A Geometric Representation of Protein Sequences

Shengyin Gu
Institute for Data Analysis and Visualization
Department of Computer Science
University of California
Davis, California, USA
gus@cs.ucdavis.edu

Olivier Poch
Laboratoire de Biologie Structurale
IGBMC (CNRS/INSERM/ULP)
Strasbourg, France
poch@titus.u-strasbg.fr

Bernd Hamann
Institute for Data Analysis and Visualization
Department of Computer Science
University of California
Davis, California, USA
hamann@cs.ucdavis.edu

Patrice Koehl
Genome Center
University of California
Davis, California, USA
koehl@cs.ucdavis.edu

Abstract

The amino acid sequence of a protein is the key to understanding its structure and ultimately its function in the cell. This paper addresses the fundamental issue of encoding amino acids in ways that the visualization of protein sequences facilitates the decoding of its information content. We show that a feature-based representation in a three-dimensional (3D) space derived from substitution matrices provides an adequate representation from which the domain content of a protein can be predicted. In addition, we show that each dimension of the feature space can be related to a physical property of the amino acids.

1. Introduction

The genetic information encoded in the genome of an organism represents the blueprint for its development and activity; its implementation depends on the functions of the corresponding gene products (i.e., nucleic acids and proteins). Among these products, proteins play a central role as they catalyze most biochemical reactions, and are responsible, among other functions, for the transport of nutrients and for signal transmission within and between cells. It is well-known that proteins function because they adopt a unique native 3D conformation. While a direct relationship between sequence similarity and conservation of 3D structure has been clearly established for proteins [3], the relationship between their 3D structures and functions is much more complex [39]. This complexity calls for more rigorous descriptions of molecular and cellular functions, and a better understanding of sequence-structure-function relationships. Efforts to unravel the latter currently focus on protein sequence analysis, as a consequence of the wealth of sequence data resulting from various genome projects. Data produced by these projects have already lead to significant improvement in predictions of both 3D structures and functions [39]. However, we still stand at the dawn of understanding the information encoded in the sequence of a gene. In this paper, we focus on protein sequence representations and show how visualization can play a role in decoding gene information content.

Proteins are heteropolymer chains of amino acids. The order in which amino acids appear defines the primary sequence of a protein. Amino acids are usually labeled using a one-letter code, and sequences are correspondingly represented as a usually long string of letters. This representation has proved very valuable, especially in the context of sequence comparisons that are performed using string matching algorithms. It does however carry limitations: letters alone poorly represent the physical and chemical properties of amino acids and as such are usually difficult to decipher. Computer programs that represent protein sequences often resort to different coloring schemes to facilitate their interpretation (e.g., ClustalX for mul-

tiple sequence alignments (MSA)[35]), or to increase their information content (e.g., the SAS server that encodes the 3D structure of a protein on its sequence using a color coding [23]). The addition of well-chosen colors improves the readability of MSA [35]; their importance however for deciphering single sequences remains limited. Note that a coloring scheme ultimately corresponds to adding dimensions to the representation of a protein sequence or a MSA in order to help decipher its information content. This concept of increased dimensions was applied to MSA using Hilbert curves [32]. It can naturally be extended to the idea of a geometric representation and visualization of individual sequences.

The concept of geometric representation of protein sequences was originally introduced by Swanson [33] who proposed a two-dimensional vector representation of the standard twenty amino acids, based on Dayhoff's mutation matrix [31]. In Swanson's representation, the two coordinates of the vectors coincide with size and hydrophobicity. Protein sequences are visualized by concatenating the vectors representing each amino acid types, yielding a vector representation of proteins (VRP). Since the original work of Swanson, other geometric representations of protein sequences have been proposed. Among those, we mention the vector diagram introduced by Yamamoto and Yoshikura [40], which represents each amino acid according to its hydrophilicity and propensity to belong to different types of secondary structures (*beta*-strands and turns). The Zp plot introduced by Feng et al. [9] represents a protein sequence in 3D space based on its hydrophobic, polar and charged residue content. The Zp plot is in fact a graphical extension of PHYSEAN, a physical sequence analysis software that takes into account physical, chemical and biological properties of amino acids [19]. Maetschke and colleagues [20] described a series of multi-dimensional encoding of amino acids, concluding that an extension of the VRP introduced by Swanson [33] to higher dimensions performed the best in identifying putative cleavage sites in proteins.

