We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 1

Motif Discovery in Protein Sequences

Salma Aouled El Haj Mohamed ,
Mourad Elloumi and Julie D. Thompson

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/65441

Abstract

Biology has become a data-intensive research field. Coping with the flood of data from the new genome sequencing technologies is a major area of research. The exponential increase in the size of the datasets produced by "next-generation sequencing" (NGS) poses unique computational challenges. In this context, motif discovery tools are widely used to identify important patterns in the sequences produced. Biological sequence motifs are defined as short, usually fixed length, sequence patterns that may represent important structural or functional features in nucleic acid and protein sequences such as transcription binding sites, splice junctions, active sites, or interaction interfaces. They can occur in an exact or approximate form within a family or a subfamily of sequences. Motif discovery is therefore an important field in bioinformatics, and numerous methods have been developed for the identification of motifs shared by a set of functionally related sequences. This chapter will review the existing motif discovery methods for protein sequences and their ability to discover biologically important features as well as their limitations for the discovery of new motifs. Finally, we will propose new horizons for motif discovery in order to address the short comings of the existent methods.

Keywords: motif discovery, bioinformatics, biological sequences, protein sequences, bioinspired algorithms

1. Introduction

Biology has been transformed by the availability of numerous complete genome sequences for a wide variety of organisms, ranging from bacteria and viruses to model plants and animals to humans. Genome sequencing and analysis is constantly evolving and plays an increasingly

important part of biological and biomedical research. This has led to new challenges related to the development of the most efficient and effective ways to analyze data and to use them to generate new insights into the function of biological systems. The completion of the genome sequences is just a first step toward the beginning of efforts to decipher the meaning of the genetic "instruction book." Whole-genome sequencing is commonly associated with sequencing human genomes, where the genetic data represent a treasure trove for discovering how genes contribute to our health and well-being. However, the scalable, flexible nature of next-generation sequencing (NGS) technology makes it equally useful for sequencing any species, such as agriculturally important livestock, plants, or disease-related microbes.

The exponential increase in the size of the datasets produced by this new generation of instruments clearly poses unique computational challenges. A single run of a NGS machine can produce terabytes of data, and even after image processing, base calling, and assembly, there will be hundreds of gigabytes of uncompressed primary data that must be stored either in flat files or in a database. Efficient treatment of all this data will require new computational approaches in terms of data storage and management, but also new effective algorithms to analyze the data and extract useful knowledge.

The major challenge today is to understand how the genetic information encoded in the genome sequence is translated into the complex processes involved in the organism and the effects of environmental factors on these processes. Bioinformatics plays a crucial role in the systematic interpretation of genome information, associated with data from other high-throughput experimental techniques, such as structural genomics, proteomics, or transcriptomics.

A widely used tool in all these stages is the comparison (or alignment) of the new genetic sequences with existing sequences. During genome assembly, short read sequences are often aligned to a reference genome to form longer contigs. Identification of coding regions then involves alignment of known genes to the new genomic sequence. Finally, functional significance is most often assigned to the protein coding regions by searching public databases for similar sequences and by transferring the pertinent information from the known to the unknown protein. A wide variety of computational algorithms have been applied to the sequence comparison problem in diverse domains, notably in natural language processing. Nevertheless, the analysis of biological sequences involves more than abstract string parsing, for behind the string of bases or amino acids is the whole complexity of molecular and evolutionary biology.

One major problem is the identification of important features, such as regulatory sites in the genomes, or functional domains or active sites in proteins, that are conserved within a family of sequences, without prior alignment of the sequences. In this context, motif recognition and discovery tools are widely used. The retrieved motifs are often compiled in databases including DNA regulatory motifs in TRANSFAC [1], JASPAR [2], or RegulonDB [3], and protein motifs in PRINTS [4], PROSITE [5], or ELM [6]. These well-characterized motifs can be used as a starting point for the identification of known motifs in other sequences. This is otherwise known as the pattern recognition problem. The challenges associated with *de novo* pattern

discovery, or the detection of previously unknown motifs [7], is far more difficult due to the nature of the motifs.

