2013 IEEE 13th International Conference on Data Mining

Statistical Inference of Protein "LEGO Bricks"

Arun S. Konagurthu
Lloyd Allison
David Abramson
Clayton School of Information Technology, Monash University
Clayton, Victoria 3800 Australia
Email: arun.konagurthu@monash.edu

Peter J. Stuckey
Department of Computing and Information Systems
University of Melbourne,
Parkville 3010 Australia

Arthur M. Lesk
Department of Biochemistry and Molecular Biology
Pennsylvania State University
University Park, Pennsylvania, 16802 USA

***Abstract*—Proteins are biomolecules of life. They fold into a great variety of three-dimensional (3D) shapes. Underlying these folding patterns are many recurrent structural fragments or building blocks (analogous to 'LEGO® bricks'). This paper reports an innovative statistical inference approach to discover a comprehensive dictionary of protein structural building blocks from a large corpus of experimentally determined protein structures. Our approach is built on the Bayesian and information-theoretic criterion of minimum message length. To the best of our knowledge, this work is the first systematic and rigorous treatment of a very important data mining problem that arises in the cross-disciplinary area of structural bioinformatics. The quality of the dictionary we find is demonstrated by its explanatory power – any protein within the corpus of known 3D structures can be dissected into successive regions assigned to fragments from this dictionary. This induces a novel one-dimensional representation of three-dimensional protein folding patterns, suitable for application of the rich repertoire of character-string processing algorithms, for rapid identification of folding patterns of newly-determined structures. This paper presents the details of the methodology used to infer the dictionary of building blocks, and is supported by illustrative examples to demonstrate its effectiveness and utility.**

I. Introduction

Proteins are molecules central to life. They are responsible for biological and cellular functions in all organisms. Each protein folds into a three-dimensional (3D) shape, determined by the intrinsic properties of its sequence or chain of amino acid residues. Among the major triumphs of modern science are the techniques to experimentally determine the 3D structures of proteins at atomic resolution. Worldwide structure determination efforts have resulted in a fast growing public database, the protein data bank (wwPDB) [1]. Currently wwPDB contains atomic coordinates of $91,550$ experimentally solved protein structures, whose size is doubling every five years. This database provides a rich source of structural and architechtural information for knowledge discovery and data mining applications that contribute to the advances made in life sciences in medicine.

Understanding the architectural principles of protein 3D structure is fundamental to biological research. It is well known that protein folding patterns contain recurrent structural themes, commonly helices and pleated sheets [2], [3]. However the identification of a canonical set of building blocks (analogous to LEGO® bricks) of protein structures still remains an important open questions in biology.

Previous investigations have sought to identify a dictionary of fragments as building blocks of protein structures [4]–[10]. (By fragment we mean a contiguous region within the folding pattern of a protein – this can be viewed as a 3D analogue of a 1D substring.) However, these approaches largely rely on *ad hoc* clustering of short fragments of some fixed-length, usually 4 to 10 amino acid residues long. The restriction of generating fixed-length fragment libraries is an artificial constraint, mainly employed to work around the difficulty of the search problem that manifests when trying to identify recurring fragments of arbitrary length. Thus, the question, *what is the canonical dictionary of fragments (of arbitrary lengths) of which all proteins are made*, remains fundamentally unsolved.

This paper addresses the above question by framing it as as a statistical inference problem. Our approach relies on the Bayesian method of minimum message length inference [11], [12], where the optimal fragment dictionary is defined objectively as the one which permits the most concise explanation (or technically, shortest lossless encoding) of the coordinates of a collection of source protein structures. To the best of our knowledge, our work is the first objective and systematic treatment of this important question, addressed using a statistically rigorous approach which investigates the compressibility of protein coordinate data.

We mine these building blocks from a collection of 8992 experimentally determined protein structures, whose coordinates were solved at atomic resolution, and available from the wwPDB [1]. These source structures are dissimilar in amino acid sequence to avoid experimental and selection bias observed within the wwPDB. In other words, the collection we use is comprehensive and unbiased, representative of all known protein folding patterns in the wwPDB.