All the methods referenced above share the idea of moving away from a simple representation of a protein sequence as a string of letters, encoding instead each amino acid as a set of values representing some of its properties. This paper draws from this concept and describes a feature-based representation of protein sequences, in which each amino acid is encoded by a unique vector of features. Our approach differs from the existing approaches described above in the way we construct our 3D vectors. The 3D vectors we compute are such that each of the three dimensions encodes a physical property of the amino acids. Key to our approach is the use of the graphical properties of our geometrical representation to identify properties of the sequence considered. We show preliminary applications to the identification of domains within protein sequences. This paper presents work in progress and more details will be provided later. In section 2, we describe 3D feature vectors for representing amino acids based on substitution matrices. Section 3 presents how these vectors can be used to represent entire sequences, as well as applications of these representations. In section 4, we conclude and allude to other applications of our graphical representation of protein sequences.

2. A Geometric Representation of Amino Acids

String representation of protein sequences is usually uninformative and can only be interpreted through the trained eyes of a protein chemist who can implicitly visualize the chemical structure of the amino acids, or by a program in which this chemical information has been encoded. One way to improve upon this is to encode properties to each amino acid representation. Swanson [33] pioneered a vector representation for protein sequences, in which each amino acid is encoded into a 2D vector whose coordinates correspond to size and hydrophobicity.

We draw from this original idea and represent amino acids as 3D vectors, in which each dimension is a feature of the amino acid. Our goal is to incorporate as many properties of an amino acid as needed into a geometric representation that allow us to visualize protein sequence properties. These properties can then be analyzed directly visually by a human, or through standard geometric procedures. This is a generalization of the 3D encoding proposed in BLOMAP [20]. In this paper, we describe one possible set of features, derived from substitution matrices. Note that the same concept can accommodate other features, such as Chou and Fasman propensities [4] of amino acids to belong to secondary structures.

2.1. Constructing Feature Vectors Based on Similarity Matrices

Common measures of similarities between amino acids are usually presented in the form of a substitution matrix, which stores the odds that any given amino acid can be replaced by any other. Schwartz and Dayhoff [31] were the first to compile such a matrix, using 71 groups of closely related proteins (i.e., with more than 85% pairwise sequence identity), and collecting the data of point accepted mutations, or

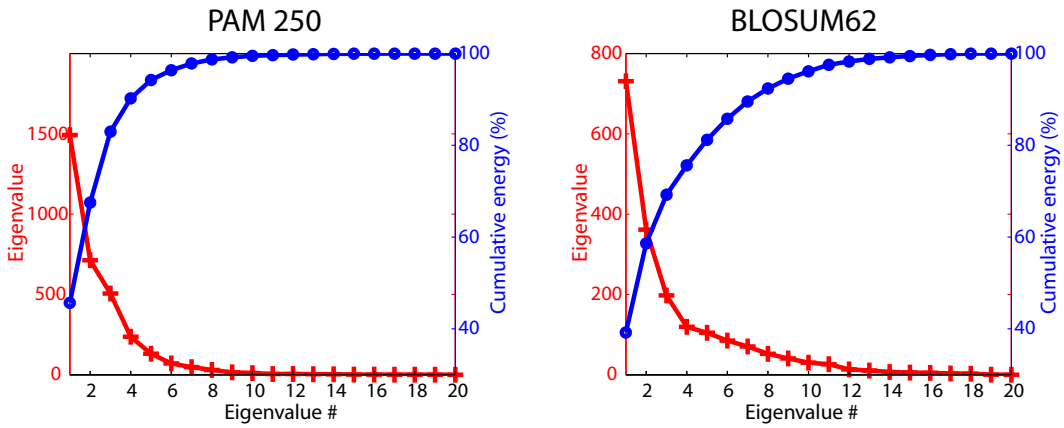

Figure 1. Principal component analysis of substitution matrices. The twenty eigenvalues (+, left axis) of the PAM250 (left panel) and BLOSUM62 (right panel) substitutions matrices as well as their cumulative energies (o, right axis) are plotted in decreasing order of amplitudes. The largest three eigenvalues account for 80% and 70% of the energy (or information content) of PAM250 and BLOSUM62, respectively.