Biological sequence motifs are defined as short, usually fixed length, sequence patterns that may represent important structural or functional features in nucleic acid and protein sequences such as transcription binding sites, splice junctions, active sites, or interaction interfaces. They occur in an exact or approximate form within a family or a subfamily of sequences. Motif discovery is therefore an important challenge in bioinformatics and numerous methods have been developed for the identification of motifs shared by a set of functionally related sequences.

Consequently, much effort has been applied to *de novo* motif discovery, for example, in DNA sequences, with a large number of specialized methods that were reviewed recently in [8]. One interesting aspect is the development of nature-inspired algorithms, for example, particle swarm optimization has been used to find gapped motifs in DNA sequences [9], while DNA motifs have been discovered using an artificial immune system (AIS) [10]. Unfortunately, far fewer tools have been dedicated to the *de novo* search for protein motifs. This is due to the combinatorial explosion created by the large alphabet size of protein sequences, as well as the degeneracy of the motifs, i.e., the large number of wildcard symbols within the motifs. Some tools, such as Teiresias [11], or the MEME suite [12], can discover motifs in both DNA and protein sequences. Other work has been dedicated to the discovery of specific types of protein motifs, such as patterns containing large irregular gaps with "eukaryotic linear motifs" with SLiMFinder [13] or phosphorylation sites [14]. Many studies have been conducted to compare these specific motif discovery tools, such as [15].

In most cases, *de novo* motif discovery algorithms take as input a set of related sequences and search for patterns that are unlikely to occur by chance and that might represent a biologically important sequence pattern. Since protein motifs are usually short and can be highly variable, a challenging problem for motif discovery algorithms is to distinguish functional motifs from random patterns that are overrepresented. One solution to this challenge is to first construct a global multiple alignment of the sequences and then search for motifs in the aligned sequences. This reduces the search space to the aligned regions of the sequences, but also severely limits the possibilities of finding new motifs.

Furthermore, existing motif discovery methods are able to find motifs that are conserved within a complete family, but most of them are still unable to find motifs that are conserved only within a subfamily of the sequences. These subfamily-specific motifs, which we will call "rare" motifs, are often conserved within groups of proteins that perform the same function (specificity groups) and vary between groups with different functions/specificities. These sites generally determine protein specificity either by binding specific substrates/inhibitors or through interaction with other protein.

In Section 2, we will provide a brief description of protein sequences and the motifs that characterize them. Then, in Section 3, the main approaches used for discovery of motifs in protein sequences will be presented. Section 3 also deals with motif recognition in protein sequences. In Section 4, the main approaches used for the more difficult problem of *de novo*

motif discovery will be presented. Finally, in Section 5, we will propose new horizons for motif discovery in order to address the short comings of the existent methods.

2. Protein sequences, active sites, and motifs

Some basic concepts in protein biology are necessary for understanding the rest of this chapter. For many readers, this will be a familiar territory and in this case, they may want to skip this section and go directly to Section 3.

The genetic information encoded in the genome sequence of any organism contains the blueprint for its potential development and activity. However, the translation of this information into cellular or organism-level behavior depends on the gene products, especially proteins. Proteins perform a wide variety of cellular functions, ranging from catalysis of reactions, nutrient transport, and signal transmission to structural and mechanical roles. A protein is composed of a single chain of amino acids (of which there are 20 different kinds), represented by their single letter codes. This "primary structure" or sequence is none other than a string of characters that we can read from left to right, i.e., from NH_2 part (*N*-terminal) to the COOH part (*C*-terminal).

Every protein molecule has a characteristic three-dimensional (3D) shape or conformation, known as its native state. The process by which a protein sequence assumes its 3D structure is known as folding. Protein folding can be considered as a hierarchical process, in which the primary sequence defines secondary structure, which in turn defines the tertiary structure. Individual protein molecules can then interact with other proteins to form complex quaternary structures. The precise 3D structure of a protein molecule is generally required for proper biological function since a specific conformation is needed that the cell factors can recognize and interact with.