Our approach discovered 1711 fragments or building blocks, ranging in length from 4 to 31 amino acid residues. This dictionary allows the efficient, lossless representation (or encoding) of any given protein structure. For a particular protein structure, the optimal lossless encoding contains a dissection (or segmentation) – that is, a designation of successive non-overlapping regions in the protein structures that match the assigned dictionary fragments – and a statement of spatial deviations (or corrections) that should be applied to the coordinates of each assigned dictionary fragment so that the coordinates corresponding to each region in the actual structure can be recovered losslessly. We note that the regions that do not efficiently match any dictionary fragment are assigned to

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.73

a 'null model'; in these cases, the spatial deviations bear the entire weight of the description – this is tantamount to stating the coordinates of the region in the source structure raw (or *as is*).

The organization of the paper is as follows. Section II gives a brief introduction to the minimum message length criterion. Section III provides the foundations of the dictionary inference problem using the MML framework. This involves designing transparent communication processes, developing lossless encoding strategies, and evolving efficient search algorithms. The details of these encoding schemes are available in the longer version of the paper available from the arχiv preprint server: refer http://arxiv.org/abs/1310.1462. The search strategy to find the optimal dictionary is described in Section IV. We conclude this paper with Section V providing various results including illustrative examples of the effectiveness and the utility of the dictionary we discover.

II. Minimum Message Length Framework

Minimum Message Length (MML) [11], [12] is a hypothesis (or model) selection paradigm which links statistical inference with information theory and data compression.

MML is a Bayesian method of inference. Formally, let E denote a mass of observed data (or evidence) and H a hypothesis on the data. From Bayes's theorem [13] we have: $P(H\ \&\ E) = P(H) \times P(E|H) = P(E) \times P(H|E)$, where $P(H)$ is the *prior* probability of hypothesis H, $P(E)$ is the prior probability of data E, $P(H|E)$ is the *posterior* probability of H given E, and $P(E|H)$ is the *likelihood*.

In the Bayesian framework, two competing hypotheses can be compared using the ratios of their posterior probabilities:

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{P(H_1)P(E|H_1)}{P(H_2)P(E|H_2)}$$

Usually, the goal of inference is to select the hypothesis with the highest posterior probability.

MML offers a complementary view of Bayesian inference by linking the probability of an event with the message length required to transmit (or communicate, explain, describe) it *losslessly*. The mathematical theory of communication [14] gives the relationship between the message length $I(e)$ to communicate an event e losslessly, and its probability $P(e)$: $I(e) = -\log P(e)$.[1]

Therefore, by applying Shannon's insight to Bayese theorem above, we get:

$$I(H\ \&\ E) = I(H) + I(E|H) = I(E) + I(H|E)$$

Similarly, two competing hypotheses can be compared as:

$$I(H_1|E) - I(H_2|E) = I(H_1) + I(E|H_1) - I(H_2) - I(E|H_2)$$

It directly follows that the best hypothesis H^* is the one for which the expression $I(H^*) + I(E|H^*)$ takes the *minimum* value. (Notice, this is equivalent to maximizing the posterior probability of the hypothesis given the data, $P(H^*|E)$.)

[1]The unit of measurement of information depends on the base of the logarithm; $\log_2$ gives message lengths measured in bits, while $\ln$ gives the same measured in nits.

MML is best understood as a communication process between an imaginary pair of transmitter (Alice) and receiver (Bob) connected by a Shannon channel. Alice's objective is to send the data E using an explanation message in a form that Bob can receive it losslessly. Alice and Bob agree on a *codebook* containing general rules of communication composed solely of common knowledge about typical, hypothetical data. Anything that is not a part of the codebook must be strictly transmitted as a part of the explanation message. If Alice can find the best hypothesis, H^* on the data, Bob will receive a decodable explanation message most economically.

Alice sends the explanation message of E in two parts. In the first part she transmits the best hypothesis, H^*, she could find on the data E taking $I(H^*)$ bits to communicate. In the second, Alice transmits the details of the observed data E not explained by the hypothesis H^*, taking $I(E|H^*)$ bits to communicate. (That is, this part correspond to the deviations of the observed data E with respect to H^*). Notice that MML inference gives a natural trade-off between hypothesis complexity ($I(H^*)$) and quality of its fit to the data ($I(E|H^*)$).

