

# Building Dictionaries Of 1D and 3D Motifs By Mining The *Unaligned* 1D Sequences Of 17 Archaeal and Bacterial Genomes

Isidore Rigoutsos<sup>1</sup> Yuan Gao<sup>2</sup> Aris Floratos Laxmi Parida

Bioinformatics and Pattern Discovery Group, Computational Biology Center, IBM TJ Watson Research Center,  
PO BOX 704, Yorktown Heights, NY 10598. <sup>2</sup> Also: Dept. Of Mathematical Sciences, The University of Memphis,  
Memphis, TN 38138. <sup>1</sup> Corresponding Author: Email: rigoutso@us.ibm.com

## Abstract

We have used the *TEIRESIAS* algorithm to carry out unsupervised pattern discovery in a database containing the *unaligned* ORFs from the 17 publicly available complete archaeal and bacterial genomes and build a 1D dictionary of motifs. These motifs which we refer to as *seqlets* account for and cover 97.88% of this genomic input at the level of amino acid positions. Each of the *seqlets* in this 1D dictionary was located among the sequences in Release 38.0 of the Protein Data Bank and the structural fragments corresponding to each *seqlet*'s instances were identified and aligned in three dimensions: those of the *seqlets* that resulted in RMSD errors below a pre-selected threshold of 2.5 Angstroms were entered in a 3D dictionary of structurally conserved *seqlets*. These two dictionaries can be thought of as cross-indices that facilitate the tackling of tasks such as automated functional annotation of genomic sequences, local homology identification, local structure characterization, comparative genomics, etc.

## Introduction

To the extent that it can be deduced from the sampling provided by the currently available amino acid sequences, the sequence space of natural proteins is sparsely populated.

Molecular families are believed to be evolving through random shift and natural selection processes. Within a family, its members typically share family-specific elements in the form of conserved amino acids. Such family specific elements have been referred to as *patterns* or *motifs* (Hodgman 1989) and can be used to describe amino acid segments of biological importance (e.g. domains, folds) (Bork and Gibson 1996; Doolittle 1995; Saraste et al. 1990).

A lot of methods and tools that discover and exploit such sequence descriptors in the form of patterns, profiles, or Hidden Markov Models (HMMs) have been proposed in the literature; these include PROSITE (Bairoch et al. 1997), BLOCKS (Henikoff and Henikoff 1996), PFAM (Sonnhammer et al. 1997), PRINTS (Attwood et al. 1998), to name a few.

Currently, the pattern determination process is based upon the identification of similar proteins, the subsequent generation of multiple alignments and, finally, the selection of the most conserved regions as significant

signatures. The difficulty with this general methodology is that it proceeds on a case by case basis (Smith and Smith 1990; Nevill-Manning et al. 1998) and is based on an underlying assumption that patterns can be found only within divergent families (Ogiwara et al. 1992). The use of alignment to identify convergently-related functional motifs such as nuclear localization signals (Boulikas 1993) is rather difficult (Horton and Nakai 1997) and the generated patterns are usually not sufficiently specific.

The *TEIRESIAS* algorithm (Rigoutsos and Floratos 1998a; Rigoutsos and Floratos 1998b) has made possible a hierarchical motif discovery and enumeration of the most frequent patterns by treating large datasets such as Swiss-Prot and NCBI's Non-Redundant database (Rigoutsos et al. 1998; Rigoutsos et al. 1999) as a whole. No subsets of sequences are formed and there is no need for any alignment prior to the discovery stage. This discovery-based approach is distinctly different from methods that first use a sequence homology search algorithm such as BLAST (Altschul et al. 1990) to form groups of homologous sequences from which motifs are subsequently derived (Harris et al. 1992; Tatusov et al. 1997; Yona et al. 1998).

Hierarchical motif discovery in such large datasets of diverse sequence composition is expected to (a) uncover previously unobserved protein features within and across family boundaries, (b) to shed light on relationships between families that have traditionally been assumed to be unrelated, and (c) to enhance our understanding of protein architecture.

We have used these patterns to derive a natural *dictionary* of re-usable elements with uses which include automated functional annotation, sensitive homology detection (Floratos et al. 1999), the analysis of phylogenetic distribution and local structure characterization (Rigoutsos et al. 1999). We use the term *seqlets* to refer to these re-usable elements.

A large number of complete archaeal and bacterial genomes has recently become available thus generating a need for producing such dictionaries through unsupervised motif discovery. Given that at the time that this work was carried out only one complete eukaryotic genome was available, we decided to not include any representative

from this phylogenetic domain so as to avoid skewing the overall composition of the database even further. The complete genome of *Caenorhabditis elegans* was reported while the experiments outlined in this work were almost completed.

We compiled a database containing the ORFs for the complete genomes of 13 Bacteria (*Escherichia coli*, *Haemophilus influenzae*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Synechocystis sp.*; *Aquifex aeolicus*, *Chlamydia trachomatis*, *Helicobacter pylori*, *Rickettsia prowazekii*, *Bacillus subtilis*, *Mycobacterium tuberculosis*, *Treponema pallidum* and *Borrelia burgdorferi*) and 4 Archaea (*Methanococcus jannaschii*, *Archaeoglobus fulgidus*, *Pyrococcus horikoshii*, and *Methanobacterium thermoautotrophicum*) (Fraser et al. 1997; Fraser et al. 1998; Andersson et al. 1998; Kawarabayasi et al. 1998; Cole et al. 1998; Smith et al. 1997; Tomb et al. 1997; Stephens et al. 1998; Kunst et al. 1997; Klenk et al. 1997; Deckert et al. 1998; Fleischmann et al. 1995; Kaneko et al. 1997; Fraser et al. 1995; Himmelreich et al. 1996; Bult et al. 1996; Blattner et al. 1997).

This database of unaligned sequences was processed with *TEIRESIAS* and produced a comprehensive 1D dictionary of seqlets that, for all practical purposes, describes the sequence space of these genomes completely. All instances of these generated seqlets were subsequently identified in a free from overrepresentation bias version of Release 38.0 of the Protein Data Bank (Bernstein et al. 1977; Abola et al. 1987) and their respective fragments were aligned in three-dimensional space. For those fragment-sets that gave rise to RMSD errors below the preset threshold of 2.5 Angstroms, the respective seqlets were associated with an *average fragment structure* and incorporated in a 3D dictionary of structural motifs.

