given protein sequence its relationship with other proteins, which basically form a structural or a functional homologous family.

To fully appreciate the complexity of the problem, one needs to imagine the whole sequence space of biologically functional sequences, where the pairwise distance between every pair of sequences is exactly the pairwise score given by some chosen method for sequence similarity (i.e., FASTA or BLAST, with any chosen substitution matrix[3]). Our task of defining functional families would sum up to drawing an outline around all proteins that belong to the same functional family, thus separating them from all sequences of other families. For single-domain proteins

that have evolved by divergent evolution, such a description is valid, as in the case of the globin and histone families. However, for multidomain proteins that may have resulted from the fusion of domains, and for those that have evolved by convergent evolution,[4] such a simplified partition of the protein sequence space is incompatible. Such groups of proteins that are related by convergence in evolution or that are linked through shared domains are abundant in the protein space. They are especially abundant in complex proteomes. In reality, methods for classifications of such cases into functional or structural families are not reliable.

Most proteins, especially those derived from newly sequenced genomes, are not yet annotated, and their evolutionary relations to other proteins are not evident. As a result, the task of expanding and generalizing the presently available classification schemes for new sequences is difficult and may result in a reduction of the quality of functional inference.

The basic principle for classifying a new sequence is to test its identity and similarity to all other sequences in the database. Should it have a sequence identity of at least 30% with another sequence (throughout most of its length), a structural or a functional inference becomes valid. More accurately, functional annotation such as catalytic function of an enzyme requires at least 40% identity in amino acids,[5] while reliable structural inference can be based on 25–30% sequence identity over at least 100 amino acids.[6] Below these levels of pairwise sequence similarity, no unsupervised search method can suggest, with a reasonable confidence, the structural or functional relatedness to the closest homologue. While methods based on BLAST-based iterative profiles, Hidden Markov Models, and methods for finding connections through intermediate sequences are superior in their sensitivity, they suffer from a higher level of false positives.[1] Specifically, PSI-BLAST, which is among the best search methods,[7] misses ~50% of

all homologies when fixing the rate of false positives to 1/1000.[8] In the case that the sequence similarity is below the level of secure inference (often referred to as the "twilight zone"), finding the protein family borders and the evolutionary remote homologues of a given protein is desirable. One should bear in mind that naive sequence-based search methods that are based on similarity scoring are often inappropriate for multidomain proteins and for proteins that are a result of convergent evolution.

The inherent limitation of detecting remote homologues was estimated based on a structural benchmark.[9] Several methodologies for navigating in the protein space for a better detection of remote homologues were proposed. These include sequence hopping,[10] transitivity of homology,[11] intermediate sequences,[1] tuned iterative–profile,[12] profile–profile,[13] jumping alignments,[14] string-based kernels,[15] and some hybrid methods.[16]

Herein we suggest that a construction of the protein sequence space into a connected graph that captures the trace of evolutionary relatedness is useful in finding remote homologues. This method achieves satisfactory results that are comparable with results that rely on additional structural and functional knowledge. We suggest that the scaffold of the protein space as presented by ProtoNet[17] can be used to infer structural and functional information for remote homologues with high reliability. The success in inferring structural and functional relatedness for selected families by the ProtoNet navigation method is discussed.

METHODOLOGY
Methods for Detecting Remote Homologues

A comparative study that tested the success in detecting remote homologues by various published strategies was published.[18] The list of protein homologues was manually selected to represent both structurally and functionally (i.e., similar enzymatic activities) related families of remote homology. The proteins in the test set were characterized by their very low pairwise sequence similarity.

For each homologous family, multiple strategies were applied and combined with the goal of unifying the proteins that belong to the same evolutionary-related group. The strategies that were used included advanced sequence-based search methods, including FASTA walk,[19] PSI-BLAST,[20] HMMer,[21] MAST,[22] and PROBE.[23] Each of the methods initiated a search from a query protein from the group of proteins, and the number of proteins from the group that were identified by the search as remote homologues was noted (Table 1). We used this set as a benchmark for testing the ability of the ProtoNet cluster-map of the protein space to perform such a unification of protein families. We evaluated and quantified the degree of specificity and sensitivity in detecting those families.

