Dictionary Building Via Unsupervised Hierarchical Motif Discovery In The Sequence Space of Natural Proteins

Isidore Rigoutsos, Aris Floratos, Christos Ouzounis, Yuan Gao, Laxmi Parida

Abstract

Using TEIRESIAS, a pattern discovery method that identifies all motifs present in any given set of protein sequences without requiring alignment or explicit enumeration of the solution space, we have explored the GenPept sequence database and built a dictionary of all sequence patterns with two or more instances. The entries of this dictionary, henceforth named seqlets, cover 98.12% of all amino acid positions and in essence provide a comprehensive finite set of descriptors for protein sequence space. As such, seqlets can be effectively used to describe almost every naturally-occurring protein. In fact, seqlets can be thought of as building blocks of protein molecules which are a necessary (but not sufficient) condition for function or family equivalence memberships. Thus, seqlets can either define conserved family signatures, or cut across molecular families and previously undetected sequence signals deriving from functional convergence. Moreover, we show that seqlets can also capture structurally-conserved motifs. The availability of a dictionary of seqlets that has been derived in such an unsupervised, hierarchical manner is generating new opportunities for addressing problems that range from reliable classification and the correlation of sequence fragments with functional categories, to faster and sensitive engines for homology-searches, evolutionary studies, and protein structure prediction.

Keywords: dictionary, sequence analysis, seqlets, patterns, motifs, functional conservation, structural conservation, GenPept, family signatures, evolution, structure prediction.

1. Introduction

Protein sequences define an immensely complex space, with an upper bound of $20^N$ possible combinations for any sequence of length $N$. For example, a 100 amino acid sequence could give rise to $20^{100}$ or $\approx 1.27\times10^{130}$ potential distinct sequences, an astronomical number. Nonetheless, protein sequence space is sparsely populated because proteins are related by divergence and form molecular families which are believed to be evolving through random drift and natural selection processes: the common elements of molecular families are particular conserved positions which have been defined as patterns or motifs (27).

Traditionally, the pattern determination process has been based upon the identification of similar proteins, the subsequent generation of multiple alignments and the selection of the most conserved sites as the representative protein domain signatures (9,35).

Up to now, it had not been possible to enumerate the most frequent patterns in a large dataset such as GenPept or Swiss-Prot (6) by treating such a dataset as a whole. Consequently, the derivation of patterns had proceeded on a
case by case basis (47). Moreover, most definitions were based on the underlying assumption that patterns can be found only within divergent families (36); the few convergently-related patterns for functional motifs (14) such as nuclear localization signals (10) were usually not specific enough and could not be readily discovered by alignment (29).

Ideally, the pattern discovery process in any such large dataset should be carried out in an unsupervised manner and without any assumptions as to potential family memberships since such assumptions only limit the set of potential answers. In fact, when one subselects a set of proteins from a given database and uses any of the available tools (30,34,48 etc.) to discover any and all of the patterns present in this set (23), one has implicitly accepted that the members of the formed set are related, and, the sequences under consideration form indeed a single set (as opposed to the union of two or more smaller sets). This is characteristic of previous work where distance-based clustering of complete sequences is the starting point (22,21,32,33,50,53).

Achieving the goal of discovering the most commonly occurring patterns in a sequence database requires at the very least the ability to: (a) handle very large datasets, (b) discover all existing patterns, (c) determine the natural boundaries of these patterns (i.e. maximality in length), and (d) guarantee that these length-maximal patterns are also as specific as possible (i.e. maximality in composition). Upon completion, one will have derived a natural pattern dictionary for the sequence space of proteins.

Similarly to dictionaries for natural languages, this bio-dictionary will contain reusable elements (seqlets) that have been identified after having processed a large body of text (protein database). Some of the dictionary seqlets are bound to appear more frequently than others: this can be either the consequence of an inherent bias in the natural usage of a given seqlet, or the reflection of a bias in the way that the protein sequence space is currently being sampled (by the ongoing genome sequencing efforts). Independent of what the case may be, the seqlets are by design reusable and expected to capture signals that have not been previously observed. Naturally, a comprehensive dictionary of seqlets is of paramount importance due to its many uses that include automated annotation, reliable classification, sensitive homology detection, structural characterization, and evolutionary studies to name just a few.

The dictionary entries can be thought of as compact descriptors for describing protein segments. This idea has been used successfully in a variety of tools (e.g. PROSITE (5), BLOCKS (24), PFAM (49), PRINTS (4), and others) in the form of patterns, profiles, and HMMs for several years already.

The novelty of what we present here derives from several facts: the method is unsupervised, it makes no use of domain-specific information to generate the results, it operates on a very large input of unaligned sequences with variable lengths, and it makes no assumptions as to potential relationships among the various sequences in the input. Furthermore, since the discovery process proceeds outward (i.e. from a seed amino acid to a combination of amino acids) it does not have to rely on any globally observed similarity to derive its results and thus it is not limited by the presence or absence of such a similarity.

Seqlets that have been obtained from large protein collections of diverse composition are undoubtedly expected to unveil previously unobserved protein features both within and across family boundaries. Furthermore, they will lead to a better understanding of protein architecture, and also discover new relationships between families which have traditionally been assumed to be unrelated.

In this work, we discuss how we build a comprehensive dictionary by having processed the most recent release of NCBI’s GenPept database (February 10, 1999). The resulting dictionary covers 98.12% of all amino acid positions in this database. The dictionary derives its importance from both the variety and significance of the
computational biology applications it can impact; these include but are not limited to automated functional annotation, local homology detection, 3-dimensional structure characterization, evolutionary studies, etc.

2. Notation And Definitions

We will denote by $\Sigma$ the alphabet of all amino acids; i.e. $\Sigma=\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. We can succinctly capture the output generated by the TEIRESIAS algorithm (38,39) with the regular expression

$$\Sigma \cup (\Sigma \Sigma^* \Sigma)^* \Sigma \cup (\Sigma \Sigma^* \Sigma)^* \Sigma \cup \Sigma \{1\}$$

i.e. the generated patterns can either be a single alphabet symbol, or strings that begin and end with a symbol or a bracket with two or more characters, and contain an arbitrary combination of zero or more residues, brackets with at least two alphabet characters, and ‘.’ characters. The ‘.’ (referred to as the “don't care character”) is used to denote a position that can be occupied by an arbitrary residue. A bracket is meant to denote a “one of ” choice; i.e. [KRH] denotes exactly one of K, R, or H. A bracket can have a minimum of 2 alphabet characters but not more than $|\Sigma|-1$.