III. Inferring the dictionary using the MML criterion

Preliminaries: Let $\mathcal{C}$ denote a *collection* of source protein structures $\{\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_{|\mathcal{C}|}\}$. In this work we use a subset of 8992 structures from the protein data bank after removing amino acid sequence bias. That is, no two structures in the collection $\mathcal{C}$ have a sequence similarity greater than 40%.

Any protein structure $\mathcal{P}$ is represented as an ordered list of (x, y, z) coordinates of its alpha Carbon (C_α) atoms along the protein backbone, denoted here as $\mathcal{P} = \{p_1, \cdots, p_{|\mathcal{P}|}\}$. All protein coordinates are defined in Angstrom units (1Å$= 10^{-10}$ meters.)

Let $\mathcal{D} = \{\mathcal{Q}_1, \mathcal{Q}_2, \cdots, \mathcal{Q}_{|\mathcal{D}|}\}$ denote a *dictionary* of fragments. Each dictionary *element* $\mathcal{Q} = \{q_1, \cdots, q_{|\mathcal{Q}|}\} \in \mathcal{D}$ is a substructural fragment (*i.e.,* a list of coordinates corresponding to a consecutive region) derived from some $\mathcal{P} \in \mathcal{C}$, of arbitrary length ($|\mathcal{Q}| < |\mathcal{P}|$).

We note that each (x, y, z) comes specified (in the protein data bank) to 3 positions after the decimal place. Since we are dealing with inference based on lossless compression, we denote $\epsilon = 0.001$ as a parameter that specifies the accuracy to which coordinate data should be stated.

Rationalizing this problem in the MML framework: In this work, any dictionary $\mathcal{D}$ of fragments is a hypothesis of building blocks on a collection of structures $\mathcal{C}$, with its observed coordinates acting as evidence for inference. Therefore, using the information-theoretic restatement of Bayes's theorem describe in Section II, we get:

$$I(\mathcal{D}\ \&\ \mathcal{C}) = I(\mathcal{D}) + I(\mathcal{C}|\mathcal{D}). \tag{1}$$

Rationalizing Equation 1 as a communication process between Alice and Bob, the measure of quality of any proposed dictionary of substructures is the total length of the explanation message that Alice transmits to Bob so that all the coordinates in the collection of source structures are received losslessly. Given that even unrelated proteins contain common, recurrent

fragments (or building blocks), Alice and Bob could reasonably hypothesize that they could apply this observation to transmit the coordinates of various structures more concisely, by using the dictionary of building blocks as the basis of communication. It is intuitive to see that the better a dictionary of fragments in terms of how well they describe (i.e., *fit*) the observed coordinates, the more economical is the description of the source structures in the collection.

To be useful for illuminating common building blocks over all proteins, the dictionary must be the same for all structures; that is, the dictionary does not change regardless of an individual structure that is being transmitted. Before transmitting the coordinates of the collection, Alice first encodes and sends Bob the canonical dictionary of fragments (taking $I(\mathcal{D})$ bits). With this information, Bob has a dictionary of substructures but not the coordinates of the source structures in the collection. Note again that Alice needs to send the dictionary only once (as a header (or first part) of the total explanation message): she need not restate it as part of the subsequent encodings of coordinates of particular protein structures being transmitted. Each subsequent message consists of the segmentation, that is, the optimal assignment of successive regions in a protein structure to dictionary fragments, plus the corrections (or vector deviations) required because each region in the source protein deviates from its assigned dictionary element. This takes $I(\mathcal{C}|\mathcal{D})$ bits.

In proposing a dictionary Alice and Bob face a tradeoff. They could use a large, all-encompassing fragment dictionary whose elements fit regions of proteins very well, leaving only small deviations in the assigned regions of the proteins to be described. In this case the explanation message length for each protein would be dominated by the explanation cost of the dictionary and the assignment of dictionary fragments to regions, because a large dictionary requires a larger message to explain itself and to nominate a dictionary element. Alternatively, they could choose a small dictionary, in which case the message length would be dominated by the transmission of the corrections (vector deviations). As described in Section II, the MML criterion provides an objective tradeoff between the dictionary complexity and its fit with the coordinate data observed in the collection.

