Computational Biology and Language

4 authors, including:

Madhavi Ganapathiraju
University of Pittsburgh
97 PUBLICATIONS **943** CITATIONS

SEE PROFILE

Narayanaswamy Balakrishnan
Indian Institute of Science
189 PUBLICATIONS **1,746** CITATIONS

SEE PROFILE

Judith Klein-Seetharaman
The University of Warwick
227 PUBLICATIONS **6,718** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Computational Electromagnetics View project

Project Intrusion Detection Systems View project

Computational Biology and Language

Madhavi Ganapathiraju[1], Narayanas Balakrishnan[2],
Raj Reddy[3], and Judith Klein-Seetharaman[4]

[1] Carnegie Mellon University, USA
madhavi+@cs.cmu.edu
[2] Indian Inst. of Science, India & Carnegie Mellon Univ, USA
balki@serc.iisc.ernet.in
[3] Carnegie Mellon University, USA
rr+@cmu.edu
[4] Carnegie Mellon University & University of Pittsburgh, USA
judithks@cs.cmu.edu

1 Introduction

Current scientific research is characterized by increasing specialization, accumulating knowledge at a high speed due to parallel advances in a multitude of sub-disciplines. Recent estimates suggest that human knowledge doubles every two to three years – and with the advances in information and communication technologies, this wide body of scientific knowledge is available to anyone, anywhere, anytime. This may also be referred to as ambient intelligence - an environment characterized by plentiful and available knowledge. The bottleneck in utilizing this knowledge for specific applications is not accessing but assimilating the information and transforming it to suit the needs for a specific application. The increasingly specialized areas of scientific research often have the common goal of converting data into insight allowing the identification of solutions to scientific problems. Due to this common goal, there are strong parallels between different areas of applications that can be exploited and used to cross-fertilize different disciplines. For example, the same fundamental statistical methods are used extensively in speech and language processing, in materials science applications, in visual processing and in biomedicine. Each sub-discipline has found its own specialized methodologies making these statistical methods successful to the given application. The unification of specialized areas is possible because many different problems can share strong analogies, making the theories developed for one problem applicable to other areas of research. It is the goal of this paper to demonstrate the utility of merging two disparate areas of applications to advance scientific research. The merging process requires cross-disciplinary collaboration to allow maximal exploitation of advances in one sub-discipline for that of another. We will demonstrate this general concept with the specific example of merging language technologies and computational biology.

Communication between researchers in these disparate fields is facilitated through use of analogies. Specifically, the analogy between words and their meaning in speech and language processing on one hand, and the mapping between

biological sequences to biological functions on the other, has proven particularly useful. Recent reviews of applications of linguistic approaches to computational biology in general can be found in references [1, 2]. Thus, we will first only briefly explain the analogy between biological sequence and natural language processing in general and then focus the remainder of the review on the use of language technologies to identify the functional building blocks in protein sequences, i.e. the "words" of "protein sequence language". Since it is not known what would be the best word equivalent, we will first describe what types of word equivalents and vocabularies have been explored using the example of one specific area of application, secondary structure prediction and analysis. In some areas of applications of language technologies to language, the words are also not known, for example in speech recognition. In these applications, identifying functional building blocks is a signal processing task and we will describe the analogy to protein sequences from this perspective. This includes first introducing proteins and protein structure in comparison to the terms used in speech processing, followed by a presentation of one specific application of signal processing techniques in computational biology, namely transmembrane helix structure prediction. This will be brought into the broader context of other applications of language technologies to the same task. Finally, we will present a sampling of a few other examples of applying language technologies to the computational biology of proteins. Additional examples can be found referenced on the website of the Center for Biological Language Modeling (BLM) in Pittsburgh, USA [3].

2   Use of Language Technologies in Computational Biology

Most functions in biological systems are carried out by proteins. Typical functions include transmission of information, for example in signaling pathways, enzymatic catalysis and transport of molecules. Proteins also play structural roles such as formation of muscular fiber. Proteins are synthesized from small building blocks, amino acids, of which there are 20 different types (see below). The amino acids are connected to form a linear chain that is arranged into a defined three dimensional structure. The precise interactions between amino acids in the three dimensional structure of a protein are the hallmark of the functions that they are able to carry out. For example, these interactions allow proteins to make contacts with small molecule ligands such as drugs. Figure 1 shows an example protein, lysozyme, to which an inhibitor ligand is bound (shown in magenta). Thus, knowing the three-dimensional shape of proteins has implications not only for the fundamental understanding of protein function, but also for applications such as drug design and discovery.

Obtaining three dimensional structures of proteins experimentally is not straight forward. X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy can accurately determine protein structures; but these methods are labor intensive, time consuming, and for many proteins are not applicable at all. Therefore, predicting structural features of proteins from a sequence is an im-
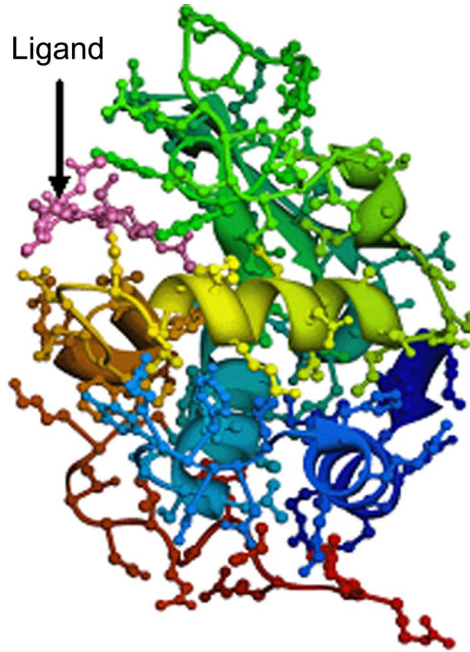
Fig. 1. An example of a protein: Lysozyme (Protein Data Bank code 1HEW). The protein is colored in rainbow color from one end to the other end. The main chain is highlighted by ribbons. Side chains extending from the main chain are shown as ball and stick representations. The magenta colored molecule is the inhibitor ligand, tri-N-acetylchitotriose. This figure illustrates how a linear protein chain folds up into a three dimensional structure thereby creating a binding site with which ligand molecules can interact. All protein figures in this paper have been created using Chimera [4].

portant topic in computational biology. Understanding the structure, dynamics and function of proteins strongly parallels the mapping of words to meaning in natural language. This analogy is outlined schematically in Fig. 2. The words in a text document map to a meaning and convey rich information pertaining to the topic of the document. Similarly, protein sequences also represent the "raw text" and carry high-level information about the structures, dynamics and functions of proteins. This information can be extracted to obtain an understanding of the complex interactions of protein within biological systems. Availability of large amounts of text in digital form has led to the convergence of linguistics with computational science, and has resulted in applications such as information retrieval, document summarization and machine translation. Thus, even though computational language understanding is not yet a reality, data availability has allowed us to obtain practical solutions that have a large impact on our lives. In direct analogy, transformation of biology by data availability opened the door to convergence with computer science and information technology.