ProtoNet Graph

We base our study on the database of ProtoNet 2.4 (http://www.protonet.cs.huji.ac.il).[17] ProtoNet is an automatically generated, hierarchical classification of all protein sequences from Swissprot (release 40.28 with 114,033

TABLE I. A Global Analysis for the Correspondence of ProtoNet Clusters with Structural and Functional Properties

Properties	# Best Clusters	CS[a]	Specificity[b]	Sensitivity[c]
SCOP Family				
All	1283	0.84	0.96	0.88
20–1000	968	0.86	0.98	0.88
SCOP Superfamily				
All	854	0.77	0.95	0.79
20–1000	607	0.78	0.97	0.80
Enzyme EC. 4th digit				
All	2157	0.75	0.88	0.84
20–1000	1581	0.76	0.88	0.85
Enzyme EC. 3rd digit				
All	206	0.50	0.87	0.55
20–1000	150	0.50	0.89	0.55

[a]Correspondence score.
[b]Specificity = $TP/(TP + FP)$.
[c]Sensitivity = $TP/(TP + FN)$. See text for definitions.

sequences). Each sequence is a vertex in a graph, and each pair of sequences defines an edge whose weight is the average pairwise BLAST E-score between those two sequences. A pairwise merging process is conducted as part of an agglomerative clustering, where the initial sequences are used as singleton clusters. The clustering process follows an order of similarity, where each merge is the most significant in terms of the lowest average E-score between two groups of sequences. The main principle applied in ProtoNet is the extensive use of restricted transitivity based on hidden intermediate sequences, coupled with a statistical and biological validation of the hierarchy of the resulting graph. The resulting structure of the graph allows one to navigate through it and to evaluate the capacity of the ProtoNet algorithm to detect remote homologues (see examples[24]).

RESULTS AND DISCUSSION
Unifying Protein Families in ProtoNet

ProtoNet is a bottom-up tree construction in which biologically relevant connections between proteins are evident.[24] In the initial steps of the clustering process, connections between closely related sequences are made but at more advanced stages of the clustering process, more questionable connections are included; at the root of the tree, all proteins are interconnected (except for a few hundred singletons). The hierarchical structure of the graph that is built by ProtoNet enables one to follow the connections between different groups of sequences when considering the progress of the clustering process. Intuitively, one can view it as a reflection of the evolutionary diverging process. As the clustering process advances, similar sequences aggregate to form groups that then unite to form subfamilies. As the clustering algorithm progresses, the families are enlarged, thus forming superfamilies and making remote homologues prevalent. As the number of pairwise connections becomes sparse during the progress of the clustering algorithm (and the statistical

significance of the pairwise connections are extremely low; i.e., BLAST *E*-score as low as 100), the representation of ProtoNet resembles a graph in which the edges connect groups of sequences (clusters) rather than individual sequences.

The properties of the graph in view of its biological content were evaluated by comparing each cluster to other classification methods such as InterPro,[25] SCOP,[26] EN-ZYMES,[27] and other external annotation sources. To evaluate the quality of ProtoNet clusters for structural and functional properties, we scored all *best clusters*. We define such a score as a *correspondence score* (CS). The CS for a certain cluster and a given keyword (i.e., *Enolase N-terminal domain-like* family, in SCOP) measures the correlation between the cluster and that keyword, using the intersect–union ratio.

Correspondence Score (cluster C for keyword K) = $|c \cap k|/|c \cup k| = TP/(TP + FP + FN)$, where c is the set of annotated proteins in cluster C, k is the set of proteins annotated with K, TP, FP, and FN standing for true positives, false positives, and false negatives, respectively. TP = the number of proteins in cluster C that have keyword annotation K; FP = the number of annotated proteins in cluster C that do not have keyword annotation K; FN = the number of proteins not in cluster C that have keyword annotation K. The cluster receiving the maximal CS for keyword K is considered the cluster that best represents K within the ProtoNet tree, and is called the *best cluster*. The score for a given cluster on keyword K ranges from 0 (no correspondence) to 1 (the cluster contains exactly all of the proteins with keyword K; maximal correspondence to the keyword).