We use the term seqlets to refer to the patterns that are discovered by processing a given input comprising many sequences. For our discussion, seqlets are captured by the above regular expression but they can of course be more general. Typically, a minimum length is enforced upon the seqlets in order to guarantee their information content; this minimum length is controlled by two input parameters: $L$ and $W$.

A seqlet $S$ is called an $<L,W>$ seqlet (with $L \leq W$) if every substring of $S$ with length $W$ contains $L$ or more non-don't care positions. The smallest length of an $<L,W>$ is obviously equal to $L$ whereas its maximum length is unbounded. Any given choice for the parameters $L$ and $W$ has a direct bearing on the degree of remaining homology among the instances of the sequence fragment that the seqlet captures: the smaller the value of the ratio $L/W$, the lower the degree of homology within any stretch of window. Also, larger values for $L$ result in seqlets with higher information content (more specific). Some specific examples of $<6,15>$ seqlets that were discovered during the processing of the GenPept database include some well known cases such as: G..G.GK[STG]TL (ATP/GTP binding P-loop), H.....HRD.K..N (Serine/Threonine protein kinases), KMSKS[LKDIR][GNDFQ]N (class I aminoacyl-tRNA synthetases), V.I.G.G...A (NAD/FAD-binding, flavoproteins), and GDG[IVAMTD]ND [AILV][PEAS][AMV][LMIF]..A (cation-transporting ATPases).

Two additional properties of the TEIRESIAS-generated seqlets are maximal composition and maximal length with respect to the processed database $D$; an $<L,W>$ seqlet $S$ is called maximal if it cannot be augmented into a more specific $<L,W>$ seqlet (by either appending/prepending a string on $\Sigma$, or dereferencing one or more don't care characters) without simultaneously reducing the seqlet’s position list. A seqlet’s position list is a list of {protein identifier, offset within protein} pairs that uniquely identify all the instances of a seqlet. The maximality properties are clearly very important in the context of discovering the natural extent of sites and domains.

A seqlet $S$ is a compact representation for a collection of strings on $\Sigma$, and in fact it defines a regular formal language $G(S)$. The elements of this language are all the strings that can be obtained from the seqlet by substituting each don’t care by an arbitrary residue from $\Sigma$, and each bracket with exactly one of its member residues. For example, the seqlet “G..G.GK[STG]TL” that describes the ATP/GTP binding P-loop defines the language $G(“G..G.GK[STG]TL”)$. The members of which include the strings GSAGSGKSTL, GYLGSGKSTL, and

---

6 Indeed if the bracketed expression contains $|\Sigma|$ characters, then it becomes equivalent to the don't care character.
GESGSGKTTL among others. A protein is said to match a given seqlet $S$ if and only if it contains at least one substring (i.e. a block of consecutive residues) that belongs to $G(S)$. Clearly, a given protein may match more than one distinct seqlets.

Two or more proteins that match the same seqlet $S$ can be thought of as sharing the region of sequence space that is captured and characterized by the seqlet $S$: the elements of the language $G(S)$ are all the potential instances of sequence fragments that can be found in this region of sequence space. Attaching ‘meaning’ to a seqlet through evaluation and analysis of the sequence fragments that it represents will automatically label the corresponding region of sequence space. This label can either be family-specific (in which case the seqlet can be used as a predicate; e.g. the above mentioned seqlets), or it can correspond to a shared element such as the P-loop (G..G.K[STG]TL). We pursue this discussion in more detail in the sequel.

Finally, in the discussion that follows we will use the term *input database* to refer to the instance of the GenPept database which we processed.

### 3. Dictionaries And Seqlets

The goal of our effort is to derive compact descriptors for the entire protein sequence space through unsupervised pattern discovery on the sequence representations of a very large collection of proteins. Clearly, the coverage of the protein sequence space will be as complete as the sampling provided by the currently available protein sequences. As a rule of thumb, the sought descriptors should be general enough to allow for small variations but not too general to lose specific meaning. The necessary number of descriptors is a function of the granularity at which one is trying to address the question.

The biological problem that we are addressing is analogous to the problem of processing an unknown natural language when the only available knowledge is the set of its most basic units. For natural languages, these most basic units can be either alphabet letters (e.g. Hebrew, Greek, English, etc.) or syllabic signs (e.g. Linear B, Japanese Kana, etc.). In the absence of any language-specific information, one attempt to understand such a language would proceed in a bottom-up manner: from the most basic unit, to the level of the vocabulary, then to the level of the syntax and semantics. Given the availability of large amounts of sample text in the unknown language, it is fairly straightforward to determine, through application of statistical or combinatorial methods, the building blocks at increasingly higher levels of the hierarchy. In fact, it was precisely this approach that led to the decipherment of the Linear B script by Michael Ventris in 1953 (42). But, when the available sample text is limited very little can be said about the corresponding language without extensive study and context-specific considerations: this is exactly the case of Linear A, a script which, although related to Linear B, continues to elude decipherment (42).

In proteins, the most basic units are the 20 natural amino acids. The sought seqlets will in essence describe the level of the hierarchy immediately above the most basic units and will correspond to the *words* of this language. We conjecture that these words are combined together into sentences and viable structures through a syntax which is imposed on seqlets. Additionally, *semantics* need to be attached to seqlets in order to build a full-fledged dictionary: to a certain extent, attaching semantic information can be carried out automatically through use of the currently available annotations for functional sites and domains. Additionally, using the crystallographic information that is available in databases such as the Protein Data Bank (1,7), a word/seqlet can be given structural meaning. Of course, it should be clear that any unsupervised discovery method will produce numerous seqlets that have not been previously observed/studied and which will remain functionally and/or structurally uncharacterized.
To recapitulate, we are given the alphabet of 20 amino acids and a large number of sentences/sequences (= a database of proteins and protein fragments) and we will apply motif discovery in order to determine as many of the words (=seqlets) in the vocabulary as possible. The discovery of these seqlets will proceed without recurrence to any domain-specific information and will only be based on the fact that something has been re-used. Once a vocabulary entry has been determined, processing of the already accumulated knowledge will attach functional and structural meaning to it, thus turning the seqlet into a complete dictionary entry. We will use seqlets to completely describe the protein sequence space given the sampling provided by the input database \( D \) under consideration. Recall that each seqlet describes a fragment of \( A \) amino acids. By design, this fragment is present in several sequences of \( D \); consequently, the number of seqlets that are required to cover the database \( D \) will be only a small fraction of the total number of amino acids in \( D \).