Optimality criterion: The optimal dictionary involves finding a dictionary of fragments that minimizes the total message length equation shown in Equation 1. To achieve this involves the following criteria:

1) Assume some dictionary $\mathcal{D}$ is given (containing arbitrary number of fragments, each of arbitrary length). According to the MML framework, the *optimal encoding of a particular protein structure* $\mathcal{P}$ using the specified dictionary $\mathcal{D}$ is defined as the combination, of minimal encoding length, of assignments of successive non-overlapping regions in $\mathcal{P}$ to fragments in $\mathcal{D}$, plus statements of spatial deviations relative to each assigned fragment per region to recover the observed coordinates in $\mathcal{P}$ losslessly.
2) Next, given the above method to optimally encode a particular protein using a specified dictionary, the *optimal encoding of a collection of protein structures* $\mathcal{C}$, all using the same fixed dictionary $\mathcal{D}$, requires the one-off statement of the dictionary $\mathcal{D}$ (as a header to the subsequent explanation message), plus the optimal encodings of each individual protein $\mathcal{P}_i \in \mathcal{C}$ using the method in Step 1.
3) Finally, given the method to optimal encode a collection of proteins in Step 2, an *optimal dictionary for the collection of protein structures* can be objectively defined as the one for which the one-off specification cost of the dictionary $\mathcal{D}$, plus the sum of the optimal encodings of all the proteins $\mathcal{P}_i \in \mathcal{C}$, yields the *shortest* explanation message.

(Refer http://arxiv.org/abs/1310.1462 for details of these steps.)

IV. The Search for the Optimal Dictionary

Equation 1 provides a rigorous objective to search for the dictionary of fragments building blocks of protein three-dimensional structures. It follows that an optimal dictionary for a collection of protein structures is one for which the statement of the dictionary, plus the sum of the optimal encodings of all the proteins in the set, is the *shortest*. That is, the objective of this work is to find a $\mathcal{D}^*$ such that:

$$I(\mathcal{D}^* \,\&\, \mathcal{C}) = \min_{\forall \mathcal{D}} I(\mathcal{D} \,\&\, \mathcal{C}) \quad \text{bits.} \tag{2}$$

Clearly, any fragment (of arbitrary size) from within any protein in the collection is a potential candidate for the dictionary. Therefore searching for the best dictionary leads to a very large optimization problem. Since the problem is computational intractable to find a provably optimal dictionary, we designed a simulated annealing algorithm in order to evolve a dictionary that iteratively converges to the best dictionary defined by the Equation 2.

Simulated Annealing is an heuristic approach which has an analogy with cooling of solids. Here, we consider each possible dictionary (of arbitrary number of fragments) as being analogous to some state of a physical system. The message length of transmitting a collection of structures using any dictionary given by Equation 1 is analogous to the internal energy of the physical system in that state.

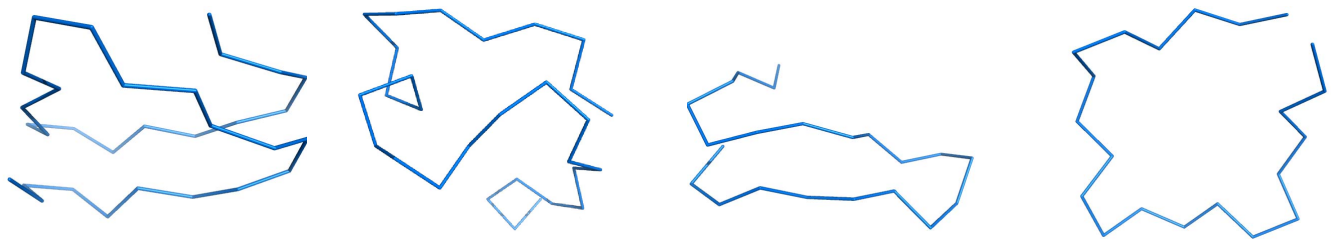
The method starts with an empty dictionary. In this state, each protein in the collection is transmitted raw, as a random coil using the null model. The strategy involves iteratively perturbing the dictionary from this initial empty state to a state where the total message length objective is minimized.