Many of the hallmarks of statistical analysis of biological sequences are similar to those of human languages. (i) Large data bodies need to be analyzed
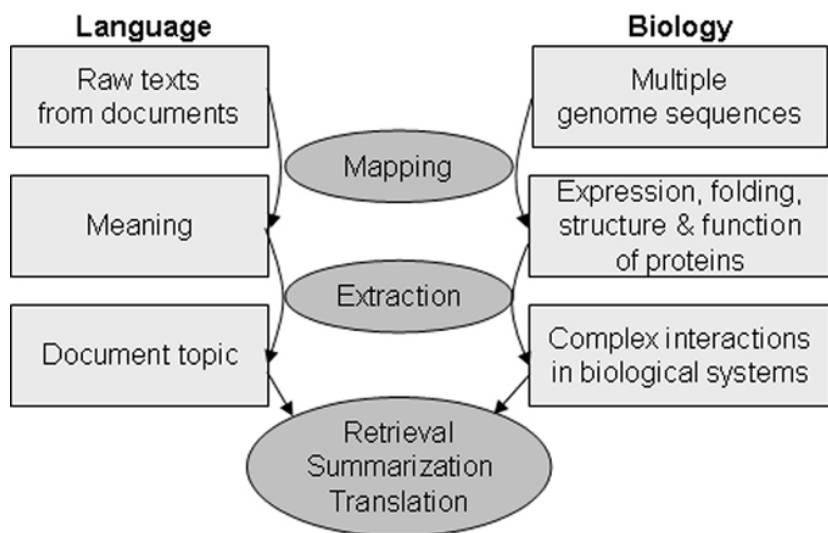
Fig. 2. Analogy between natural language and "protein sequence language". Words combine into sentences to convey meaningful content. Processing words in a document can convey the topic of the document and its meaningful content. Biological sequences are analogous to the raw texts, processing which would yield higher level information on the behavioral properties of the physical entities corresponding to the sequences.

in both cases. (ii) Fundamental units of human languages include higher order structures, paralleled by domains, subunits or functionally linked proteins. (iii) Computer-based derivation of meaning from text is analogous to the prediction of structure and function from primary sequence data. Therefore, the analogy has led to the wide application of methods used in language technologies to the study of biological sequences [1, 5, 6]. Some examples for the use of linguistic approaches for bioinformatics can be found in refs. [1, 5–12]. Recently, probabilistic language models have been used to improve protein domain boundary detection [13], to predict transmembrane helix boundaries [14] and in genome comparison [15, 16]. Finally, latent semantic analysis, a technique used in text summarization, has been used for secondary structure prediction [17] and topic segmentation of text or radio speech. The feature prediction methods of Yule's Q-statistic [18] and mutual information [19] have been applied to the membrane protein boundary prediction problem.

2.1   Protein Sequence Language

Like strings of letters and words in a text, protein sequences are linear chains of amino acids. The amino acid is one of the fundamental building blocks in protein sequences. This is illustrated schematically in Fig. 3. Each amino acid has a common component shown in Fig. 3A. In the protein chain, the amino acids are connected through this component forming the backbone of the protein. The side chains, represented by R in Fig. 3A, can be one of 20 different

types, corresponding to different chemical structures (labeled as (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y)). The main and side chains in a protein are shown in an example in Fig. 3B. Side-chains themselves have components in common among each other based on their chemical composition. Thus, we could also consider smaller chemical units than the amino acids as the functional building blocks of proteins. This would correspond to the vocabulary shown in Fig. 4. These are in fact more fundamental units than the amino acids themselves, because mutations, i.e. replacements of amino acids in protein sequences, that only exchange one single chemical group, e.g. from phenylalanine to tyrosine (OH group) can have detrimental effects on protein function.

At the other end, there are also cases where a single amino acid is not sufficient to convey a specific "meaning", but a group of amino acids does, generally referred to as a functional motif. For example, the triplet D/E R Y is a conserved motif in a specific protein family (the G-protein coupled receptors) known to encode the ability to interact with another protein (the G-protein). Finally, amino acid sequences can be replaced without loss in function, as individual amino acids or as groups of amino acids. For example, in the above triplet, the first position is not fully defined - it can be either D or E. This is due to the chemical nature of the amino acids: D and E although having different side chains, share a number of properties, most importantly, negative charge in this case. Thus, the biological vocabulary is much more flexible than the human vocabulary, because it is defined through properties with several different chemical meanings and not a single meaning as in the case of the 26 letters. There are hundreds of different scales of properties of amino acids, including size, hydrophobicity, electronic properties, aromaticity, polarity, flexibility, secondary structure propensity and charge to name just a few (see e.g., the online databases PDbase [20] and ProtScale [21]). Thus, although the 20 amino acids are a reasonable starting point to define building blocks in protein sequences, smaller, larger or uniquely encoded units may often be functionally more meaningful.
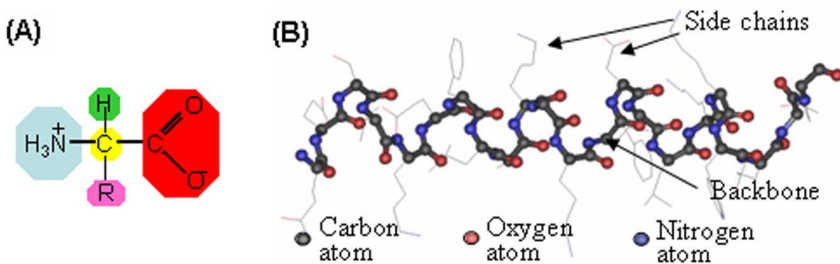

Fig. 3. (A) Chemical composition of amino acids. The composition common to all amino acids consists of a main carbon atom C_α (yellow), NH_3^+ group (blue), carboxyl group COO- (red), hydrogen atom (green) and a sidechain R (pink). The first three, along with the C_α, are common to all amino acids, where as the side chain R is different for each amino acid [17]. (B) Protein segment. A small protein segment with the composing main chain atoms is shown in ball-and-stick model. Side chains attached to the C_α are shown in grey wire frame.

-C̈-	=C̈aromatic	C̈H-	-CH$_2$-
-CH$_2$^{ring}-	-CH$_3$	=CH^{aromatic}-	C̈H^{ring}
C̈=O	-COO$^-$	=N-	-NH-
-NH$_2$	=NH$_2$$^+$	-NH$_3$$^+$	-NH^{ring}-
-OH	-SH	-S̈	

Fig. 4. Chemical Group Vocabulary: The basic chemical groups that form the building blocks of the amino acids are shown. The chemical group in each cell in the figure forms one word in the vocabulary. Thus, the size of chemical group vocabulary is 19. This vocabulary has been studied in the context of secondary structure analysis by Ganapathiraju et al [17].
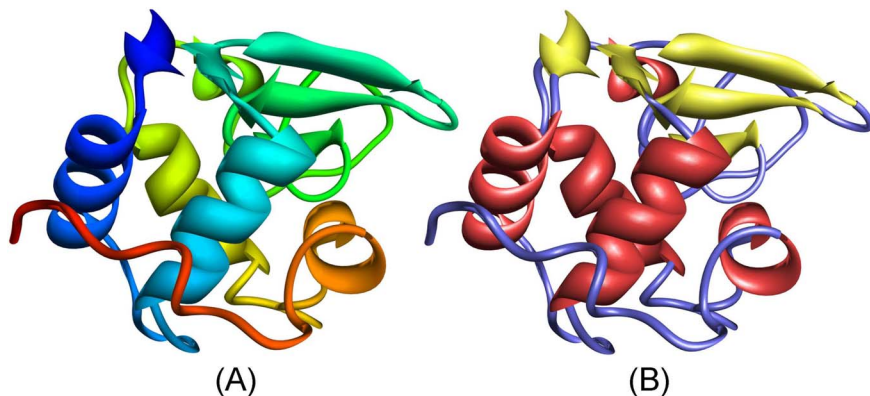

(A)   (B)

Fig. 5. Secondary structure elements in Lysozyme (PDB ID: 1HEW): Three dimensional structure of a protein is composed of smaller units (secondary structure). (A) The chain can be followed by guide of the rainbow colors. (B) The same view of the protein as in A is shown, but repeating elements helix (red), sheet (yellow) and turns and flexible loops (violet) are highlighted.