Analogously to natural language words, seqlets can be considered to fall into one of three categories depending on the amount of information that they carry: they may correspond to descriptive elements (such as nouns or verbs), to modifying elements (such as articles, adjectives, adverbs or prepositions), or to connecting elements (such as conjunctions or articles). We next examine this in more detail.

**Seqlets As Family Membership Predicates**

Typically, motifs have been obtained by operating on rather small sets of related, manually compiled proteins. Consequently, these descriptors identify regions that played a role in the specific function performed by the sequences of the set under consideration. Herein, the input set comprises a very large collection of diverse sequences and is processed in an unsupervised manner. One thus expects to identify family-specific elements as well as elements of a more elementary and reusable nature; i.e., seqlets will not only encompass individual families (e.g. active sites) but they will also go beyond family boundaries (e.g. cellular localization signals).

In general, the seqlets that we discover by processing a database such as GenPept belong to one of the following three categories:

1. **one seqlet is specific to a single protein family**: these seqlets correspond to motifs that are present in one family only.

   As an example case, let us consider the family of bacterial histone-like DNA-binding proteins. These are small DNA-binding proteins belonging to two subfamilies: the HU family of proteins that stabilize DNA from denaturation under extreme environmental conditions and the IHF (integration host factor) family with roles in recombination and transcription initiation (45). The collection of seqlets we derive from processing the input database includes

   \[F\{GLT\}[^{][FIV][RKPQA][APQES][RST][GA][RVFH][NK]P:T\]

   that is specific to the family of bacterial histone-like DNA-binding proteins. In order to further qualify the results, we use as a point of reference the relevant prosite entry (PS00045) from Release 15.0 of the Prosite database. Searching Release 36.0 of the Swiss-Prot database (on which the Prosite entry PS00045 was based), the Prosite motif gave rise to 42 true positives, 0 false positives and 1 false negative. On the other hand, searching Swiss-Prot with the above seqlet we obtain all 43 true positives (i.e. the above seqlet is present in Prosite’s 1 false negative) and no false positives.

2. **two or more seqlets are specific to a single protein family**: in this case, there are multiple seqlets that are specific to a given family.
As an example, we will use the family of DEAD box helicases (46). Two of the seqlets in the collection that we obtain from processing our input database are

(a) [IVLCTA][LFVIMK][VI][LMIFV][DE][AS][D][MLIFC]...[FGHLRW]

and

(b) [YFLH][VILMQ][HR][IVASTC][GR][TSOC][GA][RQC][GTNAK][GKT][ASKVC]

As before, we used the relevant prosite entry (PS00039) as a point of reference. Of the 94 sequences listed under PS 00039, searching Release 36.0 of the Swiss-Prot database with the reported Prosite motif captures 89 true positives, misses 5 true positives, and generates 12 false positives. A search of the same database with seqlet (a) above captures all 94 true positives without generating any false positives. Analogously, using seqlet (b), we capture 90 of the 94 true positives as well two more sequences (DBP7_YEAST and DBP9_YEAST) both of which are RNA helicases; seqlet (b) does not generate any false positives.

3. One seqlet is present in several protein families; in this category, we find seqlets that correspond to motifs preserved within members of different families. Among the best such examples is the seqlet

G..G.K[STG]TL

which is present in multiple families (44) and corresponds to an ATP/GTP binding motif (P-loop). Obviously, seqlets such as this one cannot be used to classify a sequence into a protein family. However, these seqlets have their own merit since they capture entities which are meaningful and have been derived in an unsupervised manner. In fact, the method we described has discovered the concept of the P-loop after having processed a large database containing many instances of this binding motif.

This variety is not surprising when one considers that these seqlets are obtained by treating large collections of very diverse, unaligned proteins. It is expected (and actually desired) that these seqlets model signatures that are family specific but also cross family boundaries.

Typically, and for seqlets that intersect/describe collections of proteins which are already known to be related, the significance is quantified in terms of the seqlet’s specificity and sensitivity; specificity is a measure of how many database sequences that are not family members (= false positives) the descriptor captures, whereas sensitivity is a measure of how many members of the family the descriptor does not capture (= false negatives) (51).

Specificity and sensitivity are terms which are applicable only in the context of a particular, adequately delineated family. A substantial portion of the seqlets that we generate by processing our input database captures protein regions that have not been previously reported or characterized in the literature.

Independent of whether a given seqlet has been reported before or not, we have also attempted to characterize it structurally by accessing databases of crystallographic data. Recall that a complete dictionary entry will at the very least comprise three elements: the seqlet, its functional meaning and its structural characterization.

**Seqlets As Descriptors Of Structurally Conserved Regions**

Occasionally, the seqlets that we discover by processing a large protein database describe amino acid fragments for which crystallographic data are available. In such cases, we can augment a seqlet’s dictionary entry with the structural characterization for the seqlet, by determining it as follows: given a seqlet, we identify all its sequence-instances in a database such as the Protein Data Bank (PDB); then, we extract all the corresponding 3-dimensional fragments. If we have at least two or more instances of a seqlet in the PDB, we can generate an alignment in
3-dimensional space for the respective fragments and compute the resulting RMSD error. Seqlets which gave rise to the same 3-dimensional structure (i.e. they always ‘folded’ in the same manner) will result in alignments with small RMS error values.\footnote{7}

As in the case of one-dimensional sequences, the 3-dimensional structure that is captured by a seqlet can be family-specific (one seqlet/structure-one family, multiple seqlets/structures-one family) or can cross family boundaries (one seqlet/structure-multiple families). We next give one example from each such category.

Figure 1. The structures captured by the two seqlets Y.V...TP[DE]G[DE]..[IMVL] (left) and A..PA.AA......A (right). The RMSD errors of the shown alignments are 2.28 and 0.94 Angstroms respectively. See also text for more details.

Let us first consider the seqlet Y.V...TP[DE]G[DE]..[IMVL] which is among those discovered by processing our input database: it is easy to check that all GenPept sequences containing this seqlet are ferredoxins. If we now search the PDB for instances of this seqlet, we find that it is present in ferredoxins from two different cyanobacteria, namely *Apanothece sacrum* (entry: 1FXI) and *Synechocystis* sp. (entries: 1DOX/1DOY) - clearly, only one of 1DOX/1DOY is needed for the computation of the RMS error value for the alignment of the corresponding fragments. This error value was equal to 2.28 Angstroms, and the fragments from 1FXI and 1DOX that the seqlet captures are shown aligned in Figure 1. This is an example of a seqlet that is specific enough to be used as predicate for family membership, and which corresponds to a conserved structure.