A. Perturbations

At each step the current dictionary is perturbed randomly (which is akin to sampling some new nearby dictionary state).

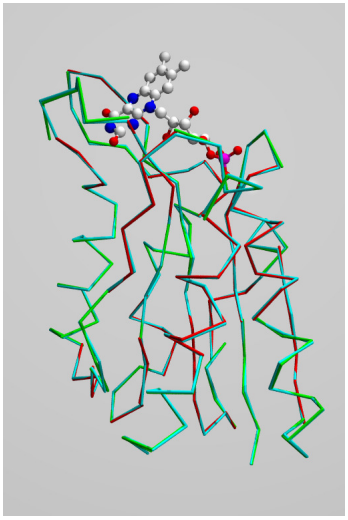
The choice of moving to the new state or remaining in the current one is decided probabilistically. Specifically, at each iteration, our method employs one of the following randomly chosen perturbations:

Add: Append to the current dictionary a new fragment. This fragment is from a randomly chosen protein from the collection, of random length.



(a) (b) (c) (d)

Fig. 1. Four fragments chosen from our dictionary our approach discovered. (a) $1\frac{2}{3}$ turns of three-fold β-helix of length 31 residues. It occurs 26 times in our collection. See wwPDB 2IC7 for one such instance. (b) An exotic β-hairpin of length 29 residues. This occurs 20 times in the collection. See wwPDB 1UJU for one such instance. (c) A long β-hairpin of length 22 residues with 47 occurrences. See wwPDB 1JO8 for one such instance. (d) 1 turn of four-fold β-helix of length 21, which occurs 19 times in the collection.



Resi	Model	RMSD (Å)	Resi	Model	RMSD (Å)
2 - 6	m1096	0.12	78 - 82	m1415	0.14
6 - 10	m1195	0.21	82 - 85	m1623	0.07
10 - 13	m1595	0.24	85 - 89	m1083	0.20
13 - 22	m0231	0.16	89 - 92	m1706	0.03
22 - 32	m0159	0.18	92 - 95	m1611	0.06
32 - 37	m0874	0.27	95 - 100	m0967	0.53
37 - 41	m1202	0.24	100 - 103	m1571	0.17
41 - 45	m1246	0.18	103 - 115	m0054	0.12
45 - 49	m1323	0.17	115 - 119	m1499	0.19
49 - 55	m0502	0.19	119 - 123	m1480	0.21
55 - 60	m0930	0.28	123 - 128	m0769	0.29
60 - 63	m1685	0.07	128 - 133	m0750	0.32
63 - 67	m1306	0.26	133 - 141	m0281	0.12
67 - 71	m1194	0.16	141 - 148	m0426	0.19
71 - 78	m0423	0.12			

Fig. 2. Dissection of the structure of flavodoxin from *Desulfovibrio vulgaris* (wwPDB entry 1J8Q). Parent structure shown in cyan; successive dictionary fragments in alternating red and green. Cofactor flavin mononucleotide (FMN) in ball-and-stick representation. Dissected regions are listed as a table below the picture.

Remove: Remove a randomly chosen fragment from the existing dictionary state.

Swap: Replace a randomly chosen fragment in the current dictionary with another randomly chosen fragment from the collection. This is equivalent to the sequence of perturbations: '`Remove`' followed by an '`Add`'.

Perturb length: Expand or shrink a randomly chosen fragment from the current dictionary state by one residue, at the randomly chosen end. (Expanding a previously chosen fragment is achieved by remembering its locus – i.e., the fragment's source protein structure and its offset in the source. This allows any existing fragment to be elongated by an additional coordinate, based on the information available in its source structure.)