During evolution, random mutagenesis events take place, which change the genomic sequences that encode proteins. There are several different types of mutation that can occur. A single amino acid can be substituted for another one. Insertions and deletions also occur, involving a single amino acid up to several hundred amino acids. Some of these evolutionary changes will adversely affect the function of a protein, e.g., mutations of active sites in an enzyme, or mutations that disrupt the 3D structure of the protein. If this happens to a protein that takes part in a crucial process for the cell, it will result in cell death. As a result, amino acids that are essential for a protein's function, or that are needed for the protein to fold correctly, are conserved over time. Occasionally, mutations occur that give rise to new functions. This is one of the ways that new traits and eventually species may come about during evolution.

By comparing related sequences and looking for those amino acids that remain the same in all of the members in the family, we can predict the sites that might be essential for function. Some examples of important functional sites include the following:

- Enzyme active sites: to catalyze a reaction, an enzyme will bind to one or more reactant molecules, known as its substrates. The active site consists of the enzyme's amino acids that

form temporary bonds with the substrate, known as the binding site, and the amino acids that catalyze the reaction of that substrate.

- Ligand-binding sites: a binding site is a region on a protein molecule where ligands (small molecules or ions) can form a chemical bond. Ligand binding often plays a structural or functional role, for example, in stabilization, catalysis, modulation of enzymatic activity, or signal transmission.
- Cleavage sites: the location on a protein molecule where peptide bonds are broken down by hydrolysis. For instance, in human digestion, proteins in food are broken down into smaller peptide chains by digestive enzymes. Many viruses also produce their proteins initially as a single polypeptide chain which is then cleaved into individual protein chains.
- Posttranslational modification sites: some amino acids in a protein can undergo chemical modification, produced in most cases by an enzyme after its synthesis or during its life in the cell. This change usually results in a change of the protein function, whether in terms of its action, half-life, or its cellular localization.
- Targeting sites: within a cell, the localization of a protein is essential for its proper functioning, but the production site of a protein is often different from the place of action. Protein targeting signals, such as nuclear or mitochondrial localization signals, can be encoded within the polypeptide chain to allow a protein to be directed to the correct location for its function.

An example of a simple functional site is the N-glycosylation site, which is a posttranslational modification where a carbohydrate is attached to a hydroxyl or other functional group of a protein molecule. The sequence motif representing this site can be indicated by N-X-S/T. The first amino acid is asparagine (N), the second amino acid can be any of the 20 amino acids (X), and the third amino acid is either serine (S) or threonine (T). This example introduces the first complication in protein motif discovery: the motifs can contain both exact and ambiguous elements. Asparagine is a necessary amino acid, since this is the site that will be glycosylated, and is represented by an exact element. The third position should be a hydroxyl-containing amino acid (serine or threonine), while the second position is a "wild card." Nevertheless, the N-glycosylation motif shown here is uninterrupted, and so it is relatively easy to recognize. The spacing between the elements in many other sequence motifs can vary considerably, but the presence of such motifs is generally detected from the structure rather than sequence and this kind of motif will not be discussed in detail here. Finally, it should be pointed out that, just because this motif appears in a protein sequence, it does not mean that the site is glycosylated. The functional implications of a motif will depend on the neighboring amino acids and the surrounding 3D context. Therefore, in practice, identifying functional motifs from a protein sequence is far from straightforward.

3. Motif recognition in protein sequences

The motif recognition problem takes as input a set of known patterns or features that in some way define a class of proteins. The goal is then to search in an unsupervised or supervised way for other instances of the same patterns. As mentioned in the Introduction, the known motifs in biological sequences are generally compiled databases that are publically available over the Internet. For example, the PRINTS database (www.bioinf.manchester.ac.uk/dbbrowser/PRINTS) contains "protein fingerprints," where a fingerprint is composed of a group of motifs that characterize a given set of protein sequences with the same molecular function. In contrast, the PROSITE (prosite.expasy.org) and ELM (elm.eu.org) databases contain single motifs that correspond to known functionally or structurally important amino acids, such as those involved in an active site or a ligand binding site. The motifs contained in these resources are generally manually curated and the entries in the databases include extensive documentation of the specific biological function associated with the sites.