PAMs. Henikoff and Henikoff [15] extended this concept to include more divergent sequences and generated the BLOSUM matrices. Several matrices have been derived, corresponding to different cutoffs in the accepted sequence identity within the BLOCKs. For example, BLOSUM62 is a substitution matrix derived from protein sequence alignments in which the sequences are at least 62% identical; it is considered to provide good performance for database search.

Substitution matrices describe each amino acid with a set of twenty numerical values (sometimes referred to as amino acid index [36]), henceforth defining a twenty-dimensional space. While such a high-dimensional space is useful for computer-guided sequence alignment methods, it is impractical for any form of visualization. Swanson was the first to embed the space corresponding to the original PAM matrix MDM78 into a plane, using a principal component analysis (PCA) approach [33]. More recently, Maetschke et al. [20] embedded the BLOSUM62 matrix into five dimensions, using the Sammon's projection technique [30], noticing that three dimensions already produce a reasonably good approximation. To further characterize which dimension is appropriate for visualizing the information content of BLOSUM62, we repeated the embedding of both PAM250 (which is very similar to the original MDM78) and BLOSUM62, using a PCA. Results are shown in Figure 1. Swanson [33] and Maetschke et al. [20] used PCA and Sammon mapping, respectively. Their methods first convert the substitution matrix into a "distance" matrix, by exponentiation of the scores included in the matrix. We kept the substitution matrix as it is. Each column of this matrix corresponds to a different amino acid, while each row is treated as a probe of a property of that amino acid. In the PCA analysis, the substitution matrix is first centered, and then the eigenvalues and eigenvectors of its covariance matrix are computed. We have found that the three largest eigenvectors account for 82% and 70% of the total "energy" (or information content) of PAM250 and BLOSUM62, respectively. These results agree with those of Swanson [33] and Maetschke et al. [20].

2.2. Information Content of Amino Acids 3D Feature Vectors

The entropy value of a substitution matrix is an information theoretic value that measures the information content [1]. In Figure 2 we show that the energy of the three largest eigenvalues of a substitution matrix is correlated with its entropy value, with correlation values of -0.91 and 0.46 for PAM and BLOSUM matrices, respectively. The difference in sign stems from the definitions of the matrices. PAM matrices with low ID numbers are computed from alignments of highly similar sequences, and as such are comparable with BLOSUM matrices of high ID numbers. They are both designed for comparisons of closely related sequences. Reversely, BLOSUM matrices with low ID numbers and PAM matrices with high ID numbers are designed

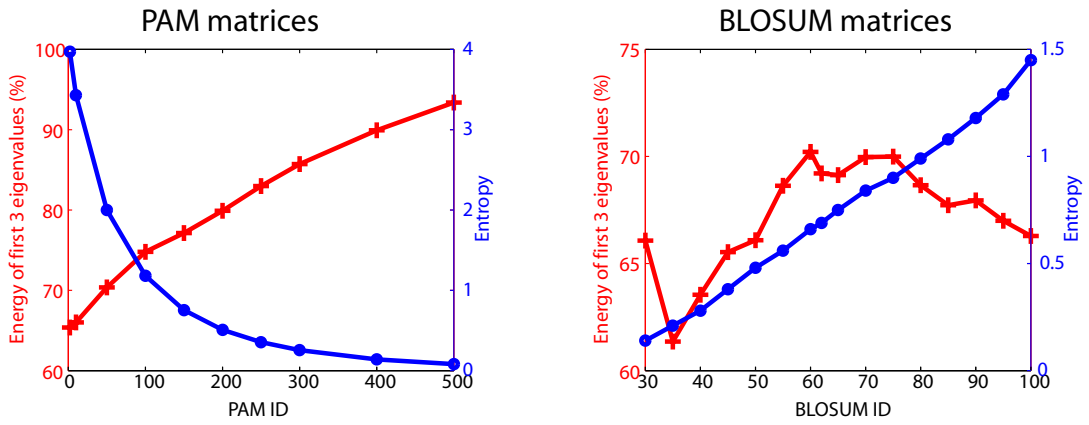

Figure 2. Information content of substitution matrices. The cumulative energy of the three largest eigenvalues (+, left axis) and the entropy (o, right axis) of substitutions matrices are plotted as functions of the matrix ID.

for comparisons of distantly related proteins. Interestingly, the energies of the three largest eigenvalues of PAM matrices are always higher than the energies of their equivalent BLOSUM matrices (with PAM250 corresponding to BLOSUM45, PAM120 to BLOSUM80 and PAM100 to BLOSUM90, based on entropy comparison).