We next examine the seqlet A..PA.AA......A which is present in many functionally diverse sequences of the processed input. This is an excellent example of a seqlet that captures a local homology shared among many sequences of distinct function. Searching the PDB for instances of this seqlet shows that it is present in cytochrome *c* from *Rhodospirillum molischianum* (entry: 2CCY) and a methylmalonyl CoA mutase from *Propionibacterium freudenreichii* (entries: 1REQ/2REQ/3REQ). Superposition of 2CCY and 1REQ results in an RMSD error that is equal to 0.94 Angstroms; the two aligned fragments are shown in Figure 1.

**Seqlets Of Low Complexity**

\footnote{7 It is very important to make sure that, prior to any RMSD error calculations, the PDB be cleaned up and does not contain redundant copies of near-identical proteins: if $n-1$ of a seqlet’s $n$ instances appear in $n-1$ sequences with remaining pair-wise homologies exceeding 90%, the computed error will be artificially biased toward small numbers. This can be easily achieved by removing all but one instances of any highly-homologous sequences from the PDB prior to identifying a seqlet’s instances.}
An additional concern relates to the presence of low complexity regions in the input sets. Traditionally, such regions have been masked prior to processing. We have decided not to filter out such regions, despite the fact that their removal greatly facilitates the seqlet discovery problem from a computational standpoint.

What led us to this decision was the existence of cases where seemingly low-complexity regions proved specific for a given family of proteins. One such example was the seqlet AL.AA..AA.A which appears in 154 sequences in GenPept. An instinctive reaction would be to dismiss this seqlet due to its apparent low complexity and its composition. However, further study reveals that this seqlet is the core component for a chemotactic transduction-specific signature. In fact, 77 of these 154 sequences share the 27-residue signature:


and are annotated as chemotactic receptors or transducers; the seqlet AL.AA..AA.A has allowed us to “zoom-in” to a region that exhibits exceptional conservation, and is apparently associated with a common functional property of some of these proteins.

Complete vs. Partial Coverage Of The Sequence Space

We have already stated that our objective is to generate and derive compact descriptors for the entire protein sequence space. In the context of building a dictionary, it is imperative that we achieve as complete a coverage of the targeted space as possible.

It can be argued that such a complete coverage is not necessary and that the (partial) coverage which is provided by collections of PROSITE-like descriptors ought to suffice. Generally speaking, this is a true statement but stems from a different objective, that of having good descriptors for protein (super-, sub-) families and their functional sites. All regions of proteins are described because the used discovery method can derive properties of sequence space that are impossible to obtain by partial covering. In this sense, our work goes beyond the mere derivation of descriptors for functional sites and protein families. Any combination of amino acids that has been reused in the sample of sequence space ought to be included in the compiled dictionary.

Before we discuss the issue of minimum size for these amino acid combinations, and in complete analogy to natural languages, we should point out that the law of diminishing returns is evident in the context of this bio-dictionary as well: although, we can account for a substantial part of the sequence space with a relatively small number of descriptors, increasingly more descriptors are needed to describe the uncovered regions.

Significance Of Seqlets In The Absence Of Domain-Specific Knowledge

It is easy to determine biological relevance for those seqlets that capture regions with known functional or structural behavior. But in the general case, the local homology captured by the seqlet may be the only information at hand. In the absence of domain-specific knowledge, one needs to consider alternative ways for characterizing the importance of a seqlet. Ideally, any proposed measure of importance should distinguish between biologically relevant and irrelevant results.

A number of measures based on either information theoretic (30,52) or statistical frameworks (2) have been proposed over the years and could be applied here. Complex relationships including phylogenetic distribution of properties cannot be described by the existing frameworks. As an example, let us consider the following two seqlets: allowing for conservative substitutions, there exist 6 instances of the seqlet AREGRKFGVGL and 29 instances of

---

8 The assumed equivalence classes are: \{G,A\}, \{D,E\}, \{K,R\}, \{M,I,V\}, \{F,Y\}, \{C\}, \{P\}, \{W\}, \{H\}, \{S,T\}, \{Q,N\}.
the seqlet GDISYSLYLIHW in the processed release of our input database. And in all cases, the function of the proteins that contain them is unknown. Both seqlets are of comparable lengths and composition and thus any relevance estimates ought to be based on the number of their occurrences in the input database. This count will of course favor the seqlet GDISYSLYLIHW over the seqlet AREGRKFGVGL. But the latter is most likely characteristic of the entire phylogenetic domain of Archaea since it appears in 4 archaeal organisms (1-Methanobacterium thermoautotrophicum, 1-Archaeoglobus fulgidus, 3-Methanococcus jannaschii, 1-Pyroccocus horikoshii); on the other hand, all 29 instances of GDISYSLYLIHW come from Caenorhabditis elegans and are likely the result of paralogous duplication.

In light of such examples the number of occurrences of a seqlet should be of lesser importance to the seqlet’s information content which is controlled by the parameters $L$ and $W$. We thus set out to discover $<L,W>$ seqlets that appear at least 2 times in the input database. As far as an educated choice for the parameters $L$ and $W$ is concerned we took the following approach: the value of $L$ was chosen to be the smallest value that provided differentiation between the processed database and a random database with the same size and amino acid distribution. The value of $W$ was chosen so that $L/W$ would correspond to the minimum sought local homology. This methodology is described in detail in (17).

### Clustering Of Related Seqlets

During the discovery process, a multitude of seqlets will be discovered that refer to the same core entity which is reused. These seqlets differ from one another in their specificity and can be thought of as the equivalent of grammatical variations in natural languages. For example, the seqlets appearing on Table 1 are all variations of the P-loop with varying degrees of specificity. Naturally, one would wish to **cluster** together seqlets that correspond to the same building block $B$.

In the context that we are examining, two simple methods for clustering seqlets can be used. The first is appearance-based clustering: here, a seqlet $S$ is put in the cluster defined by a seqlet $S_0$ if and only if $S$ ‘looks’ like $S_0$ with the extent of agreement that is necessary for membership controlled by a predefined parameter. For example, G..GSKS and G..GKTT would be put in the cluster defined by G..GKST if the minimum desired overlap was 80% of the seqlets’ positions. The second is position-based clustering: here, a seqlet $S$ is put in the cluster defined by a seqlet $S_0$ if and only if a predefined portion of the instances of $S$ have a minimum predefined overlap with the instances of $S_0$. The results depend, of course, on the amount of intersection of the position lists of $S$ and $S_0$ and the desired minimum overlap of the two instances. In both cases, a seqlet that cannot be made part of any already existing cluster can be used to start its own cluster.