B. Probability of acceptance or rejection of any perturbation

. Equation 1 gives the estimate of the negative logarithm of the joint probability of a dictionary and a collection. In Section II we have seen that the difference between the message lengths using any two different hypotheses (here, two dictionary states) gives the log-odds posterior ratio. This implies, if the total message length using a perturbed dictionary is k bits shorter (conversely, longer) than the current state, then the perturbed dictionary is 2^k times more likely (conversely, unlikely) than the perturbed state.

The simulated annealing heuristic starts with a high (time-varying) parameter t (after temperature). Let I_{current} and $I_{\text{perturbed}}$ be the total message lengths using the current and perturbed states of a dictionary. During any iteration, if $\Delta I \equiv (I_{\text{current}} - I_{\text{perturbed}}) < 0$, the perturbed state is immediately accepted as the new current state, and the procedure continued Otherwise, the perturbed state is accepted with a probability of $1/2^{\frac{\Delta I}{t}}$.

C. Cooling schedule

For simulated annealing algorithms, the variation of the temperature parameter T controls the evolution of the dictionary states. We set T to a high $10,000$ at the start. The parameter T decays at a constant rate of 0.88. For each value of

t between $10,000$ and 10, we carry out $10,000$ perturbations, while for value of T less than 10, we carry out $100,000$ perturbations to the dictionary. The stopping criterion is when the number of iterations reaches 2 million iterations (which occurs at $t = 0.246$).

D. Implementation:

A program to optimally encode a given collection of structures and then search for the best dictionary using simulated annealing was implemented in the `C++` programming language. Message Passing Interface (`MPI`) was used to parallelize this program to run on a large computational cluster.

A MapReduce model was used to distribute the encoding tasks on the large cluster. In this model, a master node accepts a collection of protein structures as input and evenly divides them into smaller subsets of structures, where each subset is distributed to worker nodes. For every perturbation of a current dictionary state, each worker node computes the sum of message lengths to optimally encode each of the structures in its allocated subset. The worker node then returns the message length back to the master node, which collects all the message lengths and combines them to compute the total message length to encode all the structures in the given collection using the current state of the dictionary.

V. Results

The work resulted in a dictionary of 1711 fragments ranging in length from 4 to 31 amino acids. This dictionary was used to dissect the entire collection of 8992 source structures. The average root-mean-square (r.m.s.) deviation of orthogonal superposition of the dictionary fragments to the assigned regions is 0.29 Å. (Note, this does not include the separate statement and application of deviations which is part of the transmitted message, which would reduce 0.29 Å to 0.) The average, over all proteins in the set, of the maximum r.m.s. deviation of any model from all regions it encodes, is 1.23 Å. In Fig. 1 shows the visualization of four fragments chosen from the dictionary we discovered, of lengths 31, 29, 22 and 21 respectively. Previous methods, due to the length constraint (see Section I), are unable to detect recurrent fragments this long.

Figure 2 shows the optimal dissection into dictionary fragments of the structure of flavodoxin from *Desulfovibrio vulgaris* (wwPDB entry 1J8Q, solved at 1.35 Å resolution [15].) To encode losslessly the flavodoxin structure, the dissection would be accompanied by the vector deviations of the C_α atoms of flavodoxin from the assigned canonical dictionary fragments. Noteworthy properties of the dissection in this example include:

1) The fits of the individual fragments of the dictionary to the structure are quite precise in almost all cases. The deviations are of the order of only tenths of an Å. (The maximum r.m.s.d. between the coordinates of a region in the protein and the assigned model is 0.53 Å; this occurs only once. All other r.m.s.d. values are $\leq$ 0.32 Å.)
2) Here the dictionary fragments account for the entire structure. No regions of the structure need be encoded as a random coil; that is, using the null model.
3) The range of lengths of the dictionary fragments appearing in the dissection of flavodoxin is from 4 to 13. Some of the segments correspond to individual fragments of secondary structure – helices and strands of sheets. Others correspond to N- or C-terminal parts of secondary structures, plus parts of the loops either preceding or succeeding them.
4) The sequence of dictionary fragments provides a one-dimensional representation of the protein folding pattern.