2.3. BLV62: Information in Each Dimension

BLOSUM62 is the preferred substitution matrix for database search. We focus on this matrix in the following. Each amino acid can be represented as a 3D vector, using its corresponding coordinates in the largest three eigenvectors of the covariance matrix of BLOSUM62. Figure 3 shows these twenty vectors, which we refer to as BLV62, all centered at the origin (with the origin being contained in the bounding box).

It is difficult to interpret the three axes of the BL vectors, as these are mathematically constructed to provide sub-components of the matrices with decreasing energy/information content. We compared the vector containing the coordinates of the twenty amino acids on the first axis corresponding to the BLV62 vectors, with 528 amino acid indices available in the AAIndex database [36]. Five of the 528 indices were selected with a correlation coefficient greater than or equal to 0.95: the "buriability" of Zhou and Zhou [41], an amino acid contact number with a cutoff of 14 Å[24], a normalized hydrophobicity scale [6], and two interactivity scales designed to correlate with hydropathy scales [2]. Note that all these indices are related to amino acid burial and their hydrophobicity. These results are in agreement with the original findings of French and Robson [10], Swanson [33] and Tomii and Kanehisa [36]. Interestingly, this behavior differs from the results described by Kinjo and Nishikawa [17], who performed spectral analysis on substitution matrices compiled from protein structure alignments, including proteins with varying levels of sequence similarities. Using the same AAIndex database that we used [36], Kinjo and Nishikawa showed that at high sequence identities hydrophobicity plays a minor role, and that the "relative mutabilities" of Dayhoff et al. [7] and Jones et al. [16] dominates. As BLOSUM62 is derived from blocks of sequences with more than 62% sequence identify, it qualifies as a high sequence identity substitution matrix. The difference between our results and those of Kinjo and Nishikawa is unclear. The best correlations between the second and third axes of the 3D BLOSUM62 vectors and an amino acid index contained in AAIndex are 0.77 and 0.75, respectively. The second axis is found to correlate well with average non-bonded energies [25], which is related to size. Interestingly, the third axis is found to correlate with computed alpha-helix propensities [18], as well as with statistics on turns in proteins [5].

3 Applications: A Geometric Representation of Protein Sequences

A sequence of a protein describes the succession of its amino acids from its N-terminal end to its C-terminal end. In the section above, we have shown that

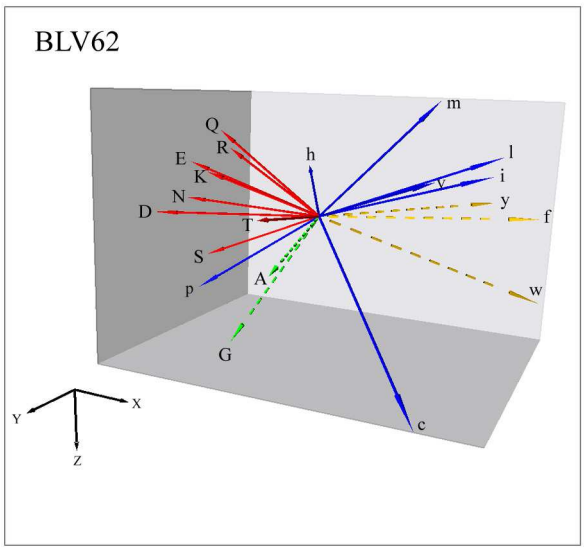

Figure 3. 3D vectors based on BLOSUM62: BLV62. This plot represents the similarity between amino acids as encoded by the BLOSUM62 matrix. The geometric proximity of amino acids correspond to their known chemical similarities. To highlight this fact, we show the known polar residues (Q, R, E, K, N, D, T, S) in solid vectors with upper-case labels, the hydrophobic residues in solid vectors with lower-case labels (m, v, l, i, h, p, c) and aromatic residues in dashed vector with lower-case labels (y, f, w). Note that the two small amino acids, A and G (in dashed vectors with upper-case labels), stand out. Note also that Cysteine (c), though non-polar, differs from other amino acids based on its ability to form disulphide bridges, usually highly conserved in proteins.