3.1. Motif representation

Over the years, a variety of motif representation models have been developed to take into account the complexity of protein motifs. The models are attempts to construct generalizations based on known functional motifs, and are used to help characterize the functional sites and to facilitate their identification in unknown protein sequences. They can be divided into two main categories.

3.1.1. Deterministic models

Consensus sequences are the simplest model for representing protein motifs. They can be constructed easily by selecting the amino acid found most frequently at each position in the signal. The number of matches between a consensus and an unknown candidate sequence can be used to evaluate the significance of a potential functional site. However, consensus sequences are limited models, since they do not capture the variability of each position. To support some degree of ambiguity, regular expressions can be used. Regular expressions are typically composed of exact symbols, ambiguous symbols, fixed gaps, and/or flexible gaps [16]. For example, the IQ motif is an extremely basic unit of about 23 amino acids, whose conserved core can be represented by the regular expression:

[FILV]Qxxx[RK]Gxxx[RK]xx[FILVWY]

where x signifies any amino acid, and the square brackets indicate an alternative.

3.1.2. Probabilistic models

Although deterministic models provide useful ways to construct human-readable representations of motifs, their main drawback is that they lose some information. For instance, in the IQ motif discussed above, the first position is usually I and both [RK] are most often R.

Probabilistic models can be used to overcome such loss of information. The position-specific scoring matrix (PSSM) [17], also known as the probability weight matrix (PWM), is undoubtedly one of the most widely used probabilistic models. This model is represented by a matrix where each entry (*i*,*a*) is the probability of finding an amino acid *a* at the *i*th position in the sequence motif. For example, for a set of motifs:

- WSEW
- WSRW
- CSKW
- CSKW
- YSKY

The corresponding PSSM is shown in **Table 1**.

Position	1	2	3	4
C	0.4	0.0	0.0	0.0
E	0.0	0.0	0.2	0.0
K	0.0	0.0	0.6	0.0
R	0.0	0.0	0.2	0.0
S	0.0	1.0	0.0	0.0
W	0.4	0.0	0.0	0.8
Y	0.2	0.0	0.0	0.2

Table 1. Example of a position specific scoring matrix (PSSM).

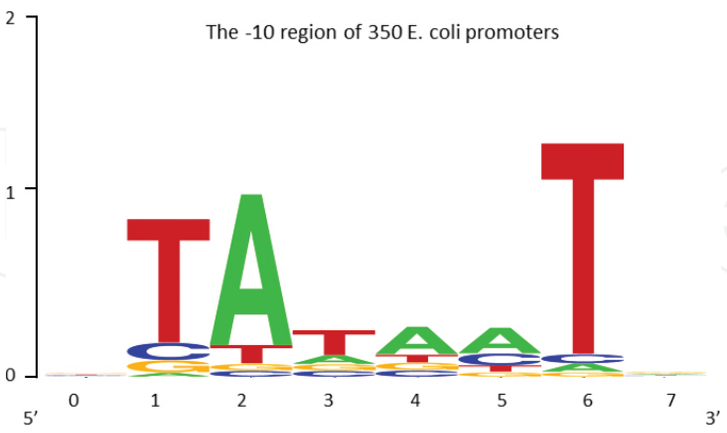

Figure 1. An example of a sequence logo for representing patterns in biological sequences. The logo represents the Pribnow box, a conserved region found upstream of the some genes in prokaryotic genomes.

Although in this example, PSSM containing entries having a value of 0, in general, pseudocounts are applied, especially when using a small dataset, in order to allow the calculation of probabilities for new motifs.

The information summarized in the PSSM can also be represented by a sequence logo [18], which is a graphical representation of the motif conservation as shown in **Figure 1**. A logo consists of a stack of letters at each position in the motif, where the relative sizes of the letters indicate their frequency in the sequences. The total height of the letters corresponds to the information content of the position, in bits.

Another widely used probabilistic model is the hidden Markov model (HMM), a statistical model that is generally applicable to time series or linear sequences. They were first introduced in bioinformatics for DNA sequences [19]. A HMM can be visualized as a finite state machine that moves through a series of states and produces some kind of output. The HMM generates a protein sequence by emitting amino acids as it progresses through a series of states. Each state has a table of amino acid emission probabilities, and transition probabilities for moving from state to state.