To study the effect of varying the vocabulary and alphabet on a typical computational biology task, secondary structure analysis and prediction, we first investigated different units for this task. Secondary structure refers to regular units of structure that are stabilized by molecular interactions between atoms within the protein, the most important interaction being the so-called Hydrogen (H) Bond. There are 7 distinct secondary structures, broadly called helix, sheet, turn and loop structures. In helix types, the designating secondary structure is formed due to H-bonds between carbonyl group and amino group of every 3rd, 4th or 5th residues, and these are called $3_{10}-$helix, $\alpha-$helix and $\pi-$helix respectively. A strand is a unit that shares long range hydrogen-bond interaction

with another strand. Two or more such interacting strands form what is called a sheet. A turn is defined as a short segment that causes the protein to bend. Loop or coil region has no specific secondary structure. Commonly, the 7 groups are reduced to 3 groups, helix, strand and loop (shown in an example in Fig. 5). To study the relevance of different vocabularies for secondary structure formation, we used the following vocabularies: (1) chemical building blocks of amino acids, (2) single amino acids from the 20 amino acid alphabet and (3) reduced alphabets based on similarities between physico-chemical properties of amino acids [17]. Latent Semantic Analysis (LSA) was used to decipher the role of the vocabulary for this task, because it is a natural language processing method that is used to extract hidden relations between words [22]. We should therefore be able to study the effects of different vocabularies on secondary structure without introducing artifacts through the differences in size and geometry in the different units studied. LSA captures semantic relations using global information extracted from a large number of documents and can therefore identify words in a text that are synonymous even when such information is not directly available. LSA was then applied to characterize segments of protein sequences with a given type of secondary structure, helix, strand or loop. Each segment was represented as a bag-of-words vector traditionally used in document processing. The word-document matrix comprising all the protein segment vectors was transformed into Eigenspace through singular value decomposition, and the protein segments were compared to each other in terms of vector representation in singular space. To compare the usefulness of this representation, protein segments were separated into training and test sets and the secondary structure of each segment in the test set was predicted based on the secondary structure of its nearest neighbors in the singular space from among the training set. When representing the amino acid sequences using the three different vocabularies, we observed that different vocabularies are better at characterizing different structure types. Helices and strands are best characterized using amino acid types with LSA, and coils are characterized better with amino acids as vocabulary and using the simple word-document matrix analysis (called VSM [23]) without LSA. Average 3-class prediction (Q_3) was found to be best using chemical groups as vocabulary and using VSM. The results demonstrate that word-document matrix analysis and LSA capture sequence preferences in structural types and can distinguish between the "meanings" of vocabularies for protein secondary structure types. Furthermore, protein sequences represented in terms of chemical groups and amino acid types provide more clues on structure than the classically used amino acids as building blocks [17].

As shown by the above study [17] and many previous studies [24], single amino acid propensities have limited ability to predict secondary structure elements. It was therefore investigated if larger segments composed of several amino acids, so-called k-mers or n-grams of amino acids are more appropriate units of protein sequence language with respect to their meaning for secondary structure [25]. However, this study found that n-grams do not capture secondary structure propensity of protein segments well. This is due to the fact that n-gram

features do not encode the types of amino acid substitutions typical for protein sequences exemplified in the motif example above. Therefore, it was investigated if a compact representation of position specific n-grams as $x\{-|+\}N$, where x is the n-gram, $\{-|+\}$ indicates whether it occurs before or after the residue under question, and N is the distance from this residue to the n-gram, may be a better representation of the protein sequence. The analogy to language can be found when classifying documents into possible topics. This task also requires identification of crucial words that can discriminate between possible topics. For example, the word 'ball' can discriminate between "science" and "sports" topics but cannot distinguish between "cricket" and "football" topics. Advances in topic detection methods for text documents have resulted in some reliable methods to identify such discriminating words. In the context of protein secondary structure prediction, there are also position specific propensities of amino acids in different secondary structure types and therefore topic detection algorithms are directly applicable to secondary structure prediction at the residue level. The application of the context-sensitive vocabulary provided results that are comparable to the current state-of-the-art methods using "black-box" classification approaches, in particular neural networks, with Q_3 accuracy of about 70%. The advantage of the use of the context-sensitive vocabulary over these "black-box" methods is that it allows analysis of the word-association matrix with singular value decomposition to identify co-occurring word pairs, corresponding to regular expressions with a specific "meaning" for secondary structure. For example, one of the most highly associated word pair corresponded to the pattern "CPxxAI". The pattern describes the loop region at the C-terminal end of a beta-sheet. Thus, the context-sensitive vocabulary encodes some of the complex dependencies between amino acids that determine formation of secondary structure.

3   Identification of Functional Building Blocks in Proteins as a Signal Processing Task

The lack of knowledge on what are the break points separating words from each other is not new to the language arena. In fact, it is found in many speech applications. In a spoken sentence, words are not separated from each other by spaces as in written text. Thus, automatic speech analysis and synthesis methods also have to deal with identification of meaningful units. The task therefore shifts from statistical analysis of word frequencies to a stronger focus on signal identification and differentiation from noise in speech recognition applications. Similarly, the task of mapping protein sequences to their structure, dynamics and function can also be seen more generally as a signal processing task. Just as the speech signal is a waveform whose acoustical features vary with time, a protein is a linear chain of chemico-physical features that vary with position in the sequence. However, while a speech sample can take unlimited continuous values, or digitized values within a given digital resolution, for proteins the value can be only one of the possible twenty, corresponding to the twenty types of amino acids (see above). Hence assigning a symbol or value to each

of the twenty amino acids is one alternative for digital representation. This is however, not a meaningful representation for signal processing. We will review below the approaches by which signal processing techniques become applicable to the identification of meaningful building blocks in protein sequences. We will demonstrate in detail using this example how scientific and technological advances in the specialized area of automatic speech recognition become relevant for the specialized area within computational biology of protein secondary structure prediction. Both areas separately have been extensively researched for several decades; the complete solution has not been accomplished; in both cases the underlying principles are understood, yet are difficult to model for decoding by a computer practically, "as the physics of simplicity and complexity meet" [26]. For a deeper understanding of protein structure and protein biochemistry, see [27] and [28]. Readers interested further in speech recognition may refer to [29, 30].

3.1   Digital Representation

Speech waveform is a superposition of signals of various different frequencies. By way of Nyquist criterion, the information in the signal can be completely captured by sampling the signal at a rate that is at least twice that of the largest frequency in the signal. Since most information in human speech is band-limited to about 8 kHz, sampling it at a rate of 16 kHz is sufficient. A typical digitized speech signal is a series of discrete-time samples of its amplitude. The amplitude of each sample is further coded into discrete levels to allow digital representation. To apply signal processing techniques to protein sequences, the protein must be represented by some numerical representation of its property at each position. To derive a meaningful representation of the protein signal, we must understand the chemical structures of the amino acids and their resulting physico-chemical properties (see above). The scales relating the 20 amino acids to each other based on these properties can be used to replace the amino acid symbols with numeric representations more similar to speech waveforms. In principle, any one of the property scales can be used, depending on the type of protein sequence analysis required. Consider the example speech utterance, "how to recognize speech with this new display", whose waveform is shown in Fig. 6A. The signal has been sampled at 16 kHz. Typically, the signal also contains background noise and therefore the pauses in between words are not entirely flat. The waveform shows how the amplitude of the sound varies as time progresses from the beginning of the utterance to the end. In contrast consider a protein. Figures 6B and 6C show how a protein may be represented as numerical signals. Figure 6B shows the protein in terms of charge and Fig. 6C shows the protein represented in terms of hydrophobicity of the amino acids. While speech is represented with respect to time, protein is represented in physical dimension from one end to the other end of the amino acid chain.