We report on the properties of the entries of these two dictionaries, the extent to which the 1D seqlets have instances in the Protein Data Bank, and the ramifications from the induced coverage. We also present several entries of the generated dictionaries and discuss uses.

The next section introduces several terms and notation that will be used throughout the discussion and also briefly outlines the properties of the used algorithm. Following that the various methods are described. Then we give details on the properties of the discovered dictionaries and show and discuss several of the seqlets.

## Terms and Background Information

In this discussion,  $\Sigma$  will denote 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 define a *seqlet* to be a string of the form

$$(\Sigma \cup \{\Sigma \Sigma^* \Sigma\}) (\Sigma \cup \{.\} \cup \{\Sigma \Sigma^* \Sigma\})^* (\Sigma \cup \{\Sigma \Sigma^* \Sigma\}) \cup \Sigma [1]$$

i.e. either 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 character ‘.’ is referred to as ‘don’t care’ and is a wild card that can replace any character from  $\Sigma$ . A bracketed expression, e.g. [GA], means a ‘one-of choice’ and in this example exactly one of G or A.

Seqlets of the above form can capture homologies in the traditional PAM or Blossum sense but also allow for amino acid substitutions not provided for by such scoring matrices.

A seqlet  $S$  is in fact a regular expression and as such it defines a language  $G(S)$ . The elements of the language are all the *strings* that can be obtained from  $S$  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 pattern “G..GSGK[ST]T” defines the language  $G(“G..GSGK[ST]T”)$  the members of which include the strings GPTGSGKST, GYLGSGKST, and GESGSGKTT among others.

A seqlet  $S$  is called an  $\langle L, W \rangle$  seqlet (with  $L \leq W$ ) if and only if every substring of  $S$  with length  $W$  contains  $L$  or more *non-don’t care* characters. A given choice for the parameters  $L$  and  $W$  has a direct bearing on the degree of remaining homology among the instances of the domain that the seqlet captures: the *smaller* the value of the ratio  $L / W$  the *lower* the degree of allowed local homology within any stretch of  $W$  amino acids (and, consequently, the higher the sensitivity of the pattern discovery process).

When processing a database  $D$  of proteins, and for a given choice for  $L$  and  $W$ , *TEIRESIAS* discovers all seqlets  $S$  which appear  $K$  or more times and additionally guarantees that they have the *maximality* property with respect to  $D$ : i.e. no reported seqlet  $S$  can be made more specific (either by appending/prepending a string on  $\Sigma$ , or by dereferencing one or more don’t care characters) without simultaneously reducing the number of locations within  $D$  where the derived result can appear).

In recent work (Rigoutsos et al. 1998; Rigoutsos et al. 1999), we described the importance of the unsupervised hierarchical discovery of seqlets beginning with large databases of *unaligned* sequences, and also addressed issues such as the specificity, sensitivity and clustering of seqlets as well as various applications. As defined above and with the maximality properties guaranteed by algorithm, seqlets can capture family-specific functional elements as well as elements of a more elementary and reusable nature. In fact, the algorithm’s ability to identify very weak patterns has proven useful in identifying motifs spanning multiple protein families.

## Methods

**Masking Of Identical Sequence Fragments Present In The Input.** Prior to discovering the seqlets in the 17 genome database we process it and mask all but the first instance of all identical amino acid fragments that are present in it. The natural language equivalent to this masking is the removal of all quasi-identical phrases (i.e. amino acid fragments) present in the sentences (i.e. ORFs) to be processed; clearly, multiple copies of such fragments do not contribute anything to the dictionaries that are built. To this end, we have used *TEIRESIAS* with a setting of  $L=8$ ,  $W=8$  and  $K=2$ . Note that conceptually similar choices have been made in (Holm and Sander 1998) which also describes an alternative method that could be used in this stage. All of the discovered seqlets were collected and their instances in the input were masked by replacing them with symbols that do not belong to the alphabet  $\Sigma$ .

**“Covering” The Input.** An important *quality* criterion for the set  $\mathcal{S}$  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. 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  $\mathcal{S}$  with an instance in the sequence  $P$ . Clearly, a given sequence may be multiply covered by more than one seqlets. In an analogous manner, 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  $\mathcal{S}$  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 sought. As described below, this is the coverage measure that we have used in our experiments. We have additionally made certain that the discovered seqlets are statistically significant: the statistical significance of a seqlet is essentially controlled by the choice of the *TEIRESIAS* parameters  $L$ ,  $W$  and  $K$ .

**1D Dictionary / Selecting The Various Parameters.** The amount of information that a seqlet carries is essentially controlled by the values of the parameters  $L$ ,  $W$  and  $K$ . As already mentioned, the ratio  $L/W$  controls the *minimum* amount of local remaining homology that a seqlet captures. A small  $L/W$  ratio will permit the discovery of weak patterns. The numerator cannot be too small (e.g.

$L=2$ ) because then the seqlets would not be specific enough. And it cannot be too large either (e.g.  $L=10$ ) since it would force us to ignore potentially interesting patterns that comprise fewer than  $L$  amino acids. An appropriate choice for  $L$  would be one that would make the distribution of  $\langle L, W \rangle$  patterns with  $L+i$  residues (and small values of  $i$ ) in the input data base  $D$  different than the corresponding distribution in a *random* database that has the same size and amino acid composition as  $D$ .

In related work (Floratos et al. 1999) 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 5 or larger. We have thus chosen  $L=6$  for our experiments. The value of  $W$  was chosen to be equal to 15 thus allowing us to capture minimum *local* homologies of 40%. Finally, the value of  $K$  was set equal to 2.

At first, it may seem that  $K=2$  is too low a threshold for statistical significance. It should be pointed out that if a seqlet consists of many alphabet symbols and bracketed expressions it can be significant even if it appears exactly twice in  $D$ . The majority of seqlets appearing exactly 2 times in the dataset we are considering consists of an average of 10 solid alphabet symbols: we have carried out Monte-Carlo simulations which indicated that such seqlets are very infrequent in random databases with size and composition similar to that of the 17 genomes. We conjecture that this is a consequence of the high average diversity encountered in the ORFs of these genomes and we are in the process of resolving this question through additional experiments and simulation.