All of the representations mentioned so far inherently assume that positions within the motif are independent of each other. However, in some cases, this strong independence assumption may not be reasonable. Markov models of higher order, permuted Markov models, or Bayesian networks can be used to capture local dependencies by considering how each position depends on the other.

3.2. Motif detection

The models described in the previous section can be applied to the task of scanning a user-submitted sequence for matches to known motifs, thus providing evidence for the function of the protein and contributing to its classification in a given protein family. Ideally, a motif model would recognize all and only the members of the family. Unfortunately, this is seldom the case in practice.

In the case of deterministic models including consensus sequences and regular expressions, the models are often either too specific leading to a large number of false negative predictions, or too degenerate resulting in many false positives. The statistical power of such models can be estimated using standard measures, such as the positive and negative predictive values (PPV and NPV, respectively).

In the case of probability matrices or HMM-based methods, a log-odds score can be calculated that is a measure of how probable it is that a sequence is generated by a model rather than by a random null model, representing the universe of all sequences (also known as the "background"). The log-odds score of a motif is defined as:

$$score(s) = \log_z \frac{P_m(s)}{P_\varnothing(s)} \tag{1}$$

where P_m is the probability that the sequence was generated by the motif model m and P_ϕ is the probability that the sequence was generated by the null model. The logarithm is usually

base 2, and the score is given in bits. A log-odds score greater than zero indicates that the sequence fits the motif model better.

4. Motif discovery in protein sequences

4.1. Methods for motif discovery

Given a set of functionally related sequences, the main aim of motif discovery algorithms is to find new and *a priori* unknown motifs that are frequent, unexpected, or interesting according to some formal criteria. The methods used to discover such motifs follow the same general schema, as shown in **Figure 2**. They can be grouped into two main categories: alignment-based methods and methods that search for motifs in unaligned sequences.

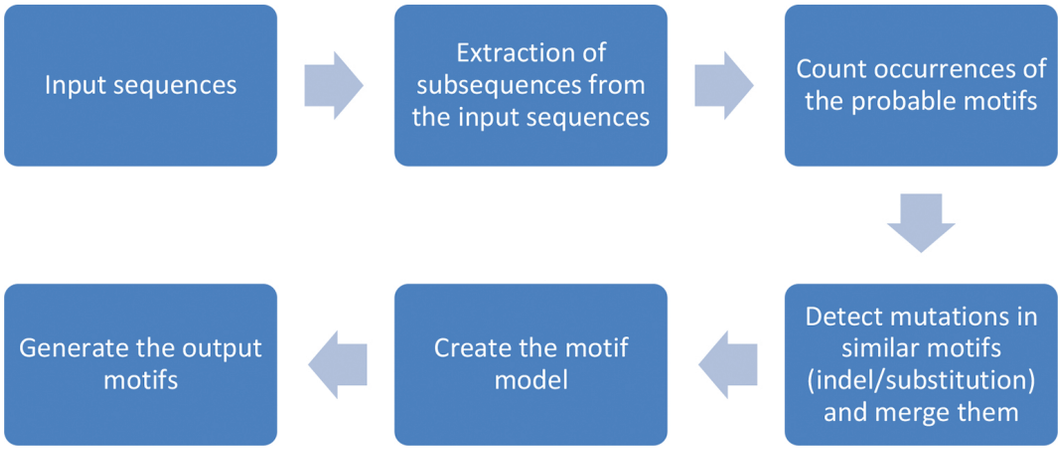

Figure 2. General motif discovery process.

4.1.1. Alignment-based methods

Alignment-based methods for motif discovery first construct a multiple sequence alignment of the set of sequences, where each sequence of amino acids is typically represented as a row within a matrix. Gaps are inserted between the amino acids so that identical or similar characters are aligned in successive columns. Once the multiple alignments are constructed, the patterns are extracted from the alignment by combining the substrings common to most of the sequences.