representing amino acids as 3D vectors improves the decoding of their properties. We extend this geometric concept to the representation of the whole sequence of a protein by direct "head-to-tail concatenation" of the vectors representing its constituent amino acids. A protein sequence then becomes a polyline in 3D space, which we refer to as the protein's *3D trace*. We describe one application of such a representation, namely the detection of domains in long protein sequences.

Large proteins do not contain a single large hydrophobic core, probably because of limitations in their folding kinetics and stability. Single compact units of more than 500 amino acids are rare. Large proteins in fact are usually organized into units with sizes around 200-300 residues, referred to as domains [28, 27]. Interestingly, while the concept of domains in proteins is well-established, there is no consensus definition of what a domain is. A domain is either defined based on sequence (regions that display a significant level of sequence similarity), function (the minimal part of a gene that is capable of performing a function) or structure (compact, spatially distinct units of protein structure) [38]. When the structure of the protein is known, its domains are usually defined by a combination of visual inspection of the structure with automated methods that take into account the globular nature of domains (for a review of existing methods, see [38]). It would be of practical interest to delineate domain boundaries in protein sequence alone, as this information would facilitate structure and function prediction. Current methods for domain prediction rely mostly on MSAs [14, 12]; these methods perform poorly on orphan sequences. Other approaches include analysis of secondary structure prediction [21], sidechain entropy [11], clusters of hydrophobic residues [12] or amino acid composition in the linker regions [13, 34].

We propose to visualize domain transition in proteins using our 3D representation of protein sequences, their 3D traces. We illustrate our approach on the sequence of *Prf*, a disease resistance gene in tomatoes [29]. The *Prf* gene encodes for a protein, PRF, of approximately 1800 residues, that contains at least three domains: an N-terminal domain, of which little is known, a nucleotide binding domain (NBS), and a Leucine Rich Repeat domain (LRR) [29]. PRF is a member of the large family of NBS-LRR proteins (for