**The Database.** The database that we used as the input to *TEIRESIAS* comprised the ORFs from the 17 complete and publicly available archaeal and bacterial genomes. In particular, the database contained: 4289 ORFs from *Escherichia coli* with 1358990 a.a.; 1709 ORFs from *Haemophilus influenzae* with 521077 a.a.; 480 ORFs from *Mycoplasma genitalium* with 174959 a.a.; 1715 ORFs from *Methanococcus jannaschii* with 483564 a.a.; 677 ORFs from *Mycoplasma pneumoniae* with 237651 a.a.; 3169 ORFs from *Synechocystis* sp. with 1033205 a.a.; 1522 ORFs from *Aquifex aeolicus* with 482512 a.a.; 2407 ORFs from *Archaeoglobus fulgidus* with 662866 a.a.; 4099 ORFs from *Bacillus subtilis* with 1216678 a.a.; 894 ORFs from *Chlamydia trachomatis* with 312553 a.a.; 1565 ORFs from *Helicobacter pylori* with 496448 a.a.; 1869 ORFs from *Methanobacterium thermoautotrophicum* with 525507 a.a.; 3918 ORFs from *Mycobacterium tuberculosis* with 1329251 a.a.; 2064 ORFs from *Pyrococcus horikoshii* with 568544 a.a.; 834 ORFs from *Rickettsia prowazekii* with 279080 a.a.; 1031 ORFs from *Trypanosoma pallidum* with 350676 a.a.; and 850 ORFs from *Borrelia burgdorferi* with 283312 a.a. I.e. 33,092 ORFs with a

grand total of 10,316,873 amino acids. Our input database and annotations were composed of the contents of NCBI's Web-site of complete genomes as they existed on December 20, 1998.

**3D Dictionary / Intersecting The Seqlets With The PDB.** Having discovered the 1D dictionary of seqlets by processing the ORFs of the 17 genomes, we proceeded and identified each one of them in the sequences of a *cleaned up* version of Release 38.0 of the Protein Data Bank (PDB). In other words, we treated each of the sequences in the cleaned up PDB as a *query* and determined which of our 1D dictionary entries are present in it. For those seqlets that appeared at least twice, we extracted the corresponding structure fragments, aligned them in 3 dimensions and computed the RMSD error of the resulting alignment. It should be clear that all the families which are *over*-represented in the PDB would generate artificially low RMSD error values for all the fragment sets in which their members participated. In fact, by artificially increasing the number of fragments to be aligned they would generate *lower* average values for the respective seqlet's RMSD error. We have thus compiled a cleaned-up version of the PDB which is free from the said overrepresentation bias. By operating on the protein sequences alone, we removed all the identical sequence fragments using the same procedure as the one used with the genomic input. The parameters were chosen to be  $L=6$ ,  $W=6$  and  $K=2$ . This procedure produced the cleaned up PDB which was used in the experiments and whose size was only 1/5th of the original PDB (540K a.a. instead of the original 2,8M a.a.). The algorithm which was used in conjunction with the cleaned up PDB to compute the RMSD error of each fragment set was:

```

- Average = 1st fragment
  while ( | delta | > EPS ) {
    - for each fragmenti
      compute alignment transformation of
      fragmenti to Average using all backbone atoms;
    end for fragmenti ;
    - compute average value NewAverage of
    all alignments with Average
    - delta = | NewAverage - Average |
    - Average = NewAverage
  }
- RMSD =  $\sum_i ( | Average - fragment_i |^2 )$ 
- report sqrt (RMSD / n) as the current seqlet's RMSD

```

Those of the seqlets that resulted in RMSD error values of 2.5 Angstroms or less were entered in a 3D Dictionary of seqlets together with the alignment of the respective structure fragments.

## Experimental Results

We next describe the results from processing the 17 genome database and give details as to the composition and contents of the resulting 1D and 3D dictionaries.

### The 1D Dictionary

During a first pass over the input, identical sequence fragments were masked as described previously. At the end of this step 74,552 patterns were discovered which covered a grand total of 1,368,687 amino acid positions, or 13.1% of the original input. After masking these positions with symbols not in  $\Sigma$ , *TEIRESIAS* was used again. Eventually a total of 2,138,771 patterns were discovered. These patterns covered 10,227,694 of the 10,316,873 amino acid positions, or 97.88% of the entire genomic input.

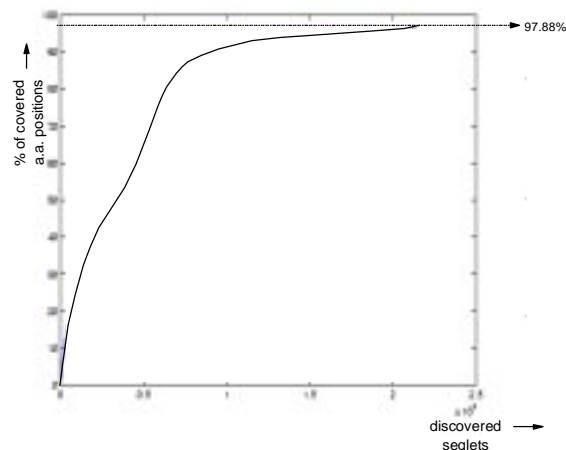


Figure 1. Coverage of the genomic input as a function of the number of discovered seqlets. Seqlets are considered in order of decreasing frequency of appearance. A total of 2,138,771 seqlets can cover 97.88% of all amino acid positions in the input.

The patterns were discovered and reported in order of decreasing frequency of appearance. In Figure 1, we are showing the induced percentage of position coverage as a function of the number of discovered seqlets: the seqlets have been considered in the order in which they are reported. As is evident from this graph, a 'law of diminishing returns' is in effect: although roughly 90.00% of all amino acid positions can be covered with approximately 1,000,000 seqlets, to cover an additional 7.88% one needs more than 1,000,000 seqlets.

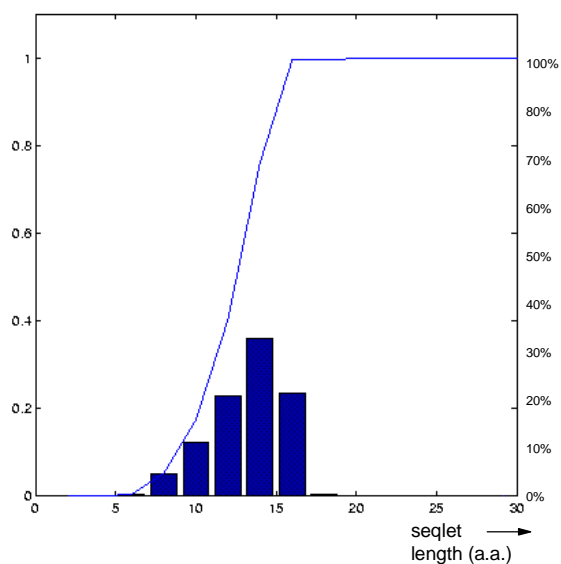


Figure 2. The probability density function and cumulative distribution for the lengths of the seqlets that are discovered when processing the input database.