For annotation keywords related to structural properties, we used all keywords based on SCOP (fold, superfamily, family, and domain levels), and for a functional view, we used the ENZYME (4 levels of EC hierarchy) and GO (in 3 categories—molecular function, cellular process, and cellular localization) annotations. A complete list of all annotations with the associated ProtoNet *best clusters* is available (www.protonet.cs.huji.ac.il/best_cluster/).

The results from such analysis confirm that clusters that perfectly match functional and structural annotations as defined by the above-listed sources are abundant within the ProtoNet graph. Representative results are shown in Table I. The results for the SCOP family level (contains 1283 keywords) is CS = 0.84 (with a higher score for proteins of a single domain; see Discussion section). By limiting the analysis for cluster size ranges from 20 to 1000 proteins, the CS is somewhat higher (0.86). The average CS for best clusters for the E.C. ENZYME fourth level in hierarchy (contains 2157 keywords) is quite high, but drops significantly for the third digit of the EC classification.

Results from the *best cluster* view point to a relatively narrow interval along the clustering process, where the connection between protein families starts to dominate. Thus, two protein families that share the same cluster in the cluster-map described at that interval may be viewed as remote homologues. This last principle guided us when

we tested for the ability of ProtoNet system to suggest remote homologues.

To compare the ability of ProtoNet to detect remote homologues to that of other available search methods, we choose to test a manually selected set of remote homologues as presented in a comparative survey.[18] In Table II, examples of proteins that share structural and functional relatedness are listed. We duplicated the comparative analysis to include ProtoNet method as an additional, testable methodology in view of the previously described 5 state-of-the-art methods for remote homologues detection.

The remote family representatives are marked by the characteristics of the relevant protein functional family (i.e., Winged helix-turn-helix (HTH) DNA-binding domains; Table II). We began each search with the query protein listed in Table II and "climbed" the cluster hierarchy until reaching the *largest* cluster in which the target proteins that are included in the defined functional family are encountered. The process is aborted, and no success is recorded when the cluster that includes the query and the target proteins annotate less than 85% of the proteins in the cluster (denoted as *TP*—true positives; unannotated proteins are not considered for such procedure). The number of protein families and representatives that were united under that search was reported. In the case that the selected largest cluster is already contaminated by unrelated identifiers or keywords, we marked it (by minus sign; Table II) to indicate a failure in the specificity (purity) of the cluster.

Evaluating the Success in Detecting Remote Homologues

Inspection of the results from Table II indicates that our simple procedure using the scaffold of ProtoNet clusters performs as well as the best method out of the 5 that were tested for almost every query. Note that the benchmark was completely independent of our work and was selected to evaluate the performance of the available 5 independent methods for detecting remote homologues. The different methods are based on exhaustive dynamic programming (FASTA), iterative profile search (PSI-BLAST, PROBE), and traditional profile-based searches (GRIBSKOV, HM-Mer). For details on the listed methods, see ref. 18).

A strong characteristic of our method is its symmetry in the results between query and target protein. This is simply due to the symmetry of the search up a cluster tree—the cluster that unites a query **A** with a target **B** is the same cluster that unites **B** with **A** when **B** is the query and **A** is the target. This symmetry property implies *consistency* of the search by the ProtoNet method. It can be appreciated from Table II in that the other methods do not perform symmetrically.

Other main characteristics of our method are the *simplicity* and *speed* of the search, which requires only basic ProtoNet navigation skills of the user, and its *robustness*, as the cluster-map platform already contains all sequences from Swissprot (ver. 40.28). Every sequence can be referred to as a query or target sequence for such a search.