   The goal of speech recognition is to identify the words that are spoken. There are several hundred thousand words in a typical language. These words are formed by a combination of smaller units of sound called phones. Recognizing
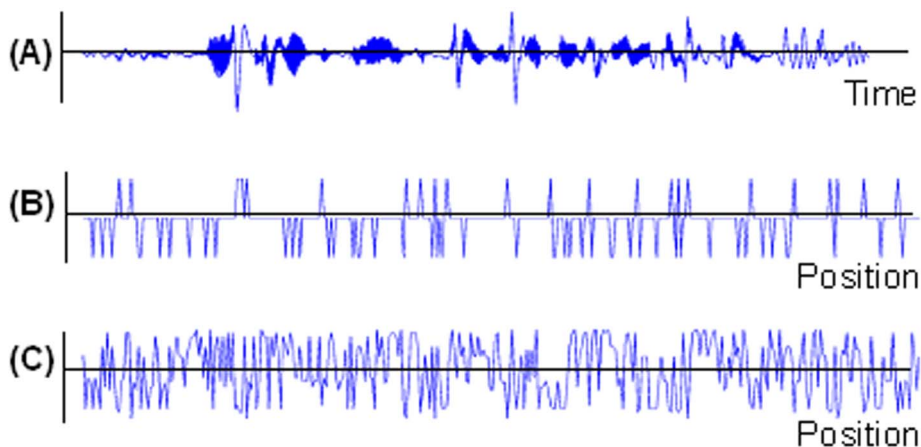
Fig. 6. Digital Representation. (A) Digital waveform of a speech signal of the utterance "how to recognize speech with this new display". The x-axis shows the time, while the y-axis shows the amplitude (loudness) of the signal. (B) Protein signal, where the sequence is represented by a numerical scale of charge of the residues. (C) Protein signal in terms of hydrophobicity of the residues. X-axis in (B) and (C) shows the residue number along the length of the protein, and y-axis is the value of the property of the residues (here charge or hydrophobicity).

a word in speech amounts to recognizing these phones. There are typically 50 phones in speech. Thus, identification of "word" equivalents in protein sequences using the signal processing approach is equivalent to "phone" identification.

3.2   Information Required to Decode the Signal

The content of a speech signal is not only dependent on the signal itself; its interpretation relies on an external entity, the listener. For example, consider the phrases:

How to recognize speech with this new display
How to wreck a nice beach with this nudist play

The two phrases are composed of almost identical phone sequences, but result in two different sentences. Spectrograms showing the frequency decomposition of the sound signals are shown in Fig. 7B and 7C, for these two sentences spoken by the same speaker.

Given the speech signal or a spectrogram, which utterance was meant by the speaker can be found by the context in which it was spoken. Thus the complete information for interpretation is not contained in the speech signal itself, but is inferred from the context. On the other hand, the linear strings of amino acids that make up a protein contain in principle all the information needed to fold it into a 3-D shape capable of fulfilling its designated function.
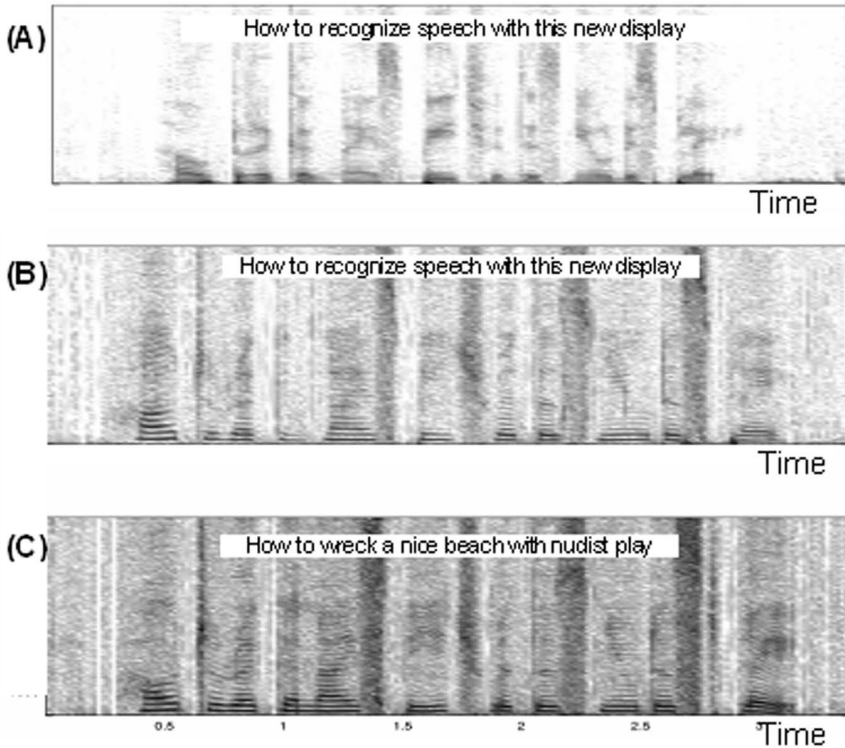
Fig. 7. Spectrograms of same utterances between different speakers and different utterances by same speaker: X-axis shows progression of time and y-axis shows different frequency bands. The energy of the signal in different bands is shown as intensity in grayscale values with progression of time. (A) and (B) show spectrograms of the same sentence "How to recognize speech with this new display" spoken by two different speakers, male and female. Although the frequency characterization is similar, the formant frequencies are much more clearly defined in the speech of female speaker. (C) shows the spectrogram of the utterance "How to wreck a nice beach with this nudist play" spoken by same speaker as in (B). (A) and (B) are not identical even though they are composed of the same words. (B) and (C) are similar to each other even though they are not the same sentences. See text for discussion.

3.3   Speaker Variability

Consider the signal characteristics of a word spoken by two different persons, especially if one is female and the other is male. Although the fundamental nature of the sounds remains the same, the overall absolute values of signal composition would be different. For example, a vowel sound would still have the same periodic nature in both utterances, but the frequency would be different. See for example, the frequency compositions of the same sentence spoken by a male and female speaker shown in Fig. 7A and 7B.

The analogy of speaker variability in the protein world can be found in the following broad categorization of proteins: the majority of a cell's proteins are
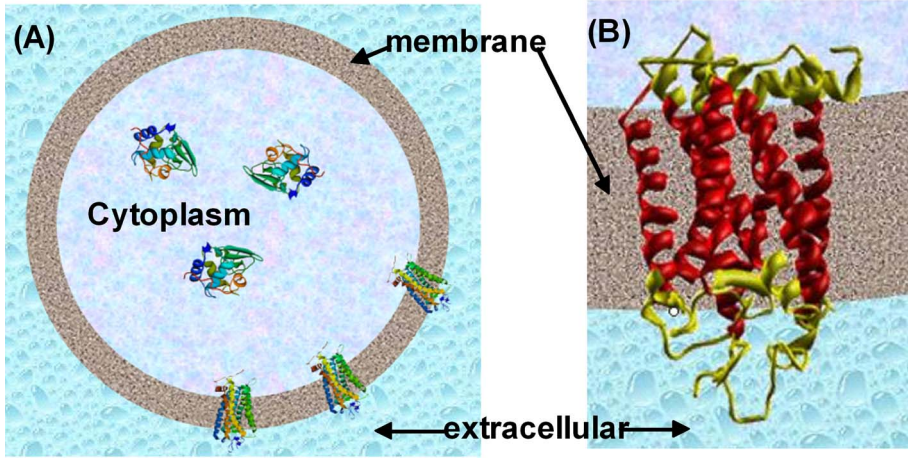
Fig. 8. Schematic of cell and soluble and transmembrane proteins. (A) A schematic of a cell: The cell is enveloped by a cell-membrane (brown) and is surrounded by water medium (blue bubbles). The medium inside the cell is made of water as well. Soluble proteins are found completely inside the cell. Membrane proteins are partly embedded in the cell-membrane. (B) Transmembrane protein Rhodopsin: It starts in the cytoplasmic region (top), traverses through the cell membrane (brown) to go into the extracellular region (bottom) and then transverses the membrane again to enter the cytoplasm. This protein has 8 helices in all, 7 of which are located mostly in the transmembrane region and extending out of it, and one helix (horizontal in picture) in the cytoplasmic region.

found inside the cells (soluble proteins), whereas, some proteins traverse through the cell membrane (membrane proteins). In contrast to soluble proteins, which are always in an aqueous environment, membrane proteins have parts that are like soluble proteins located either inside or outside of the cells, while a significant portion is located in a chemically different environment, the membrane lipid bilayer, as shown in Fig. 8. Since the environment around parts of these transmembrane proteins is different, the characteristics displayed by these parts are also different. Transmembrane helix prediction is closely related to protein secondary structure prediction given the primary sequence. Secondary structural elements described before, namely helix, strand, turn and loop are still the basic components of the three dimensional structure of membrane proteins also; however their characteristics are different from those of the soluble proteins when they are located in the membrane embedded parts. This difference may be seen as the speaker variability in speech. The intracellular, transmembrane and extracellular segments can be thought of as speech that is spoken by three different speakers.