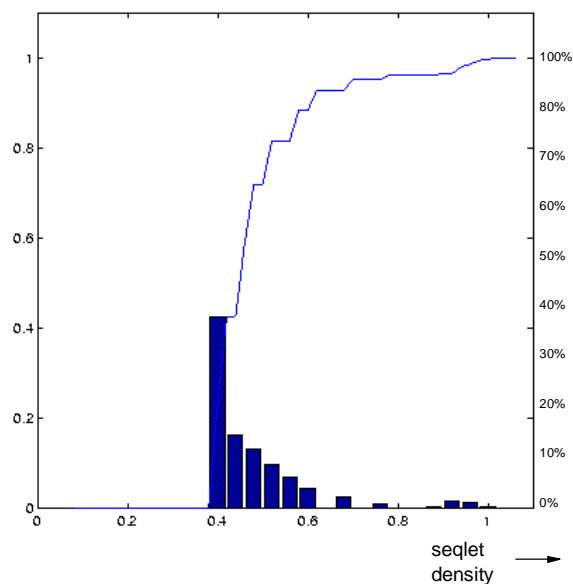


Figure 3. The probability density function and cumulative distribution for the densities of the seqlets that are discovered when processing the input database.

The *densities* and *lengths* of the discovered seqlets vary substantially. 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 Figures 2 and 3, we are showing the probability density functions and cumulative distributions for these two variables. The longest discovered seqlet spans 957 amino acid positions and corresponds to the ORFs HP0488 and HP1116 of *Helicobacter pylori*.

### The 3D Dictionary

As described above, and after having generated the comprehensive 1D dictionary of seqlets, we proceeded and identified instances of these seqlets in the free from bias version of Release 38.0 of the PDB. Of the discovered 2,138,771 seqlets, 161,641 seqlets have at least one instance in the cleaned up PDB and cover 67.98% of all the amino acid positions in it.

It is interesting to note for comparison purposes that an analogous 1D dictionary of seqlets derived from processing Release 36.0 of Swiss-Prot (a database whose contents are substantially different from the genomic input we processed here) induces an amino acid position coverage of the cleaned up PDB that equals 91.97%. This result suggests that archaeal and bacterial proteins are underrepresented in the PDB database.

Of the 161,641 seqlets that occur in the cleaned up PDB only 24,966 have two or more instances in it. This is indeed providing further proof that the cleaned up PDB is free from overrepresentation bias. But at the same time it is an obstacle since very little can be said of the structural information carried by the seqlets with single instance in the cleaned up PDB. Obviously, this situation will be resolved incrementally as more and more structures become available.

For the seqlets with two more instances in the cleaned up PDB, we computed the RMSD error of the respective fragment sets. Figure 4 shows a scatter-plot for these errors as a function of the seqlet length and density; the plot is similar to the one in (Sander and Sneider 1991). The majority of the points are in the region that the dashed line surrounds. Integrating out the length and density components gives the probability density function and cumulative distribution for the resulting RMSD error (Figure 5).

As can be seen, approximately 70.00% of the fragment groups corresponding to seqlets with two or more instances in the cleaned up PDB have RMSD error values that are less or equal to 2.5 Angstroms. This result is remarkable since it is derived from the cleaned up PDB. Also, it provides further support to our hypothesis that 1D motifs that are derived by processing large sequence databases have relevance in 3 dimensions and correspond to conserved local structures.

## Examples Of Entries And Their Use

The availability of 1D and 3D dictionaries that have been derived in such an unsupervised, hierarchical manner is creating new opportunities for efficiently addressing a number of problems in computational biology.

We have found that the discovered seqlets belong to three categories: a given seqlet is specific for a family; several distinct seqlets are specific for a family; or, a given seqlet is present in several distinct families.

Clearly, seqlets belonging to the first two categories can be associated with functional categories and used as membership predicates for the respective families. It is seqlets that fall in these categories that are particularly useful in functionally annotating ORFs of unknown function and “orphans.” As for the seqlets belonging to the third category, these are useful in identifying local homologies across proteins of different functions. We describe at length the use of seqlets for determining local homologies in (Floratos et al. 1999).

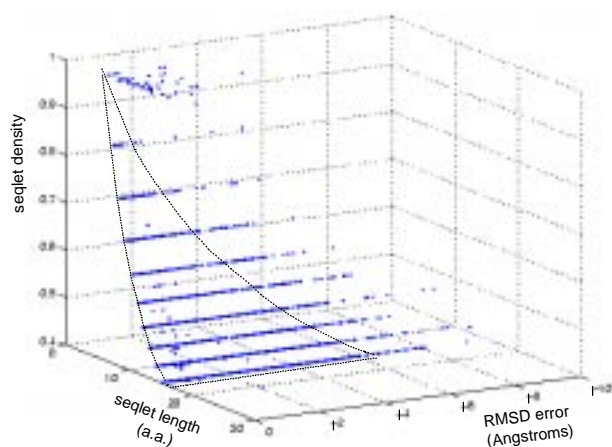


Figure 4. The distribution of the RMSD errors for the 24,966 seqlets that have two or more instances in the free from bias version of the PDB. See also text.

We next give an example of a seqlet that was used effectively to annotate several ORFs with previously unknown function. In what follows, we are making the assumption that the only available information is the annotation as it existed the day that the respective genome was submitted to GenBank.

We are aware that functional annotation of such ORFs is at the center of the work carried out by several independent research groups, and it is possible that the sequences shown below have been annotated in the meantime. The point behind these examples is not to

report the resulting annotations as novel but to demonstrate the ability of the entries in our 1D dictionary of seqlets to correctly annotate ORFs.