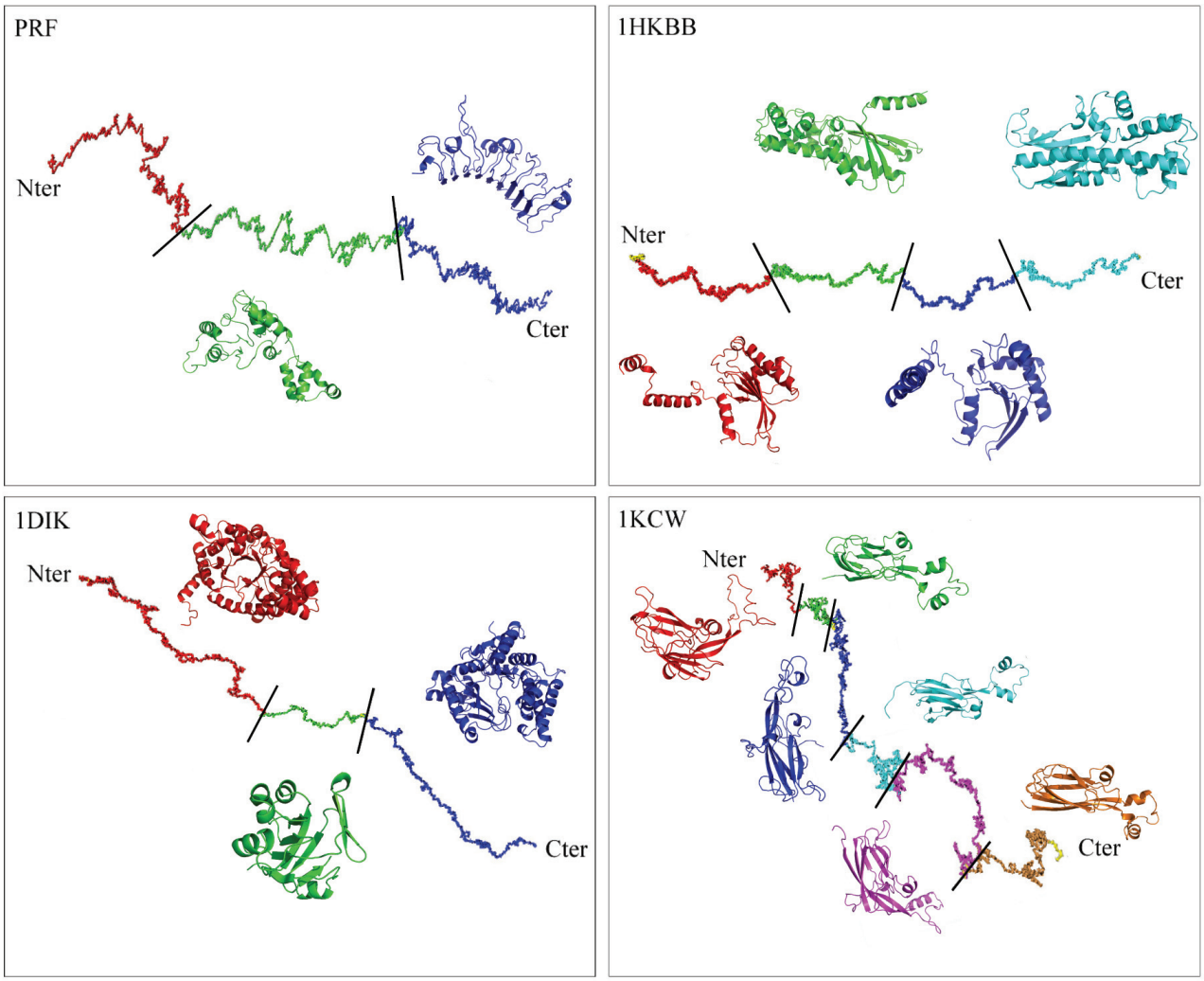

Figure 4. 3D sequence traces detect domains in proteins. **The 3D sequence trace of PRF, a disease resistance gene of tomatoes, 1DIK, a pyruvate phosphate dikinase, chain B of 1HKB (1HKBB), a hexokinase type I, and 1KCW, a ceruloplasmin are shown. Separation between known domains of these proteins (based on sequence analysis for PRF, and based on the SCOP classification of protein structures for 1DIK 1HKBB and 1KCW) are shown as line segments; they usually correspond to change in directions in the 3D trace. 1HKBB is an exception (see text for details). Model structures for each known domains are shown in cartoon representation. These models were generated using pymol (**http://www.pymol.org**) .**

review, see [22]). The 3D trace representation of PRF using the BLOSUM62 3D vectors is shown in Figure 4. Each domain transition in PRF is revealed through a change in the overall direction of its 3D sequence trace.

We applied the same procedure to three large proteins whose domain definitions are known: pyruvate phosphate dikinase (PDB code 1DIK; 884 residues), human brain hexokinase type I (PDB code 1HKB; 914 residues), and human ceruloplasmin (PDB code 1KCW; 1040 residues). Results are shown in Figure 4

According to SCOP, 1DIK contains three domains: an ATP binding domain (alpha+beta ATP grasp domain) from residue 1 to 376, a phosphohistidine domain (beta/beta/alpha domain) from residue 377 to 505, and a pyruvate kinase domain (alpha/beta tim barrel) from residue 510 to 884; all three domains are

clearly delineated on the 3D trace.

1HKB is a good example of the current limits of the use of the 3D trace for domain identification. Chain B of 1HKB contains four consecutive ribonuclease H-like motifs (alpha+beta domains): these are more difficult to distinguish based on the 3D trace only, as there are no significant changes in the overall directions of the trace.

Interestingly, results are much better on 1KCW, which contains six consecutive rubredoxin-like domains (all beta domains), in particular for domains three, four, five and six.

4 Conclusions and Perspectives

The amino acid sequence of a protein is the key to understanding its structure and ultimately its function in the cell. We have shown that amino acids can be encoded by 3D vectors, thereby allowing us to generate a geometric representation of their properties. We derived one set of 3D vectors, namely the BLV62, based on the BLOSUM62 substitution matrix, respectively. Concatenation of the vectors corresponding to the successive amino acids in a protein sequence generates a 3D trace.