As with all other domains where clustering has been applied, there is no correct clustering in the case of seqlets either. In fact, the very concept of correct clustering is vague: assuming that the distance metric has been determined and fixed, an answer begins to exist only after the selection of thresholds and parameters, and there is no rule of thumb for selecting appropriate thresholds or parameters. Although it would be desirable to generate a clustering of the seqlets we discover, one should bear in mind that assigning a seqlet to a cluster essentially commits the seqlet to a specific membership.

| G..G.GKST | G.GK.TL...L | V.L.G..G.GK |
| G..G.GK[STGDAR]TL | L.G....GKST | L.G....GKTY |
4. Methodological Details

Before we proceed with the description of the obtained results, we discuss in detail the method setup as well as all parameter choices and other experimental decisions.

Description of The Seqlet Discovery Tool

The main computational tool that we have used in this work is the Teiressas algorithm (38,39). Teiressas is a deterministic algorithm that allows one to carry out pattern discovery in biological sequences. The algorithm carries out this task without need for a data model and generates all patterns appearing \( K \) or more times in the input with the additional guarantee of pattern maximality in both composition and length. No enumeration of the solution space is necessary or indeed possible. The generated patterns are maximal in composition and length and there is no restriction as to their length, location, spatial extent, composition, or minimum number of occurrences. Finally, motifs occurring across sequences as well as “internally repeating motifs” are treated in a uniform manner.

Removal Of Identical Sequence Fragments From The Input. Prior to carrying out the seqlet discovery, we processed the input database and masked all instances of the identical amino acid fragments that are present in it except for the one instance that appears in the longest sequence among those containing the seqlet. No conservative substitutions are allowed at this stage (e.g. KPKLGL cannot match RPRMGV). Also note, that we do not ‘remove’ entire sequences from the input but only those of their portions that correspond to the recurrent fragment. The natural language equivalent to this masking is the removal of all quasi-identical phrases (in this case amino acid fragments) present in the sentences that are processed; clearly, multiple copies of such fragments do not contribute
anything to the discovery process. To this end, we have used Teiresias with a setting of \( L = 6 \), \( W = 6 \) and \( K = 2 \). Note that conceptually similar choices have been made in (28) which describes an alternative method that could be used in this stage. A third method can be found in (40). Again recall that the choice of \( L \) and \( W \) does not determine the maximum length of the removed fragments but only their minimum density (see also the discussion from page 3). After having determined the seqlets corresponding to these identical sequence fragments, their instances in the input were masked through replacement with symbols not belonging to the alphabet \( \Sigma \); the seqlets that were discovered in this phase were of course made part of the final collection.

**“Covering” The Input.** As we have already described, an important quality criterion for the set \( S \) of discovered seqlets is its ability to cover the sequences in the input set from which they were generated. One can evaluate the coverage of the input in one of two ways: either by computing the coverage of the sequences in the dataset, or by computing the coverage of the amino acids in the dataset. In the first case, we consider a sequence \( P \) in the input database \( D \) to be covered if and only if there exists at least one seqlet \( S \) in \( S \) with an instance in the sequence \( P \). Clearly, a given sequence may be multiply covered by more than one seqlets. In the second case, we consider a given position in the input database \( D \) to be covered if and only if there exists at least one seqlet \( S \) in \( S \) with an instance that includes the position under consideration. Again, a given position may be multiply covered by more than one seqlets. It should be clear that covering as many positions as possible in a given dataset is a much more demanding task than covering the sequences of this dataset. Since our goal is the comprehensive and complete description of the input that is processed, an extensive coverage at the amino acid position level is desirable. This is indeed the coverage measure that we have used to evaluate our results. Of course, one could completely cover any input with the 20 \( <L, W> = <1, 1> \) seqlets A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. Thus, we have to additionally make sure that the discovered seqlets carry information; the degree of information carried by a seqlet is essentially controlled by the choice of the Teiresias parameters \( L, W \) and \( K \).

**Parameter Selection.** As already mentioned, the ratio \( L/W \) controls the minimum amount of local remaining homology that a seqlet captures. A small \( L/W \) will permit the discovery of weak patterns. Selecting too large an \( L \) will ignore potentially interesting patterns involving fewer than \( L \) amino acids. On the other hand, too small an \( L \) is not informative since then the distribution of \( <L, W> \) patterns with \( L+i \) residues (for small values of \( i \)) in the input database \( D \) is basically similar to the corresponding distribution in a random database that has the same amino acid composition as \( D \). Elsewhere (17) we have shown that, for databases of the size we consider here, one begins distinguishing the compositional bias of \( D \) from a random database of similar size and composition when \( L \) is 6 or larger. We have thus chosen \( L = 6 \) for our experiments. The value of \( W \) was chosen to be equal to 15 and corresponds to minimum local similarity of 40%.\(^9\) Finally, the value of \( K \) was set equal to 2 (see also the relevant discussion on page 9). For an analysis of how the choice of \( L \) and \( W \) affects the coverage the reader is referred to (40).

**Evaluating Seqlets As Descriptors Of Structurally Conserved Regions.** Having discovered the 1D dictionary of seqlets by processing the input database, we identified their instances in the sequences of a ‘cleaned-up’ version of Release 38.0 of the Protein Data Bank (PDB). The corresponding structural fragments were subsequently aligned in 3-dimensional space and an RMSD error was computed for the set similar to the analysis in (43). As we have

---

\(^9\) Larger values of \( W \) are of course possible.
already discussed, failure to remove all but one of the near-identical copies of any given protein from the PDB will generate artificially small RMSD error values due to the over representation of the respective structural fragments over others. We first processed the PDB files excluding non-protein structures and generated a database \( F \) of amino acid sequences comprising those fragments for which 3D information was available: the sequence for a PDB entry that contained \( n \) contiguous regions with no structural information gave rise to \( n+1 \) fragments. We then processed \( F \) and removed from it all instances of identical sequence fragments except for the longest one in each case. The method was the same as the one used to remove identical sequence fragments from the GenPept input, and the same parameter choices were made here as well.