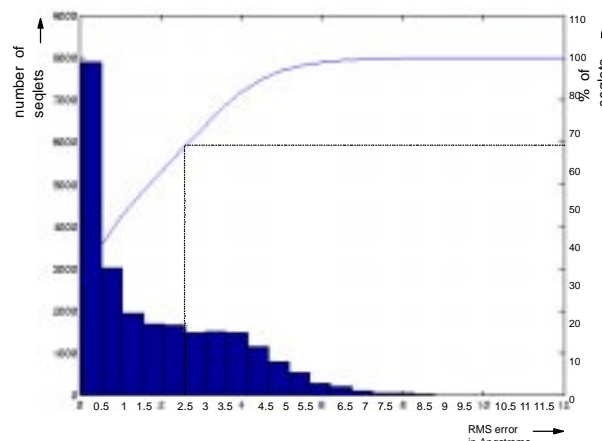


Figure 5. The histogram of the RMSD errors for the 24,966 seqlets that have two or more instances in the cleaned-up PDB. Also shown is the cumulative distribution as a function of the RMSD error value. See also text.

One of the seqlets that are discovered when we process the input database is [TFY][FY][VIS].[NDT].[NQY].[NY][FYP]TN[IVF]C...C.FC[AS]F and is present in the following ten ORFs: gi\_1500312, gi\_1651900, gi\_2983258, gi\_2983355, gi\_3328856, gi\_3329230, gi\_2650237, gi\_2313781, gi\_2621911, and gi\_2695957. Of these, gi\_3328856 and gi\_3329230 are annotated as Fe-S oxidoreductases. In Figure 8, we are showing the clustalw alignment for these 10 sequences which further supports the argument that the remaining 8 of the ORFs are Fe-S oxidoreductases. The score for the shown alignment was 32032.

Besides functional annotation, many of the seqlets in our 1D dictionary have also helped us define families of proteins. One such example is the seqlet E...[KN].....L.[YF]E....L.....F.[KR]L..[ED]..[KR]..[VI] [ED]E which occurs only four times in the processed database. The sequences that contain it are AF1530 from *Archaeoglobus fulgidus*, MJ0039 from *Methanococcus janaschii*, MTH1324 from *Methanobacterium Thermoautotrophicum* and PH1127 from *Pyrococcus Horikoshii*. A multiple sequence alignment produced using the algorithm described in (Parida, Floratos and Rigoutsos 1998) is shown in Figure 9: the homology is evident. Note that all four sequences come from the archaeal phylogenetic domain; moreover, a search of the GenPept database with this seqlet does not retrieve any more sequences. It is thus

likely that the seqlet in question is in fact a predicate for an Archaea-specific protein family.

As a matter of fact, this last example helps to point out one more use for the 1D dictionary, that of the analysis of phylogenetic distribution. Once additional complete genomes become available that will provide a more balanced sampling for the three phylogenetic domains, one can rebuild the 1D dictionary and use its cross-indexing capability to find seqlets that are specific to a single domain, shared by exactly two domains, or are universal.

Finally, we turn our attention to the entries of the 3D Dictionary, i.e. to those of the seqlets whose instances in the cleaned up PDB resulted in RMSD error values of 2.5 Angstroms or less. Two such entries are:

- Y.V...T.DG...I; this seqlet is present in 1FXI and 1DOX. The resulting RMSD value was equal to 2.28 Angstroms, and the alignment is shown in Figure 6.
- A..PA.AA.....A; this seqlet is present in 2CCY and 1REQ. The computed RMSD is equal to 0.94 Angstroms and the alignment is shown in Figure 7.

The seqlet Y.V...T.DG...I is an example of structural conservation within a family: it is present in ferredoxins from two different cyanobacteria, *Apanothece sacrum* (1FXI) and *Synechocystis* sp. (1DOX). The seqlet A..PA.AA.....A is a representative example of a 3D dictionary entry that captures a structurally conserved region present in proteins of *different function*. Indeed, 2CCY is a cytochrome c from *Rhodospirillum molischiannum*, whereas 1REQ is a methylmalonyl-CoA mutase from *Propionibacterium freudenreichii*.

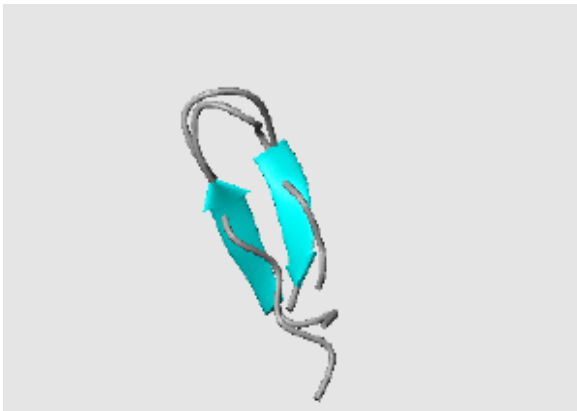


Figure 6. The structure captured by the seqlet Y.V...T.DG...I. The RMSD error of the shown alignment was 2.28 Angstroms. See also text for more details.



Figure 7. The structure captured by A..PA.AA.....A. The RMSD errors of this alignment was 0.94 Angstroms respectively. See also text for more details.

We expect that comprehensive 3D dictionaries built as described above will contribute to the ongoing efforts to tackle the problem of protein folding by allowing us to decompose any given sequence into a union of seqlets with a known and well-characterized local structure in 3 dimensions. Clearly, comprehensive dictionaries that cover as much as possible of the already sampled sequence and structure space are only a first step in that direction.

## Conclusion

We have used *TEIRESIAS* to carry out unsupervised pattern discovery in a database containing the *unaligned* ORFs from 17 publicly available complete archaeal and bacterial genomes. We have built a 1D dictionary of seqlets which accounts for and covers 97.88% of this genomic input at the level of amino acid positions. We identified instances of these seqlets in a free from bias version of Release 38.0 of the Protein Data Bank and aligned the respective 3D fragments: those of the seqlets that resulted in RMSD errors below a pre-selected threshold of 2.5 Angstroms were entered together with their average structure in a 3D dictionary of structurally conserved motifs. Both dictionaries are important because of the variety and significance of the problems that are impacted: automated functional annotation of genomic sequences, local homology detection, automated family definition, analysis of phylogenetic distribution, local 3D structure characterization, and others.

## Acknowledgments

The authors would like to thank Dan Platt for generously providing the code to align the 3D protein fragments in space, and Nikos Kyrpides for his help and invaluable insight on the problem of functional annotation.