Consider a domain specific speech recognition task where the number of speakers and the size of the vocabulary are very small. A method that is often adopted for this task is to recognize the words by speaker-specific word models. The approach adopted in transmembrane protein structure prediction is similar-

the structural elements are modeled specifically for each environment separately. Consider another domain specific task where the goal is to perform only speaker recognition out of three speakers (cytoplasmic, transmembrane and extracellular). The protein shown in Fig. 8B is an example of a transmembrane protein called rhodopsin. It consists of 8 helical segments and a beta sheet. Seven of the helices are transmembrane, one helix is soluble. A speaker-segmentation like task on this protein, would label these seven segments as transmembrane, and the rest of the protein as cytoplasmic and extracellular segments.

3.4   Signal Analysis of Transmembrane Proteins

The duration of the transmembrane segment is usually about 20-25 amino acid residues which corresponds to the 30 Å thickness of the cell membrane. (A residue is the equivalent of a sample in speech signal, whose value can be any one of the twenty amino acids). The cross-section of the cell-membrane is highly hydrophobic, thus imposing the requirement on amino acids within its environment to be predominantly hydrophobic. The properties most meaningful in this context to allow application of signal processing techniques are therefore related to hydrophobicity and polarity.

The most important mathematical tool in signal processing is the Fourier transform [31]. For a comprehensive review of signal processing methods in pro-
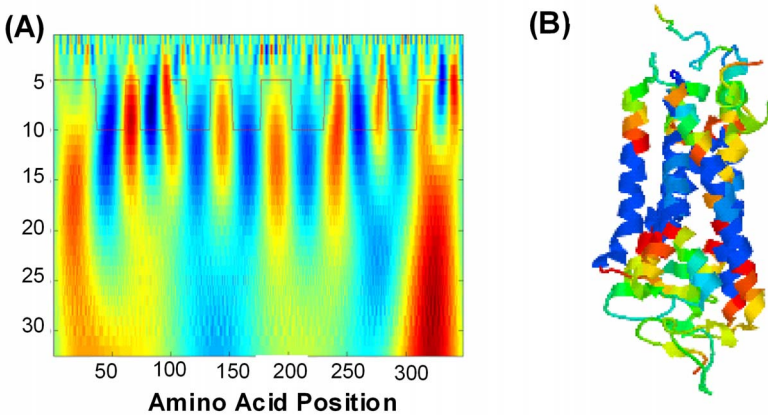

Fig. 9. Wavelet features of rhodopsin (swiss-prot id: OPSD_BOVIN) using a binary polar non-polar vocabulary: (A) Scalogram of the wavelet features: The primary sequence is mapped to polar nonpolar (1, 0 respectively) numerical scale and wavelet transform is computed at scales from 1 to 32 with the Mexican-hat analyzing function. The resulting 2D array is shown in image format after scaling the result to range between 0 and 1, with a rainbow color map VIBGYOR going from 0 to 1. The x-axis corresponds to the residue number and the y-axis corresponds to the scale at which the wavelet is computed, with the smallest scale at the top. (B) Wavelet features mapped onto the 3D structure of rhodopsin (pdb id: 1F88): The wavelet transform at a scale of 9 is normalized to a range of 0-150 and mapped onto the 3D structure of the protein, using 'temperature' field in the pdb format.

tein structure analysis, see [32]. A protein sequence is very short in length, being on an average 300 residues long. There are proteins as short as 50 residues and those that are larger than 1000 residues, but most of the proteins are a few hundred residues long. The duration of secondary structure elements are even shorter. Hence it is not suitable to use Fourier transform in the analysis of protein signal. Also, while Fourier transform can capture periodicities at any scale in the overall signal, it cannot identify the location of occurrence of periodicity. To capture local periodicities, Wavelet transform appears to be a more suitable mathematical tool [33] and has been applied earlier to speech recognition [34, 35]. Previously, the application of Wavelet transform in the context of transmembrane helix prediction has primarily been to de-noise the hydropathy signal by removing high frequency variations [36–39]. In the work presented here, wavelet transform is used to derive features from amino acid sequences.

In order to facilitate the use of signal analysis for the transmembrane helix prediction problem, polar/non-polar characteristics are mapped polar = 1, non-polar = 0. Other mappings such as by electronic properties, viz., mapping from strong electron donor to strong electron acceptors to numerical values +2 to -2, have also been studied. However, the best results were observed empirically by the choice of polar/non-polar representation. Application of wavelet transform to the polar/non-polar representation of one particular membrane protein, bovine rhodopsin (Swissprot ID: OPSD_BOVIN), is shown in Fig. 9. The numerical mapping of the sequence with polar/non-polar property is the same as shown in Fig. 6. A standard analysis function, Mexican-hat, at scales from 1 to 32 has been applied to this protein signal, resulting in a continuous wavelet transform of the protein sequence.

The wavelet transform gives rise to patterns that are distinct between the transmembrane regions from non transmembrane regions. An image representation of the wavelet transform, called the scalogram is shown in Fig. 9A. Superimposed on the scalogram is the location of transmembrane and non-transmembrane regions. Further, the wavelet transformed signal at different scales is also mapped onto the 3-dimensional structure of the protein, to visually analyze the distribution of feature values in different segments of the protein, here for scale 9 in rhodopsin (Fig. 9B).

3.5   Formal Analysis of the Features Derived Using Wavelet Methodology

Comparing the scalogram of a transmembrane protein in Fig. 9A to the spectrograms of speech in Fig. 7, it can be seen that the durational characteristic of transmembrane segments is very similar to that of phones in speech. The observations are very similar from one sample (or frame) to the next; there is an onset period and offset period from the transmembrane segment. In the absence of such durational feature, a classifier would have been suitable to classify the protein residues as transmembrane or non-transmembrane. However, to capture the time (or position) specific characteristics of the wavelets with respect to transmembrane domains, hidden Markov modeling (HMM) like architecture is
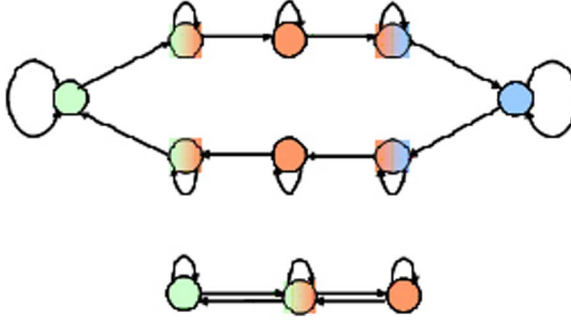
Fig. 10. HMM topology used for transmembrane prediction. The fully green state corresponds to the cytoplasmic loop, blue state to the extracellular loop, and the fully orange state to the core of the transmembrane region. The shaded states of green/orange and blue/orange colors correspond to transmembrane regions nearer to the lipid bilayer on cytoplasmic and extracellular sides. Although positive-inside rule [40] applies to the loop region thus characterizing cytoplasmic loops differently from extracellular loops, no distinction has been made in this work between cytoplasmic and extracellular loops. Hence, the topology shown on top reduces to that on the bottom, with just three states.

best suited. Here, we considered an HMM with a simple architecture as shown in Fig. 10. Each state is modeled with a mixture of 8 Gaussians. The vector of wavelet coefficients computed for scales 4 to 16 at each residue position in the protein, is considered the feature vector corresponding to that residue. In Fig 9A, the feature vectors correspond to columns in the 2D image of wavelet coefficients, considering only rows 4 to 16. The data set used is the set of 160 proteins [41]. The data set is available as 10 disjoint sets so that separate data may be used for training and testing. We used the first set (numbered 0) for testing, and the remaining sets for training. The accuracy of classification of each residue as transmembrane or non-transmembrane is found to be 80.0% (Q_2). Q_2 refers to the percentage of residues that have been classified correctly into the two states transmembrane and non-transmembrane. Although hidden Markov models have been used earlier towards transmembrane prediction, what is unique here is the demonstration of the use of wavelet coefficients as feature vectors. Within the speech recognition framework, wavelets have traditionally been used for speech enhancement (similar to hydrophobicity smoothing in case of transmembrane prediction), but a recent paper has demonstrated the use of wavelet coefficients as features for phoneme classification [35].