\[
- \text{Average} = 1\text{st fragment} \\
\text{while} (|\text{delta}| > \text{EPS}) \{ \\
- \text{for each fragment,} \\
\quad \text{compute alignment transformation of fragment, to Average} \\
\quad \text{using all backbone atoms;} \\
\quad \text{end for fragment,} ; \\
- \text{compute average value NewAverage of all alignments} \\
\quad \text{with Average} \\
- \text{delta} = |\text{NewAverage} - \text{Average}| \\
- \text{Average} = \text{NewAverage} \\
\} \\
- \text{RMSD} = \sum (|\text{Average} - \text{fragment}|^2) \\
- \text{report} \sqrt{\text{RMSD} / n} \text{ as the current seqlet’s RMSD} \\
\]

Figure 2. Pseudo-code for the algorithm used to compute the RMSD errors for the 3D fragment alignments.

and thus the obtained results are expected to be qualitatively similar to those that would be obtained for the entire dictionary. For each one of the sub-selected seqlets, we determined all its instances in PDBclean. We distinguish three possibilities: no instance, exactly one instance, or at least two instances in PDBclean. For those seqlets that appeared two or more times, we extracted the corresponding three-dimensional fragments, aligned them in 3 dimensions and computed the RMSD error of the resulting alignment using all backbone atoms (not only \( C_a \)). Pseudo-code for the algorithm that was used appears in Figure 2.

5. Seqlet Discovery In 1D and 3D

In this section, we present the experimental results of our seqlet discovery on NCBI’s GenPept database. This database comprised 387,451 sequences containing a grand total of 119,952,906 amino acids. It is the version of the database which existed on NCBI’s Web-site on February 10, 1999.
We first used TEIRESIAS to identify all seqlets capturing identical sequence fragments in the input; for each such seqlet we masked all of its instances except for the one appearing in the longest sequence among those containing the seqlet. This step identified 2,676,860 seqlets; after the masking, there was a grand total of 49,551,012 un-masked amino acid positions left in the input and these were processed by TEIRESIAS using the parameter settings (\(L=6\), \(W=15\), \(K=2\)). This gave rise to an additional 23,126,507 seqlets. The final collection of 25,803,367 seqlets resulted in a coverage of 117,702,444 amino acids positions or 98.12% of the original input at the amino acid level. The un-covered regions were typically of small size (3-4 amino acids) and located near either the N- or C-terminus. The Table below summarizes these results.

<table>
<thead>
<tr>
<th>Database</th>
<th>Total # Sequences</th>
<th>Total # Original Amino Acid Positions</th>
<th>% Original Amino Acid Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenPept (NCBI)</td>
<td>387,451</td>
<td>119,952,906</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Database</th>
<th>Total # Discovered Seqlets</th>
<th>Total # Covered Amino Acid Positions</th>
<th>% Covered Amino Acid Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>GenPept (NCBI)</td>
<td>25,803,367</td>
<td>117,702,444</td>
<td>98.12%</td>
</tr>
</tbody>
</table>

Table 2. The characteristics of the processed input set as well as of the discovered set of seqlets and the induced coverage.

Considering the discovered seqlets in order of decreasing frequency of appearance in the input, we plotted the induced coverage of the input’s amino acid positions as a function of the seqlets that were considered. As expected, the coverage rises very quickly early on when frequently appearing seqlets are considered, and becomes increasingly slower as the less frequent seqlets are encountered; the resulting curve is shown in Figure 3. The presence of a ‘law of diminishing returns’ is evident: although, approximately 2.7M seqlets can cover 50% of the input, to cover an additional 48.12% more than 20M seqlets are necessary.

We next report on the densities and lengths of the discovered seqlets. We define a seqlet’s length to be the number of amino acid positions it spans. We define a seqlet’s density to be the ratio of the number of non-don’t care positions in the seqlet over the seqlet’s length. Clearly, the smallest possible length is equal to the value of \(L\), whereas the smallest possible (local) density is approximately equal to \(L/W\). In Figure 4, we are showing the probability density functions and the cumulative distributions for these two variables. The vast majority of the discovered seqlets span fewer than 50 positions and have local densities equal to 50% or less.
Figure 3. Coverage of the input database as a function of the number of discovered seqlets. Seqlets are considered in order of decreasing frequency of appearance. A total of 25,803,367 seqlets can cover 98.12% of all amino acid positions in the input.

We next randomly sub-selected 3,868,719 seqlets, i.e. approximately 15% of the discovered collection of seqlets and searched PDBclean for their instances. Of the 404,127 seqlets found in PDBclean, 73,473 had two or more occurrences. In Figure 5, we show the histogram of the occurrences for the 404,127 seqlets. This histogram is providing very strong evidence that PDBclean is free from any overrepresentation bias. We next computed the RMSD error values for the 73,473 seqlets with two or more occurrences and the corresponding histogram is shown in Figure 5 (right); as can be seen here, approximately 60% of the time the computed RMSD error is below 2.5 Angstroms. This is a remarkable result when one considers that PDBclean was free of bias, and that this result has been obtained automatically from the processed data without any recurrence to biology-specific information.

Figure 4. The probability density functions and the cumulative distributions for the lengths (left) and the densities (right) of the seqlets that are discovered when processing the input database.
We should stress here that among the 330,654 seqlets with unique instances in PDBclean, there exists a subset that captures structurally conserved regions and thus the effective percentage of seqlets which carry useful structural information is much higher than 55%. However, it is not clear how to identify this subset in an automated fashion: in fact, a seqlet with multiple identical instances in PDB carries the same amount of information as a seqlet that has a unique instance in PDB. This is a question that can only be addressed incrementally as the structures of more proteins become available. And of course, incorporating domain-specific knowledge will further facilitate this task. As can be seen from the Figure, the histogram of the RMSD errors is fairly flat between 2.5 and 5.0 Angstroms, an indication that domain knowledge can be brought to bear.

![Figure 5. Left: The histogram of the seqlets’ occurrences in PDBclean. Right: The histogram of the computed RMSD error for those seqlets that had multiplicities 2 or higher (see also text).](image)

### 6. Dictionary Applications

The availability of a set of seqlets that have been derived in such an unsupervised, hierarchical manner is creating new opportunities for efficiently addressing a number of problems in computational biology. In this section we showcase several such uses: functional annotation of previously uncharacterized sequences, fast and efficient identification of local homologies, evolutionary studies, and characterization of local 3D structure.