## References

- Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; and Weng, 1987. *J. Protein Data Bank*. In Crystallographic Databases-Information Content, Software Systems, Scientific Applications, Eds. F. H. Allen, G. Bergerhoff, and R. Sievers, Data Commission of the International Union of Crystallography, Bonn-Cambridge-Chester, pp. 107-132.
- Altschul S.F.; Gish, W.; Miller, W.; Myers, E.W.; and Lipman, D.J. 1990. Basic local alignment search tool. *Journal Of Molecular Biology*; 5(3):403-410.
- Altschul, S.F.; Boguski, M.S.; Gish, W.; and Wootton, J.C. 1994. Issues in searching molecular sequence databases. *Nature Genetics*, 6, 119-129.
- Andersson, S.G.E.; Zomorodipour, A.; Andersson, J.O.; Sicheritz-Ponten, T.; Alsmark, U.C.M.; Podowski, R.M.; Naeslund, A.K.; Eriksson, A.S.; Winkler, H.H.; and Kurland, C.G. 1998. The Genome Sequence of *Rickettsia prowazekii* and the Origin of Mitochondria. *Nature* 396, 133-140.
- Attwood T.K.; Beck, M.E.; Flower, D.R.; Scordis, P.; and Selley, J.N. 1998. The PRINTS protein fingerprint database in its fifth year. *Nucleic Acids Research*; 26(1):304-308.
- Bairoch A. and Apweiler, R. 1998. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Research*; 26(1):38-42.
- Bairoch A.; Bucher, P.; and Hofmann, K. 1997. The PROSITE database, its status in 1997. *Nucleic Acids Research*; 25(1):217-221.
- Bernstein, F.C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, Jr.; E. F.; Brice, M.D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; and M. Tasumi. 1977. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *Journal Of Molecular Biology*; 112, 535-542.
- Blattner F.R.; Plunkett III, G.; Bloch, C.A.; Perna, N.T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J.D.; Rode, C.K.; Mayhew, G.F.; Gregor, J.; Davis, N.W.; Kirkpatrick, H.A.; Goeden, M.A.; Rose, D.J.; Mau, B.; and Shao, Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331):1453-1474.
- Bork P. and Gibson, T.J. 1996. Applying motif and profile searches. *Methods In Enzymology*; 266:162-184.
- Boulikas, T. Nuclear localization signals (NLS). 1993. *Crit. Rev. Eukaryot. Gene Expr.*; 3(3):193-227.
- Bult C.J.; White, O.; Olsen, G.J.; Zhou, L.; Fleischmann, R.D.; Sutton, G.; Blake, J. A.; FitzGerald, L.M.; Clayton, R.A.; Gocayne, J.D.; Kerlavage, A.R.; Dougherty, B.A.; Tomb, J.F.; Adams, M.D.; Reich, C.I.; Overbeek, R.; Kirkness, E.F.; Weinstock, K.G.; Merrick, J.M.; Glodek, A.; Scott, J.L.; Geoghagen N.S.M.; and J.C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, 273 (5278) :1058-1073.
- Cole, S.T.; Brosch, R.; Parkhill, J.; Garnier, T.; Churcher, C.; Harris, D.; Gordon, S.V.; Eiglmeier, K.; Gas, S.; Barry III, C.E.; Tekaaia, F.; Badcock, K.; Basham, D.; Brown, D.; Chillingworth, T.; Connor, R.; Davies, R.; Devlin, K.; Feltwell, T.; Gentles, S.; Hamlin, N.; Holroyd, S.; Hornsby, T.; Jagels, K.; Krogh, A.; McLean, J.; Moule, S.; Murphy, L.; Oliver, S.; Osborne, J.; Quail, M.A.; Rajandream, M.A.; Rogers, J.; Rutter, S.; Seeger, K.; Skelton, S.; Squares, S.; Sqaes, R.; Sulston, J.E.; Taylor, K.; Whitehead, S. and Barrell, B.G. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393 (6685), 537-544.
- Deckert, G.; Warren, P.V.; Gaasterland, T.; Young, W.G.; Lenox, A.L.; Graham, D.E.; Overbeek, R.; Snead, M.A.; Keller, M.; Aujay, M.; Huber, R.; Feldman, R.A.; Short, J.M.; Olson, G.J. and Swanson, R.V. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* 392, 353-358.
- Doolittle, R.F. 1994. Convergent evolution: the need to be explicit. *Trends Biochem Sci.*; (1):15-18.
- Doolittle, R.F. 1995. The multiplicity of domains in proteins. *Annual Review Of Biochemistry*; 64:287-314.
- Fleischmann, R.D.; Adams, M.D.; White, O.; Clayton, R.A.; Kirkness, E.F.; Kerlavage, A.R.; Bult, C.J.; Tomb, J.F.; Dougherty, B.A.; Merrick, J.M. et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496-512.
- Floratos, A.; Rigoutsos, I.; Parida, L.; Stolovitzky, G.; and Gao, Y. Sequence Homology Detection Through Large-Scale Pattern Discovery. 1999. *Proceedings 3rd Annual ACM International Conference on Computational Molecular Biology (RECOMB)*, France.
- Fraser C.M.; Gocayne, J.D.; White, O.; Adams, M.D.; Clayton, R.A.; Fleischmann, R.D.; Bult, C.J.; Kerlavage, A.R.; Sutton, G.; Kelley J.M. et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270 (5235) :397-403.
- Fraser, C.M.; Norris, S.J.; Weinstock, G.M.; White, O.; Sutton, G.G.; Dodson, R.; Gwinn, M.; Hickey, E.K.; Clayton, R.; Ketchum, K.A.; Sodergren, E.; Hardham, J.M.; McLeod, M.P.; Salzberg, S.; Peterson, J.; Khalak, H.; Richardson, D.; Howell, J.K.; Chidambaram, M.; Utterback, T.; McDonald, L.; Artiach, P.; Bowman, C.; Cotton, M.D.; Fujii, C.; Garland, S.; Hatch, B.; Horst, K.; Roberts, K.; Watthey, L.; Weidman, J.; Smith, H.O. and Venter, J.C. 1998. Complete Genome Sequence of *Treponema pallidum*, the Syphilis Spirochete. *Science* 281, 375-388.
- Fraser, C.M.; Casjens, S.; Huang, W.M.; Sutton, G.G.; Clayton, R.A.; Lathigra, R.; White, O.; Ketchum,