3.6   Membrane Helix Boundary Prediction Using N-Gram Features

The above work on transmembrane helix boundary prediction using signal processing techniques borrowed from language technologies strongly complements other applications of language technologies to the same task. As with phoneme identification, other language technologies applications use segmentation ap-

proaches, for example in document classification and topic detection. Traditionally, these methods have relied on n-gram features and statistical associations between them. We have complemented the above study with n-gram approaches to address the membrane protein boundary prediction problem.

(1) Similar to topic segmentation in natural language, we applied Yule's measure of association [42] to this problem based on its use in natural language processing [43]. Given a text with n different words, an n x n table of Yule values for every pair of words is computed. The distribution of Yule values in the table differs for different categories of text, indicating the positions of the boundaries. In a model application to the G-Protein Coupled Receptor (GPCR) family of membrane proteins, we found that Yule values can differentiate between transmembrane helices and loops connecting the helices [18].

(2) Using n-gram features but a different association measure, Mutual Information, it was also shown that language technologies can discover known functional building blocks, the transmembrane helices, without prior assumption on the length, type or properties of these building blocks. While the above Yule statistics required prior knowledge in the form of a training set for examples of transmembrane versus non-transmembrane applications, using mutual information, no such knowledge was required. Computing Mutual Information statistics on the entire dataset of a membrane protein family, the GPCR family, without prior knowledge on the positions of extracellular-transmembrane and cytoplasmic-transmembrane boundaries, can rediscover these boundaries, as shown in Fig. 11 [19]. In topic segmentation, topic boundaries are indicated by minima in Mutual Information. Similarly, in membrane proteins sequences, both membrane-cytoplasmic and membrane-extracellular boundaries are detected with high accuracy [19].
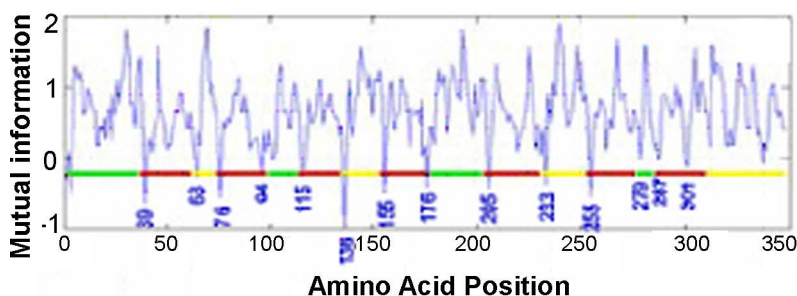

Fig. 11. Mutual information values along the rhodopsin sequence using different datasets GPCR to generate mutual information values [19]. Horizontal lines use the same color code as in Figure 1 indicating the positions of the segments belonging to each of extracellular, cytoplasmic and helices domains based on expert knowledge. The positions of breakpoints indicated by mutual information minima are shown as blue labels. The figure is JKS's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in [19] http://doi.acm.org/10.1145/967900.967933.

(3) Finally, an n-gram language modeling approach has also been adopted. The method builds a language model for each 'topic' representing transmembrane helices and loops and compares their performance in predicting the current amino acid, to determine whether a boundary occurs at the current position. The language models make use of only n-grams probabilities, but surprisingly still produced promising results [14, 18, 19].

4   Other Applications of Language Technologies in Computational Biology of Proteins

4.1   Genome Comparison

The secondary structure and transmembrane helix prediction tasks are only two examples of many tasks in computational biology where language technologies are relevant. For example, the features most often used in language technologies are word n-grams and there are many other tasks where n-grams do form meaningful building blocks. Probably the most widely known application of n-grams in computational biology is their use in the BLAST algorithm, where they enhance computational efficiency in sequence searching in the initial step [44]. However, n-grams have also proven useful in a number of other bioinformatics areas. The distributions of n-grams in biological sequences have been shown to follow Zipf's
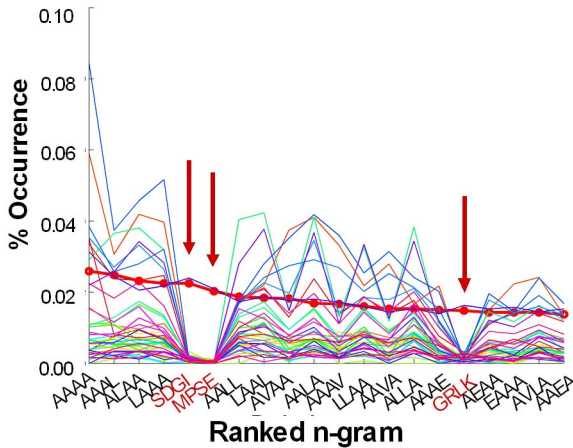

Fig. 12. Distribution of amino acid n-grams with n=4 in Neisseria meningitidis in comparison to the distribution of the corresponding amino acids in 44 other organisms [61]. N-grams of Neisseria are plotted in descending order of their frequency in the genome (in bold red). Numbers on x- indicate the ranks of the specific n-grams in Neisseria. Frequencies of corresponding n-grams from genomes of various other organisms are also shown (thin lines). The second thin line closely following the bold red line corresponds to a different strain of Neisseria meningtidis. Arrows indicate the positions of 4-grams that are over-represented in Neisseria, but are rare in other genomes. The figure is reproduced from [61] with permission from the publisher.

law [45–51]. Zipf's law states that the frequency of a word is related to its rank by a power law [52, 53]. While there is some debate as to the meaning of this observation for biological sequences [45–51], the Zipf plot of n-gram frequencies has found application in identification of genome signatures [16].

The Zipf-like analysis of protein sequences allows addressing the question of whether the sequences in proteins of different organisms are statistically similar or if organisms may be viewed as representations of different languages. We compared the n-gram frequencies of 44 different organisms using the n-gram comparison functions provided by the Biological Language Modeling Toolkit. (1) A simple Markovian uni-gram (context independent amino acid model from the proteins of Aeropyrum pernix was trained. When training and test sets were from the same organism, a perplexity (a variation of cross-entropy) of 16.6 was observed, whereas data from other organisms varied from 16.8 to 21.9. Thus, even the simplest model can automatically detect the differences in amino acid usage of different organisms. (2) We developed a modification of Zipf-like analysis that can reveal specific differences in n-grams in different organisms. First, the amino acid n-grams of a given length were sorted in descending order by frequency for the organism of choice. An example is shown in Fig. 12 for Neisseria meningitidis for n=4. Remarkably, there are three n-grams (shown by red arrows in the figure) that are among the top 20 most frequently occurring 4-grams in Neisseria, but that are rare or absent in any of the other genomes.

These highly idiosyncratic n-grams suggest "phrases" that are preferably used in the particular organism. These phrases are highly statistically significant, not only across organisms, but also within Neisseria itself. In particular, the 4-grams SDGI and MPSE are highly over-represented as compared to the frequencies expected based on the uni-gram distributions in Neisseria [16]. (3) While it is not known if these "phrases" correspond to similar or different substructures of proteins, we found that amino acid neighbor preferences are also different for different organisms, suggesting the possibility for underlying subtle changes in the mapping of sequences to structures of proteins.