One use is the definition of protein families through novel, specific, and sensitive seqlets that have not been previously observed, and their subsequent use in automatically carrying out functional annotation of hypothetical ORFs.\(^\text{10}\) Clearly, the characterization and definition of new protein families is possible. On Table 2 we show several seqlets that can be used as *predicates* for deciding membership in the respective families; they are contained in the bio-dictionary that we have computed by processing the input database and to the best of our knowledge they have not been reported previously.

<table>
<thead>
<tr>
<th>Seqlet</th>
<th>Protein Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>QQ[EK][GK]Q.[EK].PEQQ</td>
<td>involucrin</td>
</tr>
<tr>
<td>CRNF[DN]....P[WY]C</td>
<td>hepatocyte growth factor</td>
</tr>
<tr>
<td>[GA]..C[ED].C....R..C</td>
<td>protein kinase C</td>
</tr>
</tbody>
</table>

\(^\text{10}\)Naturally, in the absence of the ability to carry out functional assays in a wet-laboratory setting, one needs to ‘corroborate’ such annotations with multiple sequence alignments.
Table 3. Several examples of seqlets that can be used as family predicates within the GenPept database.

Although descriptors such as shown on this Table are very specific for a given family, they may not necessarily be sensitive enough to capture all the members of the corresponding family. In this case, use of composite descriptors (35,24,49,19,20) is more appropriate. More recently, we imbedded the use of composite descriptors in a Bayesian framework (41).

Figure 6. Using the seqlet D.....DLKT[ML]D.G.[VIL]E.[ML]A....A.V.I.G[VL][GA]...[ST]I as an anchor point to discover sequence homology. Also shown are the domains that can be identified in the sequences that are involved (see also text).

Even when a seqlet is not specific enough to functionally characterize a protein, it can still provide an anchor point for identifying local homologies. One of the seqlets in our discovered bio-dictionary is D.....DLKT[ML]D.G.[VIL]E.[ML]A....A.V.I.G[VL][GA]...[ST]I and appears only 4 times in the sequence database: MTH1474 (gi|2622589 / 430 a.a. / Methanobacterium thermoautotrophicum), AF1305 (gi|2649276 / 394 a.a. / Archaeoglobus fulgidus), MJ1447 (gi|1592092 / 381 a.a. / Methanococcus janaschii), and PH1938 (gi|3258382 / 406a.a. / Pyrococcus horikoshii). This seqlet allows to discover the approximately 200 amino acid long homology that involves the C-termini of MTH1474, AF1305, MJ1447 and the N-terminus of PH1938. The actual arrangement as well as the region captured by the seqlet is shown in Figure 6. This case also helps us highlight two additional points. First, functionally important entities have been reused and fused during evolution: HUMS_BACSU, an experimentally characterized D-arabino 3-hexulose 6-phosphate formaldehyde lyase, has been fused with two different components to give rise to distinct proteins: (a) the family comprising MTH1474, AF1305, and MJ1447, and (b) the protein PH1938.11 Second, wrong annotations can be propagated if the necessary care is not exercised: all four proteins, i.e. MTH1474, AF1305, MJ1447 and PH1938, appear annotated in GenPept as D-arabino 3-hexulose

11 Moreover, the second component (190 a.a. fragment) of PH1938 is itself the single component of the family comprising the orthologs MJ1247, AF1796, MTH249 and MTH1546.
6-phosphate formaldehyde lyases. However, these annotations may not be correct; the only thing that can be said of these sequences is that their N- or C-terminus is similar to the D-arabino 3-hexulose 6-phosphate formaldehyde lyase HUMS_BACSU (Figure 6).

But it is not only the sensitivity in homology identification that the seqlets can afford but also the speed. In fact we conjecture that as more and more sequences become publicly available, the bio-dictionary will eventually saturate and comprise seqlets that cover the entire space of natural sequences. At that point, homology searches will correspond to searches through the fixed size dictionary of seqlets instead of the ever increasing sequence databases. Currently, the accumulated databases provide a sampling which may correspond to a proper subset of the actual sequence space: unless more sequences and complete genomes become available, it is difficult to state with certainty whether this is the case. Because of its importance, the topic of using seqlets as anchor points for discovering sequence homologies in biological databases as well as details on building search engines is being discussed at length (17).

An additional use of seqlets is in the analysis of phylogenetic distribution (37). Recall that by processing the input database as a whole we have in essence compared sequences that would have not otherwise been studied together. The seqlets not only allow us to cut across families but also across the three phylogenetic domains. By augmenting the sequence label in the input database with one of the 7 tags: ‘Archaea,’ ‘Bacteria,’ ‘Eukaryota,’ ‘Other,’ ‘Unclassified,’ ‘Viroids,’ or ‘Viruses’, answering questions such as

- what is the set of motifs that is unique to and characterizes a given phylogenetic domain?
- which motifs are shared by exactly two phylogenetic domains?
- which motifs are universal?

amounts to determining in which sequences a candidate seqlet can be found. And the fact that the bio-dictionary provides exhaustive and complete coverage of the sequence space without applying any statistical filtering allows us to provide exhaustive answers to such questions. We next present examples of such domain-specific seqlets that were discovered during the processing of the input database.

An Archaea-specific seqlet is [MLIV][MLIV][KR][MLIV][DE][DE][MLIV][MLIV][GA][GA]. This seqlet is present in the archaeal family of chaperonins; chaperonins are a special kind of chaperones which are double-ring, oligomeric containers and provide closed compartments that shield folding proteins from the cellular environment. Instances of this particular seqlet are present in the 11 archaeal species: Archaeoglobus fulgidus, Desulfurococcus, Haloferax volcanii, Methanobacterium thermoautotrophicum, Methanococcus jannaschii, Methanococcus thermolithotrophicus, Methanopyrus kandleri, Pyrococcus horikoshii, Pyrococcus sp., and Thermococcus sp.

A Bacteria-specific seqlet is [FY][MLIV][GA][DE][MLIV][FY][GA][GA][P][MLIV][GA]. It is found in cytochrome D oxidase proteins and its 16 instances appear in 7 bacterial species: Azotobacter vinelandii, Bacillus subtilis, Escherichia coli, Haemophilus influenzae, Klebsiella pneumoniae, Mycobacterium tuberculosis, and Streptomyces coelicolor.