TABLE II. Comparing the Success in Detecting Remote Homologues by Alternative Methods

Functional family	Query sequence	A	B	C	D	E	PN	Cluster (no. of proteins)	Remarks
Winged HTH DNA-binding domains (1opc, 1aoy, 1dprB, 1hst, 1ecl, 1bgw, 1smtA, 1lea, 1fokA)	1. ompr_ecoli (1opc)	2	1	1	1	2	2−	226420 (1088)	Low *TP,* Detected: 1b9w
	2. argr_ecoli	1	1	1	1	1	1		
	3. 1lea	/	/	/	/	/	3	226170 (348)	*TP* = 1 for SF Detected: 1smtA, 1dprB
Nucleic acid–binding proteins with OB-fold	1. 1ah9	1	1	1	1	1	1		
1. S1 domains, translation initiation factor IF1 (1ah9)	2. cspa_ecoli	1	1	1	1	0	1		
2. Cold shock protein (1mjc) (1asyA, 1lylA, 1cuk, 3ullA, 1cmkA, 1pfsA)	**3. 1asy**	/	/	/	/	/	2	210265 (143)	Detected: 1lyl
1. Pertusis toxin	1. tox3_borpe	1	1	0	1	0	1		
2. Aerolysin (domain)	2. lems_human	Ex	1	0	1	1	1		
3. LINK domains									
4. CTL domain									
1. Chromo domain	mod3_human	1	1	1	1	1	1		
2. DNA-binding proteins (1sap)	1sap	1	1	1	1	1	1		
1. Mite allergen	def2_derfa	1	1	1	1	1	1		
2. Immunoglobulin-like									
Urease**	ure1_kleae	4 [+6]	7	10	1	3	10	225659 (253)	*TP* = −86%
1. PAPS reductase	1. mt16_yeast	1	3	2	2	3	3−	226597 (841)	Low *TP* for all three; high *TP* for 1 and 2 in cluster 212193
2. ATP sulphurylase	2. nade_bacsu	1	3	3	1	1			
3. N-type ATP Ppases (1nsyA, 1gpmA)	3. cysd_ecoli	3	3	3	3	3			
1. HIT (4rhn)	1. fhit_human	2 [+1]	3	3	3	1	2	221320 (48)	Could not find Ap$_4$A proteins through search. High *TP*
2. Ap$_4$A hydrolase									
3. GalT (1hxpA)	2. gal7_ecoli	1	1	3	1	1			
1. Covalent NAD-binding oxidases (1ahu)	1. 1ahu	1	2	2	0	1	2	225895 (79)	High *TP*
2. UDP-*N*-acetylenolpyruvoylglucosamine (2mbr)	2. murb_ecoli	1 [+1]	1	1	2	0			
Coenzyme A transferase	1. X81440	2	2	2	2	2	/	225850 (51)	X81440 not present in Swissprot. 2nd query unified this family
1. Glutaconate	2. atoa_ecoli	2	2	1	1	1	2 (*)		
2. Succinyl, acetate, 3-oxoadipate, butyrate, etc.									
1. Dioxygenase (1han)	1. bphc_burce	1	2	2	1	2	2	222839 (70)	*TP* = 1 for SF Including also 1byl from this SF
2. Glyoxylase (1froA)	2. igul_human	1	2	2	2	1			

Functional families that unify two or more sequence families were combined, as in Holm.[18] The success in detection is marked by the number of successes in view of the predetermined representatives for remote homology. Alternative methods for remote homologue detection were applied over single query proteins, and an enumeration of the sequence families that were recognized by each of the methods was recorded (a result of 1 means the query is recovered; zero indicates failure to identify it). A search in ProtoNet tree was performed in the same way by climbing the cluster hierarchy, starting from a query protein and determining the cluster in which the protein identifier of the target proteins reached > 85% *TP*. The methods applied are abbreviated as follows: A. FASTA walk; B. PSI-BLAST; C. PROBE; D. GRIBSKOV; and E. HMMer (according to Holm[18]); PN, ProtoNet. The mark (−) indicates contamination by unrelated structural families. The mark (/) indicates a query examined solely by ProtoNet. Ex, the number of protein exploded. (*), the entry X81440 could not be matched in ProtoNet s proteins, so we indicated the smallest cluster that unites both families of that query. (**), discussed in detail in the text.

While performing our searches, we were able to find in some functional families more examples of remote homologues that were not queried in the original benchmark (marked by underline, Table II). These examples were found as a by-product of the search in ProtoNet, and were added as additional representatives to enrich the list in Table II.