TABLE I. Clustering of models in the identified dictionary

Class	Size	Code	Description
1	616	e	short extended regions
2	301	t	short non-hairpin turns (some $\beta-$bulges)
3	4	t	short non-hairpin turns
4	325	h	short helices
5	164	h	short helices
6	167	E	extended regions (some with hooks at end)
7	6	E	extended regions (some curved)
8	13	T	non-hairpin turns (some $\beta-$bulges)
9	50	b	shorter $\beta-$hairpin
10	30	B	$\beta-$hairpins
11	3	B	$\beta-$hairpins, unequal arms ('shepherd's crook')
12	1	B	$\beta-$hairpins, unequal arms ('shepherd's crook')
13	3	H	irregular alpha helix (plus C_α-only)
14	17	H	long alpha helices
15	2	Ω	long wide loops (Ω loop)
16	1	Ω	long wide loops (Ω loop)
17	2	Ω	long wide loops (Ω loop)
18	1	C	double $\beta-$hairpin ('paper clip')
19	1	B	long twisted $\beta-$hairpins
20	2	Ω	helix-strand-helix-strand / wide (Ω) loops
21	1	3	$1\frac{2}{3}$ turns of three-fold $\beta-$helix
22	1	4	1 turn of four-fold ($\beta-$helix)

Clustering of the dictionary fragments: To further rationalize the 1711 dictionary fragments, we clustered the dictionary fragments into coarse structural classes with the UPGMA method [16] using the Mahalanobis distance [17] computed from the following characterising properties:

1) the number of backbone hydrogen bonds between residues separated by 4 in the sequence (to group helices which demonstrate this periodicity),
2) the distance between the C_α atoms of N- and C-terminal residues,
3) the cosine of the angle between the C_α atom of the N-terminal residue, the C_α atom of the middle residue (or, for fragments containing even numbers of residues, the average position of the C_α atoms of the two middle residues),
4) the r.m.s. deviation of a fit of the C_α atoms to a straight line, and
5) the average value of the cosine of the dot products of $C\rightarrow O$ vectors of successive residues. Table I shows the clusters, and suggested class codes.

One-dimensional representation of protein folds. We consider as a case study the assignment of fragments to clusters which reveals structural patterns in Drosophila lebanonensis *alcohol dehydrogenase* (wwPDB code 1SBY).

The details of the dissection of amino acid residues 1-183 of 1SBY into dictionary fragments are available in the longer version of our paper available from http://arxiv.org/abs/1310.1462. The sequence of class symbols derived from the dissection, converted to upper case to suppress the distinction between short and long versions of the same substructure, affords a more perspicuous representation of this dissection:

EETTHTEEHEETHHEEETHHTTEEHTHHTEE

(E = strand, H = helix, T = non-hairpin turn (Table 1).)

This is an instance of the regular expression:

$$(\mathtt{E+T+H+T+E+HE*TH+T*E+})\{2\}$$

This sequence from the dissection provides a concise one-dimensional representation of the folding pattern. It captures the duplication of the two $\beta-\alpha-\beta-\alpha-\beta$ substructures (and the points of insertion of non-pattern elements) but not their symmetrical spatial disposition. Nevertheless, the dissection provides a faithful signature of the NAD-binding domain folding pattern [18].

The rich repertoire of algorithms on character strings is applicable. For example, the string could be used to design regular expressions for probing collections for similar structures.

More generally, standard regular-expression-matching algorithms permit application of the linear representations to identify common folding patterns in a set of structures, or, specifically, to classify a newly-determined structure, as for example in SCOP [19] and CATH [20]. The representation can also identify variations and deviations from standard folding patterns in known families. For instance, some NAD-binding domains (including *Drosophila lebanonensis* alcohol dehydrogenase) contain extra helices and/or hairpins [21] and this would be revealed by a 'sequence' alignment of the dissections of members of this family.

VI. Conclusion

This work introduces a novel method to infer the dictionary of building blocks of protein structures. This work fall squarely into one of the most important cross-disciplinary areas of modern science, where biology and computing meet. The approach described in this paper is a successful demonstration of rigorous statistical inference applied to an important data mining problem in structural Bioinformatics. The knowledge of this dictionary direct us to a number of avenues for further research. These include, for instance, straightforward generalisations, such as inclusion of backbone and even sidechain atoms. The exploration of the linear representation of folding patterns, and its potential correlation with sequence signals is a particularly attractive challenge. Working out the grammar associated with sequences of classes of dictionary fragments can illuminate different levels of folding architectures. The dictionary fragments themselves might well be applicable in approaches to predictions of protein structure, the holy grail of bioinformatics.