4.2   Protein Family Classification

Another important task in computational biology is protein family classification. G-protein coupled receptors (GPCRs) are a superfamily of proteins and particularly difficult to classify into families due to the extreme diversity among its members. A comparison of BLAST, k-NN, HMM and SVM with alignment-based features has suggested that classifiers at the complexity of SVM are needed to attain high accuracy [54]. However, we were able to show that the simple Decision Tree and Naïve Bayes classifiers in conjunction with chi-square feature selection on counts of n-grams perform extremely well, and the Naïve Bayes classifier even outperforms the SVM significantly [55]. We also generalized the utility of n-grams for high-accuracy classification of other protein families using the Naïve Bayes approach [55, 56]. In line with these observations, Wu and co-workers have observed that neural networks perform well with n-gram features in the protein family classification task [57].

4.3   Prediction of Protein Folding Initiation Sites

The demonstrated success of language technologies for a number of typical computational biology tasks suggests that these methods may also prove useful in studies of tasks that have been studied less extensively. Prediction of folding initiation sites in proteins is a formidable task that requires novel approaches. We investigated if inverse frequencies may correlate with experimentally determined folding initiation sites in the protein folding model system, lysozyme. Our hypothesis was based on the observation that in natural languages, rare words carry the most relevant meaning of a text. Shown in Fig. 13 are inverse tri-gram frequencies plotted along the lysozyme sequence. Indeed, we observed a correlation between the locations of rare trigrams and the location of residual structure in the unfolded protein as evidenced by maxima in relaxation rates measured in NMR spectroscopic experiments. The statistical significance of this observation remains to be established by extension to other proteins, but lysozyme is the only protein for which the locations of folding initiation sites are known. However, the steady growth in the size of the protein databank will allow a systematic comparison between the sequences of n-grams and the number of structures that each n-gram can occur in. Such statistics are already beginning to be reported in the I-sites database [59] and in the analysis of sequences encoding certain types of structures [60]. Thus, it is expected that n-gram analysis may significantly contribute to the protein tertiary structure prediction problem in the future.
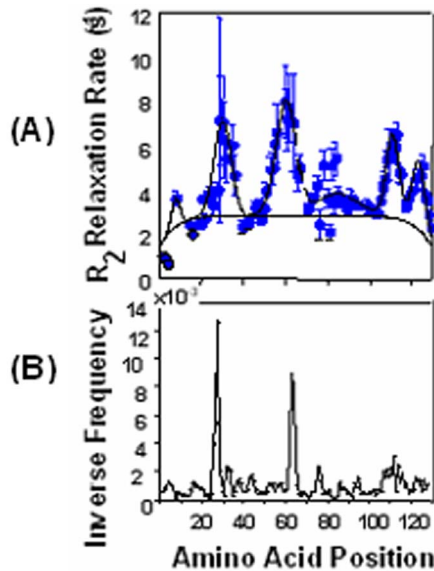

Fig. 13. Location of folding initiation sites in model protein Lysozyme (see Fig. 1) A. Transverse relaxation rates [58]. Large values above the black line indicate the presence of residual structure. B. Inverse trigram frequency in human lysozyme. The figure is reproduced from [61] with permission from the publisher.

5 Biological Language Modeling Toolkit and Website

A large number of linguistic methods for protein sequence analysis are provided at `http://flan.blm.cs.cmu.edu/`.

6 Conclusions

Here, we have shown that the use of an intuitive analogy allows direct application of methods developed in one specialized area of research to that of another. In particular, we demonstrated the use of language and speech technologies for a variety of computational biology problems. We described the major hurdle in the use of this analogy, the identification of functional equivalents of "words" in protein sequences with the long-term goal of preparing a dictionary for "protein sequence language". Although we are far from building such a dictionary, we demonstrate that a number of different vocabularies can provide meaningful building blocks in protein sequences. The utility of these vocabularies depends on the specific type of application in computational biology, and we provided examples from secondary structure prediction of soluble and of membrane proteins, of motif identification in genomes, protein family classification and protein folding and tertiary structure. Vocabularies range from individual chemical groups, to single amino acids, to combinations of amino acids (n-grams) with and without context information to chemical property representation. Automatic identification of functional building blocks using speech recognition and topic boundary detection methods both independently identified secondary structure elements as major functional building blocks of protein sequences.

Acknowledgements

References

1. Searls, DB: "The Language of Genes" Nature. volume 420. issue 6912. (2002) 211-7
2. Bolshoy, A: "DNA Sequence Analysis Linguistic Tools: Contrast Vocabularies, Compositional Spectra and Linguistic Complexity." Appl Bioinformatics. volume 2. issue 2. (2003) 103-12
3. Biological Language Modeling Project: http://www.cs.cmu.edu/~blmt/
4. Huang, CC and Couch, GS and Pettersen, EF and Ferrin, TE: "Chimera: An Extensible Molecular Modeling Application Constructed Using Standard Components" http://www.cgl.ucsf.edu/chimera. PSB1996: Pacific Symposium on Biocomputing. (1996) 50-61
5. Baldi, P: Bioinformatics. MIT Press. (1998)

6. Durbin, R and Eddy, S and Krogh, A and Mitchison, G: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press. (1998)

7. Bolshoy, A and Shapiro, K and Trifonov, E and Ioshikhes, I: "Enhancement of the Nucleosomal Pattern in Sequences of Lower Complexity." Nucl. Acids. Res. volume 25. issue 16. (1997) 3248-3254

8. Burge, C and Karlin, S (1997): Prediction of Complete Gene Structures in Human Genomic DNA. J Mol Biol 268(1), 78-94

9. Baxevanis, AD and Ouellette, BFF: Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins. Wiley-Interscience. (1998)

10. Bussemaker, HJ and Li, H and Siggia, ED: "Building a Dictionary for Genomes: Identification of Presumptive Regulatory Sites by Statistical Analysis." Proc Natl Acad Sci U S A. volume 97. issue 18. (2000) 10096-100

11. Gibas, C and Jambeck, P: Developing Bioinformatics Computer Skills. O'Reilly & Associates. (2001)

12. Troyanskaya, OG and Arbell, O, Koren and Y, Landau, GM and Bolshoy, A (2002): Sequence Complexity Profiles of Prokaryotic Genomic Sequences: A Fast Algorithm for Calculating Linguistic Complexity. Bioinformatics. May 18(5), pp 679-688

13. Coin, L and Bateman, A and Durbin, R: Enhanced Protein Domain Discovery by Using Language Modeling Techniques from Speech Recognition. Proc Natl Acad Sci USA. Apr 15; 100(8):4516-20. Epub 2003 Mar 31

14. Cheng, BYM and Carbonell, J and Klein-Seetharaman, J: Application of Topic Segmentation Techniques to Protein Sequences: Identification of Transmembrane Helix Boundaries in Gpcrs. (2004) In: Proceedings of the 8th International Conference on Spoken Language Processing, Jeju Island, Korea (Submitted)

15. Ganapathiraju, M and Klein-Seetharaman, J and Rosenfeld, R and Carbonell, J and Reddy, R: "Rare and Frequent Amino Acid N-Grams in Whole-Genome Protein Sequences." RECOMB'02: The Sixth Annual International Conference on Research in Computational Molecular Biology. Washington, USA. (2002)

16. Ganapathiraju, M and Weisser, D and Rosenfeld, R and Carbonell, J and Reddy, R and Klein-Seetharaman, J: "Comparative N-Gram Analysis of Whole-Genome Sequences" HLT2002: Human Language Technologies Conference. California, USA. (2002)

17. Ganapathiraju, M and Klein-Seetharaman, J and Balakrishnan, N and Reddy, R: "Characterization of Protein Secondary Structure Using Latent Semantic Analysis." IEEE Signal Processing magazine, May 2004 issue 15 (2004) 78-87

18. Ganapathiraju, M and Weisser, D and Klein-Seetharaman, J: "Yule Value Tables from Protein Datasets" SCI2004: World Conference on Systemics Cybernetics and Informatics. Florida, USA. (2004)

19. Weisser, D and Klein-Seetharaman, J: Identification of Fundamental Building Blocks in Protein Sequences Using Statistical Association Measures (2004)