The protein family of the urease-related hydrolases serves to illustrate the traces of diversifying proteins throughout evolution. Several methods were applied to unify the members of the urease family.[28] All methods used structural alignment, sequence alignment, and profile-based search.[10] The entire group, as defined in the original study,[10] contains 12 protein families that share structurally similar active-site architecture. Of those proteins, 9 are members of the *metallo-dependent hydrolases* SCOP superfamily, 2 are members of a related superfamily named *composite domain of metallo-dependent hydrolases*, and an additional one belongs to a different superfamily (*arylphosphatase*). Of the hydrolases that are in the database, ProtoNet defined 10 out of 12 in a relatively pure cluster (A225659, 253 proteins). This cluster best represents the *metallo-dependent hydrolases* superfamily hierarchy.

Inspecting this cluster obtained in ProtoNet shows that it includes all previously described urease-related superfamily proteins, as well as additional yet undefined related members, as shown in Table III. The different representatives in this large family all share an enzymatic activity marked by EC 3.5.-.-, with the exception of proteins having a different catalytic activity (EC 3.8.-.-.). Interestingly, the evolutionary and functional relatedness between those two apparently different activities was confirmed.[29] Additionally, not only is the high coverage of the urease-related superfamily (in cluster A225659) validated, but monitoring the descendant clusters in the cluster-map provides a good description of the partition of this functional family into more refined subfamilies. Relationships between the subfamilies can be proposed, and the divergence process of the family is traceable.

Figure 1 shows the pairwise similarity within proteins of the cluster A225659 (253 proteins) as a matrix reflecting all-against-all pairwise similarity BLAST E-scores. Darkest color indicates a maximal BLAST E-score (E-score = 0). The gradient in color intensity indicates a decrease in the similarity score. White indicates an *E*-score that is worse than 100. The domination of the white color reflects the very low pairwise similarity among proteins in the cluster. Among all possible pairwise connections, ~70% are worse than *E*-score 100, indicating the sequence remoteness resulting from the BLAST search. The order of the proteins in the matrix reflects the construction of the ProtoNet tree for this subgraph (starting at the bottom right toward the top left). An interactive illustration (color-coded) of the weak connections between all 253 proteins in this cluster is provided at www.protonet.cs.huji.ac.il/homologues/.

In addition to the examples described in Table III and in Figure 1, additional remote homologues are easily de-

tected. For example, the proteins that share the histone core proteins H2A, H2B, H3, and H4 are linked together in a subtree of ProtoNet with a variety of transcription factors, such as the transcriptional initiator TFIIB, TFIID, TATA box, CAAC box binding protein, and more. Despite extremely low sequence similarity (cluster A224843, 284 proteins), all these proteins share an identical structural fold.[30]

In another very interesting case, similarity between alcohol dehydrogenases and crystalline of the eye and some proteins in synaptic vesicles in the nerve system has been proposed.[31] Such homology is apparent at the sequence similarity level only through some weak connecting intermediate sequences. Proteins of the alcohol-dehydrogenases (also polyol-, threonine-, archaeon glucose-dehydrogenases), eye lens zeta-crystallins, *Escherichia coli* quinone oxidoreductase, synaptic vesicle VAT-1 proteins, enoyl reductases of mammalian fatty acid, and yeast erythronolide synthases share the same structural fold, with only very few conserved amino acids throughout this diverged superfamily.[32] This superfamily was termed medium-chain dehydrogenases/reductases (MDRs). A cluster in ProtoNet with 226 proteins (cluster A220748) combines all the MDR proteins that carry very different functional roles but nevertheless share a high degree of structural and biophysical properties. By using the interactive visualization tools within ProtoNet (http://www.protonet.cs.huji.ac.il/), the connectivity among the proteins of the MDR superfamily is shown. Note that the intermediate sequences that connect specific families are traceable by inspecting the pairwise matrix (as in Fig. 1, and see www.protonet.cs.huji.ac.il/homologues/).