References

[1] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide protein data bank," *Nature Structural & Molecular Biology*, vol. 10, no. 12, pp. 980–980, 2003.

[2] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain," *Proceedings of the National Academy of Sciences*, vol. 37, no. 4, pp. 205–211, 1951.

[3] L. Pauling and R. B. Corey, "The pleated sheet, a new layer configuration of polypeptide chains," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 37, no. 5, p. 251, 1951.

[4] R. Unger, D. Harel, S. Wherland, and J. Sussman, "A 3D building blocks approach to analyzing and predicting structure of proteins," *Proteins*, vol. 5, pp. 355–373, 1989.

[5] R. J. Rooman, J. Rodriguez, and S. J. Wodak, "Automatic definition of recurrent local structure motifs in proteins," *J. Mol. Biol.*, vol. 213, pp. 337–350, 1990.

[6] A. C. Camproux, P. Tuffery, J. P. Chevrolat, J. F. Boisvieux, and S. Hazout, "Hidden markov model approach for identifying the modular framework of the protein backbone," *Protein Engineering*, vol. 12, pp. 1063–1073, 1999.

[7] C. Micheletti, F. Seno, and A. Maritan, "Recurrent oligomers in proteins: An optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies," *Proteins*, vol. 40, pp. 662–674, 2000.

[8] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt, "Small libraries of protein fragments model native protein structures accurately," *J. Mol. Biol.*, vol. 323, pp. 297–307, 2002.

[9] I. Friedberg and A. Godzik, "Connecting the protein structure universe by using sparse recurring fragments," *Structure*, vol. 13, pp. 1213–1224, 2005.

[10] A. Joseph, G. Agarwal, S. Mahajan, J. Gelly, L. Swapna, B. Offmann, F. Cadet, A. Bornot, M. Tyagi, H. Valadié *et al.*, "A short survey on protein blocks," *Biophys. Rev.*, vol. 2, pp. 137–145, 2010.

[11] C. Wallace and D. Boulton, "An information measure for classification." *Comp. J.*, vol. 11, pp. 185–194, 1968.

[12] C. Wallace, *Statistical and Inductive Inference by Minimum Message Length*. Berlin: SpringerVerlag, 2005.

[13] T. Bayes and R. Price, "An essay towards solving a problem in the doctrine of chance," *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, 1763.

[14] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Technical Jrnl.*, vol. 27, pp. 379–423, 1948.

[15] R. Artali, G. Bombieri, F. Meneghetti, G. Gilardi, S. Sadeghi, D. Cavazzini, and G. Rossi, "Comparison of the refined crystal structures of wild-type (1.34 Å) flavodoxin from *Desulfovibrio vulgaris* and the s35c mutant (1.44 Å) at 100 K," *Acta Crystallogr D. Biol Crystallogr*, vol. 58, pp. 1787–92, 2002.

[16] W. Li, "Simple method for constructing phylogenetic trees from distance matrices," *Proc. Nat. Acad. Sci.*, vol. 78, pp. 1085–1089, 1981.

[17] P. Mahalanobis, "On the generalised distance in statistics," *Proc. Nat. Inst. of Sci. of India*, vol. 2, pp. 49–55, 1936.

[18] A. Lesk, "NAD-binding domains of dehydrogenases," *Curr. Opin. Struct. Biol.*, vol. 5, pp. 775–783, 1995.

[19] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures." *J. Mol. Biol.*, vol. 247, pp. 536–540, 1995.

[20] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton, "CATH – a hierarchic classification of protein domain structures." *Structure*, vol. 5, pp. 1093–1108, 1997.

[21] T. Przytycka, R. Aurora, and G. Rose, "A protein taxonomy based on secondary structure," *Nat. Struct. Biol.*, vol. 6, pp. 672–682, 1999.