Substitution matrices provide the odds that any given amino acid can be replaced by any other for a given amount of time. Among all existing substitution matrices, BLOSUM62 occupies a special position as it is the default matrix used for protein sequence database search. Using PCA, we have shown that BLOSUM62 can be projected into a 3D space, without significant loss of information. The three principal axes correlate best with hydrophobicity, number of contacts (which relates to size), and propensities to belong to an α-helix or a turn, respectively. While the dependence for the first two axes was already described for MDM78 [33], the dependence of the third axis on secondary structure was not previously described. We believe that this is of importance as it clearly adds a structural information onto the sequence representation. We will further study this correlation.

We have focused this research on the visualization of protein sequences in 3D space, using the novel 3D trace concept. Simple visual analysis of this 3D polygonal representation provides access to structural properties of the corresponding protein, such as its partitioning into domains. We will extend this approach. In particular we are interested in developing a quantification of the information contained in the 3D trace. We will consider applying wavelet transforms to extract geometric signatures of the 3D trace. Wavelet transforms have already been applied to the analysis of protein sequences (e.g., [8, 26]. In these approaches, the protein sequences are converted to numerical sequences, using an amino acid electron-ion interaction potential [37]. Interestingly, this amino acid index (included into AAIndex), is not correlated to the three principal axes of BLOSUM62 to any significant extend.

There are many ways to combine the 3D vectors corresponding to the amino acids into a complete representation for the entire sequence of a protein. We have relied on probably the simplest of such representations, i.e., the concatenation of the vectors. We will further investigate which other graphical representations support highly effective visual and quantitative extraction of the information contained in a protein sequence.

In this paper, we have encoded amino acids into 3D vectors derived from a substitution matrix (BLOSUM62). Note that this concept can be generalized to other properties. It is possible, for example, to represent each amino acid using a vector that contains its propensities to belong to a helix, a β strand, or a turn. Such vectors, and the corresponding 3D traces, should prove useful for predicting the structural classes of a protein. We are currently developing this representation.

5. Acknowledgments

This work was supported in part by the National Science Foundation under contracts CCF-0625744 and a large Information Technology Research (ITR) grant.

References

[1] S. Altschul. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, 219:555–565, 1991.

[2] U. Bastolla, M. Porto, H. Roman, and M. Vendruscolo. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins: Struct. Func. Bioinfo.*, 58:22–30, 2005.

[3] C. Chothia and A. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J.*, 5:823–826, 1986.

[4] P. Chou and G. Fasman. Prediction of protein conformation. *Biochemistry*, 13:211–245, 1974.

[5] P. Chou and G. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.*, 47:45–148, 1978.

[6] H. Cid, M. Bunster, M. Canales, and F. Cazitua. Hydrophobicity and structural classes in proteins. *Prot. Eng.*, 5:373–375, 1992.

[7] M. Dayhoff. A model of evolutionary changes in proteins. *Atlas of Protein Sequence and Structure*, 5:345–352, 1978.

[8] C. H. de Trad, Q. Fang, and I. Cosic. Protein sequence comparison based on the wavelet transform approach. *Protein Eng*, 15:193–203, 2002.

[9] Z. Feng and C.-T. Zhang. A graphic representation of protein sequence and predicting the subcellular locations of prokaryotic proteins. *Int. J. Biochem. Cell Biol.*, 34:298–307, 2002.

[10] S. French and B. Robson. What is a conservative substitution. *J. Molec. Evol.*, 19:171–175, 1983.

[11] O. Galzitskaya and B. Melnik. Prediction of protein domain boundaries from sequence alone. *Protein Sci.*, 12:696–701, 2003.

[12] R. George and J. Heringa. Snapdragon: a method to delineate protein structural domains from sequence data. *J. Mol. Biol.*, 316:839–851, 2002.

[13] R. George and J. Heringa. An analysis of protein domain linkers: their classification and role in protein folding. *Prot. Eng.*, 15:871–879, 2003.

[14] X. Guan and L. Du. Domain identification by clustering sequence alignments. *Bioinformatics*, 14:783–788, 1998.

[15] S. Henikoff and J. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. (USA)*, 89:10915–10919, 1992.