The ProtoNet tree allows for the testing of the connectivity among proteins, thus detecting remote homologues. However, our examples are based on analyzing proteins from the Swissprot database (~114,000 proteins). As the number of protein sequences currently available is at least 10 times higher, the relevance and the robustness of our method toward the larger database should be re-evaluated. We developed an extended version of ProtoNet that includes over 1 million proteins (ProtoNet 3.0), based on Swissprot 41.12 combined with TrEMBL 24.8. Many of the proteins in this combined data set are still poorly annotated. In ProtoNet 3.0, nodes in the tree that are candidates for representing clusters of remote homologues are eventually much larger. We tested the robustness of the results in the extended version of ProtoNet. In general, despite a ~9-fold increase in the database size and a reduced quantity and quality of the associated annotations, the performance of the system is only slightly reduced. We illustrate it for the urease superfamily that was discussed above. A cluster representing the urease-related superfamily (A293766, ProtoNet 3.0) is composed of 1508 proteins that include 16 out of the 17 groups (as in Table III) with a very high sensitivity (i.e., collecting all proteins from the database that belong to this enzymatic group). One additional urease-related enzymatic group (*N*-acyl-D-amino acid deacylase) that was missed from the Swissprot based cluster (Table II) is detected in cluster
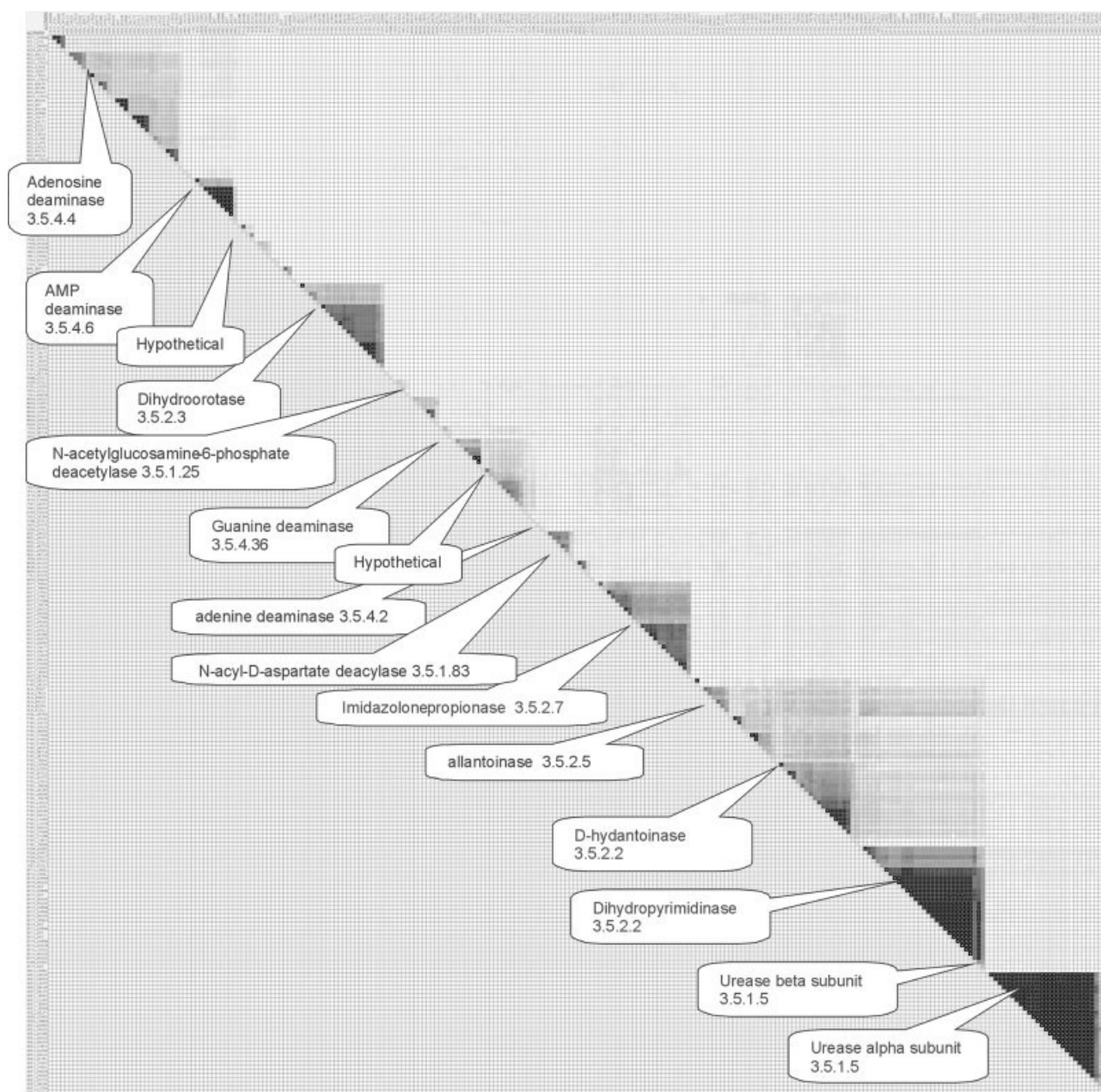
Fig. 1.   All-against-all BLAST *E*-score in a subtree of ProtoNet for urease-related proteins. Representative enzymatic activity of some of the proteins is shown in the callouts. The matrix shows all 253 proteins in cluster A225659. Black indicates maximal BLAST *E*-score (0). Dark gray indicates very significant pairwide similarity. A gradient of black to white parallels a decrease in the similarity score.