- K.A.; Dodson, R.; Hickey, E.K.; Gwinn, M.; Dougherty, B.; Tomb, J.-F.; Fleischmann, R.D.; Richardson, D.; Peterson, J.; Kerlavage, A.R.; Quackenbush, J.; Salzberg, S.; Hanson, M.; van-Vugt, R.; Palmer, N.; Adams, M.D.; Gocayne, J.D.; Weidman, J.; Utterback, T.; Watthey, L.; McDonald, L.; Artiach, P.; Bowman, C.; Garland, S.; Fujii, C.; Cotton, M.D.; Horst, K.; Roberts, K.; Hatch, B.; Smith, H.O. and Venter, J.C. 1997. Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*. *Nature* 390, 580-586.
- Gonnet G.H.; Cohen M.A.; and Benner, S.A. 1992. Exhaustive matching of the entire protein sequence database. *Science*, 256:1443-1445.
- Harris, N. L.; Hunter, L.; and States, D. J. 1992. *Proceedings 10th National Conference on Artificial Intelligence*, San Jose, CA.
- Henikoff J.G. and Henikoff, S. 1996. Blocks database and its applications. *Methods In Enzymology*; 266:88-105.
- Himmelreich R.; Hilbert, H.; Plagens, H.; Pirkl, E.; Li, B.C.; and Herrmann, R. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Research*; 24(22):4420-4449.
- Hodgman T.C. 1989. The elucidation of protein function by sequence motif analysis. *Comput Appl. Biosci.* 5(1):1-13.
- Holm, L. and Sander, C. 1998. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics.* 14(5):423-9.
- Horton P. and Nakai, K. 1997. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proceedings International AAAI Conference On Intelligent Systems For Molecular Biology*; 5:147-152.
- Jonassen, I.; Collins, J.F.; and Higgins, D.G. 1995. Finding flexible patterns in unaligned protein sequences. *Protein Science*; 4(8):1587-1595.
- Kaneko T. and Tabata, S. 1997. Complete genome structure of the unicellular cyanobacterium *Synechocystis* sp. PCC6803. *Plant Cell Physiol.*; 38(11):1171-1176.
- Kawarabayasi, Y.; Sawada, M.; Horikawa, H.; Haikawa, Y.; Hino, Y.; Yamamoto, S.; Sekine, M.; Baba, S.; Kosugi, H.; Hosoyama, A.; Nagai, Y.; Sakai, M.; Ogura, K.; Otsuka, R.; Nakazawa, H.; Takamiya, M.; Ohfuku, Y.; Funahashi, T.; Tanaka, T.; Kudoh, Y.; Yamazaki, J.; Kushida, N.; Oguchi, A.; Aoki, K.; Yoshizawa, T.; Nakamura, Y.; Robb, F.T.; Horikoshi, K.; Masuchi, Y.; Shizuya, H. and Kikuchi, H. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaeobacterium, *Pyrococcus horikoshii* OT3 (supplement). *DNA Res.* 5, 147-155.
- Klenk, H.P.; Clayton, R.A.; Tomb, J.; White, O.; Nelson, K.E.; Ketchum, K.A.; Dodson, R.J.; Gwinn, M.; Hickey, E.K.; Peterson, J.D.; Richardson, D.L.; Kerlavage, A.R.; Graham, D.E.; Kyripides, N.C.; Fleischmann, R.D.; Quackenbush, J.; Lee, N.H.; Sutton, G.G.; Gill, S.; Kirkness, E.F.; Dougherty, B.A.; McKenney, K.; Adams, M.D.; Loftus, B.; Peterson, S.; Reich, C.I.; McNeil, L.K.; Badger, J.H.; Glodek, A.; Zhou, L.; Overbeek, R.; Gocayne, J.D.; Weidman, J.F.; McDonald, L.; Utterback, T.; Cotton, M.D.; Spriggs, T.; Artiach, P.; Kaine, B.P.; Sykes, S.M.; Sadow, P.W.; D'Andrea, K.P.; Bowman, C.; Fujii, C.; Garland, S.A.; Mason, T.M.; Olsen, G.J.; Fraser, C.M.; Smith, H.O.; Woese, C.R. and Venter, J.C. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390 (6658), 364-370.
- Krause, A. and Vingron, M. 1998. A set-theoretic approach to database searching and clustering. *Bioinformatics*, 14(5).
- Kunst, F.; Ogasawara, N.; Moszer, I.; Albertini, A.M.; Alloni, G.; Azevedo, V.; Bertero, M.G.; Bessieres, P.; Bolotin, A.; Borchert, S.; Borriss, R.; Boursier, L.; Brans, A.; Braun, M.; Brignell, S.C.; Bron, S.; Brouillet, S.; Bruschi, C.V.; Caldwell, B.; Capuano, V.; Carter, N.M.; Choi, S.K.; Codani, J.J.; Connerton, I.F.; Cummings, N.J.; Daniel, R.A.; Denizot, F.; Devine, K.M.; Dusterhoft, A.; Ehrlich, S.D.; Emmerson, P.T.; Entian, K.D.; Errington, J.; Fabret, C.; Ferrari, E.; Foulger, D.; Fritz, C.; Fujita, M.; Fujita, Y.; Fuma, S.; Galizzi, A.; Galleron, N.; Ghim, S.Y.; Glaser, P.; Goffeau, A.; Golightly, E.J.; Grandi, G.; Guiseppi, G.; Guy, B.J.; Haga, K.; Haiech, J.; Harwood, C.R.; Henaut, A.; Hilbert, H.; Holsappel, S.; Hosono, S.; Hullo, M.F.; Itaya, M.; Jones, L.; Joris, B.; Karamata, D.; Kasahara, Y.; Klaerr-Blanchard, M.; Klein, C.; Kobayashi, Y.; Koetter, P.; Koningstein, G.; Krogh, S.; Kumano, M.; Kurita, K.; Lapidus, A.; Lardinois, S.; Lauber, J.; Lazarevic, V.; Lee, S.M.; Levine, A.; Liu, H.; Masuda, S.; Mauel, C.; Medigue, C.; Medina, N.; Mellado, R.P.; Mizuno, M.; Moestl, D.; Nakai, S.; Noback, M.; Noone, D.; O'Reilly, M.; Ogawa, K.; Ogiwara, A.; Oudega, B.; Park, S.H.; Parro, V.; Pohl, T.M.; Portetelle, D.; Porwollik, S.; Prescott, A.M.; Presecan, E.; Pujic, P.; Purnelle, B.; Rapoport, G.; Rey, M.; Reynolds, S.; Rieger, M.; Rivolta, C.; Rocha, E.; Roche, B.; Rose, M.; Sadaie, Y.; Sato, T.; Scanlan, E.; Schleich, S.; Schroeter, R.; Scoffone, F.; Sekiguchi, J.; Sekowska, A.; Seror, S.J.; Serror, P.; Shin, B.S.; Soldo, B.; Sorokin, A.; Tacconi, E.; Takagi, T.; Takahashi, H.; Takemaru, K.; Takeuchi, M.; Tamakoshi, A.; Tanaka, T.; Terpstra, P.; Tognoni, A.; Tosato, V.; Uchiyama, S.; Vandenbol, M.; Vannier, F.; Vassarotti, A.; Viari, A.; Wambutt, R.; Wedler, E.; Wedler, H.; Weitzenegger, T.; Winters, P.; Wipat, A.; Yamamoto, H.; Yamane, K.; Yasumoto, K.; Yata, K.; Yoshida, K.; Yoshikawa, H.F.; Zumstein, E.; Yoshikawa, H. and Danchin, A. 1997. The complete genome

- sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249-256.
- Linial, M.; Linial, N.; Tishby, N.; and Yona, G. 1997. Global self-organization of all known protein sequences reveals inherent biological signatures. *Journal Of Molecular Biology*; 268 (2) :539-556.
- Neuwald, A.F. and Green, P. 1994. Detecting Patterns In Protein Sequences. *Journal Of Molecular Biology*, pp. 698-712.
- Nevill-Manning, C.G.; Wu, T.D.; and Brutlag, D.L. 1998. Highly specific protein sequence motifs for genome analysis. *Proceedings National Academy Of Sciences, USA*, 95(11):5865-5871.
- Ogiwara A.; Uchiyama, I.; Seto Y.; and Kanehisa, M. 1992. Construction of a Dictionary of sequence motifs that characterize groups of related proteins. *Protein Engineering*; (6):479-488.
- Parida, L.; Floratos, A.; and Rigoutsos, I. 1998. MUSCA: An Algorithm for Constrained Alignment of Multiple Data Sequences. *Proceedings 9th Workshop on Genome Informatics*, Tokyo, Japan.
- Rigoutsos, I. and Floratos, A. 1998a. Combinatorial pattern discovery in biological sequences: the *Teiresias* algorithm. *Bioinformatics*, 14(1):55-67.
- Rigoutsos, I. and Floratos, A. 1998b. Motif Discovery Without Alignment Or Enumeration. *Proceedings 2nd Annual ACM International Conference on Computational Molecular Biology (RECOMB)*, New York, NY.
- Rigoutsos, I.; Floratos, A.; Ouzounis, C.; Parida, L.; Stolovitzky, G.; and Gao, Y. 1998. Unsupervised Hierarchical Motif Discovery In the Sequence Space Of Natural Proteins. Technical Report RC 21218. IBM TJ Watson Research Center.
- Rigoutsos, I.; Floratos, A.; Ouzounis, C.; Gao, Y; and Parida, L. 1999. Dictionary Building Via Unsupervised Hierarchical Motif Discovery In the Sequence Space Of Natural Proteins. Forthcoming..
- Sander, C. and Schneider, R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56-68.
- Saraste M.; Sibbald, P.R.; and Wittinghofer, A. 1990. The P-loop - a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.*; 15(11):430-434.
- Smith R.F. and Smith, T.F. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proceedings National Academy Of Sciences, USA*, 87:118-122.
- Smith, H.; Annau, T.; and Chandrasegaran, S. 1990. Finding Sequence Motifs In Groups Of Functionally Related Proteins. *Proceedings National Academy Of Sciences, USA*, vol. 87, pp. 826-830.
- Smith, D.R.; Doucette-Stamm, L.A.; Deloughery, C.; Lee, H.-M.; Dubois, J.; Aldredge, T.; Bashirzadeh, R.; Blakely, D.; Cook, R.; Gilbert, K.; Harrison, D.; Hoang, L.; Keagle, P.; Lumm, W.; Pothier, B.; Qiu, D.; Spadafora, R.; Vicare, R.; Wang, Y.; Wierzbowski, J.; Gibson, R.; Jiwani, N.; Caruso, A.; Bush, D.; Safer, H.; Patwell, D.; Prabhakar, S.; McDougall, S.; Shimer, G.; Goyal, A.; Pietrovski, S.; Church, G.M.; Daniels, C.J.; Mao, J.-i.; Rice, P.; Nolling, J. and Reeve, J.N. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H: functional analysis and comparative genomics. *Journal Of Bacteriology* 179, 7135-7155.
- Sonnhammer, E.L.; Eddy, S.R.; and R. Durbin. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405-420.
- Stephens, R.S.; Kalman, S.; Lammel, C.J.; Fan, J.; Marathe, R.; Aravind, L.; Mitchell, W.P.; Olinger, L.; Tatusov, R.L.; Zhao, Q.; Koonin, E.V. and Davis, R.W. 1998. Genome Sequence of an Obligate Intracellular Pathogen of Humans: *Chlamydia trachomatis*. *Science*. 282(5389):754-9.
- Tatusov, R.L.; Koonin, E.V.; and Lipman, D.J. 1997. A genomic perspective on protein families. *Science*, 278(5338):631-637.
- Tomb, J.-F.; White, O.; Kerlavage, A.R.; Clayton, R.A.; Sutton, G.G.; Fleischmann, R.D.; Ketchum, K.A.; Klenk, H.P.; Gill, S.; Dougherty, B.A.; Nelson, K.; Quackenbush, J.; Zhou, L.; Kirkness, E.F.; Peterson, S.; Loftus, B.; Richardson, D.; Dodson, R.; Khalak, H.G.; Glodek, A.; McKenney, K.; Fitzgerald, L.M.; Lee, N.; Adams, M.D.; Hickey, E.K.; Berg, D.E.; Gocayne, J.D.; Utterback, T.R.; Peterson, J.D.; Kelley, J.M.; Cotton, M.D.; Weidman, J.M.; Fujii, C.; Bowman, C.; Watthey, L.; Wallin, E.; Hayes, W.S.; Borodovsky, M.; Karp, P.D.; Smith, H.O.; Fraser, C.M. and Venter, J.C. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388 (6642), 539-547.
- Wooton, J. and Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology*, 266, 554-571.
- Yona, G.; Linial, N.; Tishby, N.; and Linial, M. 1998. A map of the protein space - an automatic hierarchical classification of all protein sequences. *Proceedings International AAAI Conference On Intelligent Systems For Molecular Biology*. Montreal, Canada.