[16] D. Jones, W. Taylor, and J. Thornton. The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8:275–282, 1992.

[17] A. Kinjo and K. Nishikawa. Eigenvalue analysis of amino acid substitution matrices reveals a sharp transition of the mode of sequence conservations in proteins. *Bioinformatics*, 20:2504–2508, 2004.

[18] P. Koehl and M. Levitt. Structure-based conformational preferences of amino acids. *Proc. Natl. Acad. Sci. (USA)*, 96:12524–9, 1999.

[19] I. Ladunga. Physean: Physical sequence analysis for the identification of protein domains on the basis of physical and chemical properties of amino acids. *Bioinformatics*, 15:1028–1038, 1999.

[20] S. Maetschke, M. Towsey, and M. Boden. Blomap: an encoding of amino acids which improves signal peptide cleavage site prediction. *Asia Pacific Bioinformatics Conference*, pages 141–150, 2005.

[21] R. Marsden, L. McGuffin, and D. Jones. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.*, 11:2814–2824, 2002.

[22] L. McHale, X. Tan, P. Koehl, and R. Michelmore. Plant nbs-lrr proteins: adaptable guards. *Genome Biology*, 7:212, 2006.

[23] D. Milbrurn, R. Laskowski, and J. Thornton. Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Prot. Engineering*, 11:855–859, 1998.

[24] K. Nishikawa and T. Ooi. Radial locations of amino acid residues in a globular protein: Correlation with the sequence. *J. Biochem.*, 100:1043–1047, 1986.

[25] M. Oobatake and T. Ooi. An analysis of non-bonded energy of proteins. *J. Theor. Biol.*, 67:567–584, 1977.

[26] T. Riaz, K.-B. Li, F. Tang, and A. Krishnan. Cmdwave: Conserved motifs detection using wavelets. *In Silico Biology*, 5:0038, 2005.

[27] J. Richardson. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, 34:167–339, 1981.

[28] G. Rose. Hierarchic organization of domains in globular proteins. *J. Mol. Biol.*, 134:447–470, 1979.

[29] J. Salmeron, G. Oldroyd, C. Rommens, and S. S. et al. Tomato prf is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the pto kinase gene cluster. *Cell*, 86:123–133, 1996.

[30] J. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, C-18:401–409, 1969.

[31] R. Schwartz and M. Dayhoff. Matrices for detecting distant relationships. *Atlas of Protein Sequence and Structure*, 5:345–352, 1978.

[32] N. Shah, S. Dillard, G. Weber, and B. Hamann. Volume vizualization of multiple alignment of large genomic dna. In T. Moeller, B. Hamann, and R. Russell, editors, *Mathematical foundations of scientific visualization, computer graphics, and massive data exploration*. Springer Verlag, Heidelberg, Germany, 2007. to appear.

[33] R. Swanson. A vector representation for amino acid sequences. *Bull. Math. Bio.*, 46:623–639, 1984.

[34] T. Tanaka, Y. Kuroda, and S. Yokoyama. Characteristics and prediction of domain linker sequences in multi-domain proteins. *J. Struct. Funct. Genomics*, 4:79–85, 2003.

[35] J. Thompson, T. Gibson, F. Plewniak, F. Jeanmougin, and D. Higgins. The clustalx windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids. Res.*, 25:4876–82, 1997.

[36] K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Prot. Eng.*, 9:27–26, 1996.

[37] V. Veljkovic, I. Cosic, B. Dimitrijevic, and D. Lalovic. Is it possible to analyze dna and protein sequences by the method of digital signal processing? *IEEE Trans. Biomed. Eng.*, 32:337–341, 1985.

[38] S. Veretnik, P. Bourne, N. Alexandrov, and I. Shindyalov. Toward consistent assignment of structural domains in proteins. *J. Mol. Biol.*, 339:647–678, 2004.

[39] J. Watson, R. Laskowski, and J. Thornton. Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, 15:275–284, 2005.

[40] K. Yamamoto and H. Yoshikura. A new representation of protein structure: vector diagram. *CABIOS*, 2:83–88, 1986.

[41] H. Zhou and Y. Zhou. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins: Struct. Func. Genet.*, 54:315–322, 2004.