A293766. This example illustrates the potential of the ProtoNet tree to detect remote homologues even in a relatively poorly annotated database. It is worth noticing that among the 1508 proteins, most of them (58%) have no Prosite annotation, and 21.1% are marked as hypothetical proteins. Thus, the inference of functional groups is a direct byproduct of navigating the ProtoNet tree.

A careful inspection of the families with remote homologues showed that the performance of ProtoNet is very satisfactory for single-domain proteins (as in the case of histones, urease, and MDR proteins); a much lower success rate can be obtained for proteins that are multido-

main. The relatively low scores in the *best cluster* analysis (www.protonet.cs.huji.ac.il/best_cluster/) best reflect such instances. While ProtoNet is based on a whole protein clustering, and remote homology (especially from a structural perspective) is often associates with a domain level, it is appropriate to use this methodology with a critical objective.

In summary, we offer the use of ProtoNet's cluster-map as a platform for searching for remote homologues and for identification of weak connections between functional families. We propose a simple search procedure for finding remote homologues for a query sequence and compare it to

TABLE III. The Enzymatic Groups Associated with ProtoNet Cluster A225659

Enzyme Name	no. of Proteins in the Cluster/no. in Database	E.C.
(1) Adenosine deaminase	28/28	3.5.4.4
(2) Adenine deaminase	6/6	3.5.4.2
(3) Allantoinase	6/6	3.5.2.5
(4) AMP deaminase	11/11	3.5.4.6
(5) N-isopropylammelide isopropylamidohydrolase	1/1	3.5.99.4
(6) Atrazine chlorohydrolase	1/1	3.8.1.8
(7) Hydroxydechloroatrazine ethylamonohydrolase	1/1	3.5.99.3
(8) Cytosine deaminase	1/3	3.5.4.1
(9) Dihydropyrimidinase 2	10/10	3.5.2.2
(10) Guanine deaminase	7/8	3.5.4.3
(11) Imidazolonepropionase	23/23	3.5.2.7
(12) N-acetylglucosamine-6-phosphate deacetylase	6/6	3.5.1.25
(13) N-acyl-D-aspartate deacylase	1/1	3.5.1.83
(14) D-aminoacylase	1/1	3.5.1.81
(15) N-acyl-D-glutamate deacylase	1/1	3.5.1.82
(16) Dihydroorotase (DHOase)	49/49	3.5.2.3
(17) Urease (Urea amidohydrolase)	78/78	3.5.1.5

The annotations for 17 different groups are extracted from the ENZYME database. Note that for almost all groups, the coverage is complete, with no additional proteins carrying the same enzymatic activity reported in the database used.

a benchmark of performance for different methods that rely on information beyond the elementary pairwise similarity. We show that in almost all cases, our search method performs as well as the best method. Our method is very fast and consistent due to the fact that it relies on a predetermined scaffold of the protein space that is in accord with the process of evolutionary diversification.