20. PDBase: http://www.scsb.utmb.edu/comp_biol.html/venkat/prop.html In Silico Biol, volume 4, issue 2. (2004) 0012

21. ProtScale: http://www.expasy.org/tools/protscale.html In Silico Biol, volume 4, issue 2. (1992) 0012

22. Landauer, T and Foltx, P and Laham, D: "Introduction to Latent Semantic Analysis." Discourse Processes, volume 25, issue 5212. (1998) 259-284

23. Berry, MW and Browne, M: Understanding Search Engines: Mathematical Modeling and Text Retrieval. Soc for Industrial & Applied Math. (1999)

24. Rost, B: "Review: Protein Secondary Structure Prediction Continues to Rise" J Struct Biol, volume 134, issue 2-3. (2001) 204-18

25. Liu, Y and Carbonell, J and Klein-Seetharaman, J and Gopalakrishnan, V: "Comparison of Probabilistic Combination Methods for Protein Secondary Structure Prediction." Bioinformatics, volume 16, issue 4. (2004) 376-82

26. Frauenfelder, H and Wolynes, PG: "Proteins: Where the Physics of Simplicity and Complexity Meet." Physics Today, volume 47, issue 15. (1994) 58-64

27. Carl-Ivar Branden, JT: Introduction to Protein Structure. Garland Publishing. (1999)

28. Voet, D and Voet, JG: Biochemistry. J. Wiley & Sons. (1995)

29. Rabiner, L and Juang, B-H: Fundamentals of Speech Recognition. Pearson Education POD. (1993)

30. Deller, JR and Hansen, JHL and Proakis, JG: Discrete-Time Processing of Speech Signals. Wiley-IEEE press. (1999)

31. Proakis, JG and Manolakis, D: Digital Signal Processing: Principles, Algorithms and Applications. Macmillan USA. (1992)

32. Giuliani, A and Benigni, R and Zbilut, JP and Webber, CL Jr and Sirabella, P and Colosimo, A: "Nonlinear Signal Analysis Methods in the Elucidation of Protein Sequence-Structure Relationships." Chem Rev, volume 102, issue 5. (2002) 1471-92

33. Graps, A: "An Introduction to Wavelets." Computational Science and Engineering, IEEE [see also Computing in Science & Engineering], volume 2, issue 2. (1995) 50-61

34. Tan, BT and Fu, M and Spray, A and Dermody, P: "The Use of Wavelet Transforms in Phoneme Recognition." ICSLP96: Fourth International Conference on Spoken Language Processing. (1996) 148-55

35. Gupta, M and Gilbert, A: "Robust Speech Recognition Using Wavelet Coefficient Features." ASRU01: IEEE Workshop on Automatic Speech Recognition and Understanding. (2001) 50-61

36. Lio, P and Vannucci, M: "Wavelet Change-Point Prediction of Transmembrane Proteins." Bioinformatics, volume 16, issue 4, (2000) 376-82

37. Fischer, P and Baudoux, G and Wouters, J: "Wavpred: A Wavelet-Based Algorithm for the Prediction of Transmembrane Proteins." Comm. math. sci, volume 1, issue 1, (2003) 44 - 56

38. Pashou, EE and Litou, ZI and Liakopoulos, TD and Hamodrakas, SJ: "Wavetm: Wavelet-Based Transmembrane Segment Prediction" In Silico Biol, volume 4, issue 2. (2004) 0012

39. Qiu, J and Liang, R and Zou, X and Mo, J: "Prediction of Transmembrane Proteins Based on the Continuous Wavelet Transform." J Chem Inf Comput Sci, volume 44, issue 2. (2004) 741-7

40. von Heijne, G: "Membrane Protein Structure Prediction. Hydrophobicity Analysis and the Positive-inside Rule" J Mol Biol, volume 225, issue 2. (1992) 487-94

41. Sonnhammer, EL and von Heijne, G and Krogh, A: "A Hidden Markov Model for Predicting Transmembrane Helices in Protein Sequences." Proc Int Conf Intell Syst Mol Biol, volume 6, issue 6912. (1998) 175-82

42. Bishop, YMM and Fienberg, SE and Holland, PW: Discrete Multivariate Analysis. The MIT Press, Cambridge, Massachusetts and London, England (1975)

43. Cai, C and Rosenfeld, R and Wasserman, L: "Exponential Language Models, Logistic Regression, and Semantic Coherence." Proc. NIST/DARPA Speech Transcription Workshop. (2000) 10096-100

44. Altschul, SF and Gish, W and Miller, W and Myers, EW and Lipman, DJ: Basic Local Alignment Search Tool. (1990) J Mol Biol 215(3):403-10. Related Articles, Links

45. Mantegna, RN and Buldyrev, SV and Goldberger, AL and Havlin, S, Peng and CK, Simons, M and Stanley, HE: "Linguistic Features of Noncoding DNA Sequences." Phys Rev Lett, volume 73, issue 23. (1994) 3169-72

46. Konopka, AK and Martindale, C: "Noncoding DNA, Zipf's Law, and Language" Science, volume 268, issue 5212. (1995) 789

47. Chatzidimitriou-Dreismann, CA and Streffer, RM and Larhammar, D: "Lack of Biological Significance in the 'Linguistic Features' of Noncoding DNA – a Quantitative Analysis." Nucleic Acids Res, volume 24, issue 9. (1996) 1676-81

48. Israeloff, NE and Kagalenko, M and Chan, K: "Can Zipf Distinguish Language from Noise in Noncoding DNA?" Physical Review Letters, volume 76, issue 11. (1996) 1976

49. Strait, BJ and Dewey, TG: "The Shannon Information Entropy of Protein Sequences." Biophys J, volume 71, issue 1. (1996) 148-55

50. Tsonis, AA and Elsner, JB and Tsonis, PA: "Is DNA a Language?" J Theor Biol, volume 184, issue 1. (1997) 25-9

51. Li, W: "Statistical Properties of Open Reading Frames in Complete Genome Sequences." Comput Chem, volume 23, issue 3-4. (1999) 283-301

52. Zipf, GK: "Selective Studies and the Principle of Relative Frequency in Language." ICSLP96: Fourth International Conference on Spoken Language Processing. (1932) 3544-57

53. Miller, GA and Newman, EB: "Tests of a Statistical Explanation of the Rank-Frequency Relation for Words in Written English." American Journal of Psychology, volume 71, issue 23. (1958) 209-218

54. Karchin, R and Karplus, K and Haussler, D: Classifying G-Protein Coupled Receptors with Support Vector Machines. Bioinformatics 18(1):147-59 (2002)

55. Cheng, BYM and Carbonell, JG, and Klein-Seetharaman, J. (2004) Protein Classification Based on Text Document Classification Techniques. Proteins: Structure, Function and Bioinformatics (in press)

56. Vries, J and Munshi, R and Tobi, D and Klein-Seetharaman, J and Benos, PV and Bahar, I: A Sequence Alignment-Independent Method for Protein Classification. (2004) J Appl Bioinformatics (in press)

57. Wu, C and Whitson, G and McLarty, J and Ermongkonchai, A and Chang, TC: Protein Classification Artificial Neural System. (1992) Protein Science 1(5):667-677

58. Klein-Seetharaman, J and Oikawa, M and Grimshaw, SB and Wirmer, J and Duchardt, E and Ueda, T and Imoto, T and Smith, LJ and Dobson, CM and Schwalbe, H: "Long-Range Interactions within a Nonnative Protein." Science, volume 295, issue 5560. (2002) 1719-22

59. Simons, KT and Bonneau, R and Ruczinski, I and Baker, D: Ab Initio Protein Structure Prediction of Casp III Targets Using Rosetta. (1999) Proteins 1999; Suppl. 3:171-6

60. Kuznetsov, IB and Rackovsky, S: On the Properties and Sequence Context of Structurally Ambivalent Fragments in Proteins. (2003) Protein Science 12(11):2420-33

61. Ganapathiraju, M and Manoharan, V and Klein-Seetharaman, J: "BLMT: Statistical Sequence Analysis using N-Grams." J. Applied Bioinformatics, volume 3, issue 2. (2004)