The seqlet [MLIV][P][MLIV][GA][ST][GA][GA][GA][GA] is an example of a seqlet that is present in two phylogenetic domains, those of Archaea and Bacteria. It is contained in imidazole glycerol phosphate synthases identified in 12 archaeal and bacterial species: Aquifex aeolicus, Archaeoglobus fulgidus, Bacillus subtilis, Buchnera aphidicola, Escherichia coli, Haemophilus influenzae, Klebsiella pneumoniae, Mycobacterium capsulatus, Mycobacterium tuberculosis, Salmonella typhimurium, Sulfolobus solfataricus, and Synechocystis.
Finally, the seqlet P[GA][DE][QN][MLIV][GA][FY][QN][MLIV][KR][GA][MLIV] is present in the archaeal and eukaryotic phylogenetic domains: It is found in elongation factors (EF-Tu and EF-1a) from archaeal and eukaryotic species: Archaeoglobus fulgidus, Desulfurococcus mobilis, Dinenympha exilis, Diplomonad ATCC50330, Giardia intestinalis, Haloarcula marismortui, Halobacterium halobium, Hexamita inflata, Methanobacterium thermoautotrophicum, Methanococcus jannaschii, Methanococcus vannielii, Pyrococcus horikoshii, Pyrococcus woesei, Spironucleus vortens, Sulfolobus acidocaldarius, Sulfolobus solfataricus, Telotrochidium henneguyii, Tetrahymena pyriformis, Thermococcus celer, and Thermoplasma acidophilum.

Obviously, the biases here are strong since all of the above mentioned proteins may have only been sequenced in different species and for a variety of reasons.

Figure 7. Examples of supersecondary structures captured by seqlets. Left: a β−α−β structure captured by the seqlet [DI]...G[GA]G...G[MLIV]A.A...A.LG.[KR]...[MLIV]...K and shared by two difference enzymes (1GRG / E.C. 1.6.4.2 and 2TSY_B / E.C. 4.2.1.20). Right: in this example, [KR][MLIV],[MLIV],[GA],[GA]...[GA]...[GA]...[MLIV]....[ST].[DE][MLIV],[MLIV][MLIV][DE]...[DE] is the seqlet which captures this β−α−β structure that is shared by 1FCD_A and 5LDH. See also text for more details.

We conclude this section with the discussion on the uses of the seqlets in the context of 3D-structure characterization and prediction. As we have already seen, the seqlets of the bio-dictionary that we have compiled also capture 3-dimensional information. In Figure 1, we can see that the seqlets can capture basic secondary structure elements such as beta-hairpin, alpha-helices, and turns.

But seqlets can in fact capture super-secondary structures equally effectively. In Figure 7 (left) we show the 3-dimensional structure-fragment that is shared by two distinct enzymes; the shown structure is captured by the seqlet [DI]...G[GA]G...G[MLIV]A.A...A.LG.[KR]...[MLIV]...K. In PDBclean, the seqlet appears in 1GRG and 2TSY_B. In 1GRG, a gluthione reductase from Homo sapiens (E.C. 1.6.4.2) the seqlet spans amino acid positions 22 through 53 (DYLVI\text{GGG}SGL\text{ASARRAELGR\text{AAVVE}}H\text{K}). In 2TSY_B, a tryptophan synthase from Salmonella typhimurium (E.C. 4.2.1.20), the seqlet is found on the B chain between amino acid positions 106 and 137 (I\text{IAET-GAGQHGVASALASALLGLKRIYMGA}).

[KR][MLIV],[MLIV],[GA],[GA]...[GA]...[GA]...[MLIV]....[ST].[DE][MLIV],[MLIV][MLIV][MLIV][DE]...[DE] is the seqlet corresponding to our second example and the corresponding structure is shown in Figure 7 (right). In
PDBclean the seqlet appears in 1FCD_A and 5LDH. In 1FCD_A, a flavocytochrome c sulfide dehydrogenase from *Chromatium vinosum*, the seqlet spans amino acid positions 4 through 38 (KVVVVGGGTGGGATAAKYKLADPSIEVT-LIEPNTD). In 5LDH, a lactate dehydrogenase from *Sus scrofa* (E.C. 1.1.1.27) the seqlet spans positions 23 through 57 (KITVVGVQGMAISILGKSLDELALVDVLED).

And similarly to the case of sequence homology, the seqlets can also be used as anchors to derive extended structural homology: beginning with seqlets that correspond to low RMSD errors, and thus capture structurally conserved blocks, we can extend the seqlet outward by allowing the *sequences* containing the seqlet to ‘decide’ how far to go. Frequently, the extended sequence fragment corresponds to an extended structural agreement. One such example is the *helix-turn-helix* shown in Figure 8. The seqlet that gave rise to it is E.A.R..G.S.P and it is shared by the sequences 2EZI and 1PYS of PDBclean. 2EZI is the Mu end of the DNA binding domain of *Mu* phage transposase; 1PYS is a phenylalanyl-tRNA synthetase from *Thermus thermophilus*. Treating the seqlet as an anchor point we can extend it outward as follows (the positions captured by the seqlet are shown in boldface):  

>2EZI 206-RLEIAREHGSISPRAITA(RI)QQLD-232  
>1PYS 407-RPaYANRLGTSIP-EAEQIAILRKLGL-432  

The homology can be extended to the left and mainly to the right of the seqlet and spans the positions 206-232 in 2EZI and 407-432 in 1PYS. The structures that correspond to this last homology are shown in Figure 8.

Figure 8. Using the seqlet as an anchor point at the *sequence* level, we can derive extended *structural* agreements (see also text).

It is expected that pairing the seqlets of the bio-dictionary with the 3D fragments that they capture will allow us to generate good characterizations of local 3D structure directly from the 1D amino acid sequence. Furthermore, and assuming that we can tessellate any given sequence in its entirety with structurally characterized seqlets, it is in principle possible to provide coarse estimates of the *global* 3D structure by combining the structures of the individual seqlets (In Preparation).

7. Conclusion

We have used a pattern discovery algorithm to process the GenPept database as a whole and explore the sequence space of natural proteins. We have built a bio-dictionary of reusable elements referred to as seqlets that can account for and cover 98.12% of all amino acid positions in the processed database. The seqlets have been shown to capture both functional and structural information as well as similarity signals across families. The
availability of such a dictionary has numerous applications in computational biology and several such applications have been discussed in detail.

8. Acknowledgments

We would like to thank Dan Platt for his help with the code which computes the RMSD error. We would also like to thank Dr. Nikos Kyrpides for his help with the homology example and his invaluable insight to the problem of functional annotation. Finally, our thanks go to the anonymous reviewers whose comments helped to significantly improve the presentation in this manuscript.

References