Partitioning the protein space into homologous families is motivated by Structural and Functional Genomics initiatives. As seen from the benchmark and the additional examples, the identification of remote homologues for a given sequence extends our ability in predicting structural relatedness and in many instances also the functional characteristics. We propose a robust method for finding such remote homologues and defining structural and functional superfamilies.

ACKNOWLEDGMENTS

REFERENCES

1. Sternberg MJ, Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
2. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 1997;269:423–439.
3. Stark A, Sunyaev S, Russell RB. A model for statistical significance of local similarities in structure. J Mol Biol 2003;326:1307–1316.
4. Saier MH Jr. Phylogenetic approaches to the identification and characterization of protein families and superfamilies. Microb Comp Genomics 1996;1:129–150.
5. Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002;318:595–608.
6. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
7. Jones DT, Swindells MB. Getting the most from PSI-BLAST. Trends Biochem Sci 2002;27:161–164.
8. Karwath A, King RD. Homology induction: the use of machine learning to improve sequence similarity searches. BMC Bioinformatics 2002;3:11.
9. Muller A, MacCallum RM, Sternberg MJ. Benchmarking PSI-BLAST in genome annotation. J Mol Biol 1999;293:1257–1271.
10. Holm L, Sander C. An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. Proteins 1997;28:72–82.
11. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
12. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
13. Yona G, Levitt M. (2002). Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. J Mol Biol 2002;315:1257–1275.
14. Spang R, Rehmsmeier M, Stoye J. A novel approach to remote homology detection: jumping alignments. J Comput Biol 2002;9:747–760.
15. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. Bioinformatics 2004;20:467–476.
16. Jaroszewski L, Rychlewski L, Godzik A. Improving the quality of twilight-zone alignments. Protein Sci 2000;9:1487–1496.
17. Sasson O, Vaaknin A, Fleischer H, Portugaly E, Bilu Y, Linial N, Linial M. ProtoNet: hierarchical classification of the protein space. Nucleic Acids Res 2003;31:348–352.
18. Holm L. Unification of protein families. Curr Opin Struct Biol 1998;8:372–379.
19. Pearson WR. Using the FASTA program to search protein and DNA sequence databases. Methods Mol Biol 1994;24:307–331.
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.

21. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
22. Bailey TL, Gribskov M. Methods and statistics for combining motif match scores. J Comput Biol 1998;5:211–221.
23. Neuwald AF, Liu JS, Lipman DJ, Lawrence CE. Extracting protein alignment models from the sequence database. Nucleic Acids Res 1997;25:1665–1677.
24. Sasson O, Linial N, Linial M. The metric space of proteins—comparative study of clustering algorithms. Bioinformatics 2002; 18(Suppl 1):S14–S21.
25. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley R, Courcelle E, Durbin R, Falquet L, Fleischmann W, Gouzy J, Griffith-Jones S, Haft D, Hermjakob H, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Orchard S, Pagni M, Peyruc D, Ponting CP, Servant F, Sigrist CJ. InterPro: an integrated documentation resource for protein families, domains and functional sites. Brief Bioinform 2002;3:225–235.
26. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
27. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000;28:304–305.
28. Heger A, Holm L. Sensitive pattern discovery with "fuzzy" alignments of distantly related proteins. Bioinformatics 2003;19(Suppl 1):I130–I137.
29. Sadowsky MJ, Tong Z, de Souza M, Wackett LP. AtzC is a new member of the amidohydrolase protein superfamily and is homologous to other atrazine-metabolizing enzymes. J Bacteriol 1998;180: 152–158.
30. Baxevanis AD, Landsman D. Histone Sequence Database: new histone fold family members. Nucleic Acids Res 1998;26:372–375.
31. Linial M, Levius O. VAT-1 from Torpedo is a membranous homologue of zeta crystallin. FEBS Lett 1993;315:91–94.
32. Persson B, Zigler JS Jr, Jornvall H. A super-family of medium-chain dehydrogenases/reductases (MDR): sub-lines including zeta-crystallin, alcohol and polyol dehydrogenases, quinone oxidoreductase enoyl reductases, VAT-1 and other proteins. Eur J Biochem 1994;226:15–22.