One of the first automatic methods for the identification of conserved positions in a multiple alignment was the AMAS program [20], using a set-based description of amino acid properties. Since then, a large number of different methods have been proposed. For example, Al2Co [21] calculates a conservation index at each position in a multiple sequence alignment using weighted amino acid frequencies at each position. The DIVAA method [22] is based on a statistical measure of the diversity at a given position. The diversity measures the proportion of the 20 possible amino acids that are observed.

The advantage of the alignment-based approach is that no upper limit has to be imposed on the length of the motifs. Moreover, these algorithms usually do not need as input a maximum threshold value for the motif distance from the sequences. In general, this approach works well if the sequences are sufficiently similar and the patterns occur in the same order in all of the sequences. Unfortunately, this is not usually the case and therefore most methods for motif discovery in protein sequences assume that the input sequences are unaligned.

4.1.2. Alignment-free methods

The vast majority of motif discovery methods in bioinformatics are alignment-free approaches that do not rely on the initial construction of a multiple sequence alignment. Instead, they generally search for patterns that are overrepresented in a given set of sequences. The simplest solution is to generate all possible motifs up to a maximum length l, and then to search separately for the approximate occurrences of each motif in the set of sequences. Once a list of candidate patterns is obtained, the ones with the highest significance scores are selected. This approach guarantees to find all motifs that satisfy the input constraints. Moreover, the sequences can be organized in suitable indexing structures, such as suffix trees, etc., so that the time needed by the algorithm to search for a single motif is linear in the overall length of the sequences.

This simplistic approach has an evident disadvantage: the number of candidate motifs, and therefore the time required by the algorithm, grows exponentially with the length of the sequences. Computing a significance score for each motif further increases the time required by the algorithm. A number of more efficient tools have been developed to address these issues and in the next chapter, we will discuss some of the more widely used ones.

4.2. Tools for motif discovery

In this section, we will present of the programs that are specifically designed to search for motifs in protein sequences that are biologically significant. The search for motifs in a set of unaligned sequences is a complex problem because many factors come into play, such as the precise start and end boundaries of the motif, the size variability (presence of gaps or not), or stronger or weaker motif conservation during evolution.

De novo motif discovery programs are generally based on one of the following three algorithms:

- Enumeration is a method that involves counting all substrings of a certain length (known as words or *k*-mers) and then seeking overrepresentations. Such exhaustive motif finding approaches are guaranteed to report all instances of motifs in a set of sequences. However, the exponential complexity of such searches means that the problem quickly becomes intractable for large alphabets.
- Deterministic optimization is based on the expectation-maximization (EM) algorithm that estimates the likelihood of a motif from existing data in two stages repeated iteratively. The first uses a set of parameters to reconstruct the hidden motif structure. The second uses this structure to reestimate the parameters. This method allows finding alternate sequences representing the motif and updating the motif model.

- Probabilistic optimization is an iterative method in which a random subsequence is extracted from each sequence to build an initial model. In each subsequent iteration, the *i*th sequence is removed and the model is recalculated. Then, a new motif is extracted from the *i*th sequence. This process is repeated until convergence.

Below, and in **Table 2**, we present the most used motif discovery programs and discuss their advantages and limitations.

Teiresias [11] is based on an enumeration algorithm. It operates in two phases: scanning and convolution. During the scanning phase, elementary motifs with sufficient support are identified. These elementary motifs constitute the building blocks for the convolution phase. They are combined into progressively larger motifs until all the existing maximal motifs are generated.

MEME [12] is an example of a deterministic optimization algorithm. It allows discovery of motifs in DNA or protein sequences based on expectation maximization (EM). MEME discovers at least three motifs, each of which may be present in some or all of the input sequences. MEME chooses the width and number of occurrences of each motif automatically in order to minimize the "E-value" of the motif, i.e., the probability of finding a similarly well-conserved pattern in random sequences. With default parameters, only motif widths between 6 and 50 are considered, but the user have the possibility to change this as well as several other parameters (options) of the motif discovery.

Pratt [23] is based on probabilistic optimization. It first searches the space of motifs, as constrained by the user, and compiles a list of the most significant sequences that matches at least the user-defined minimum number of sequences. If the user has not switched off the refinement, these motifs will be input to one of the motif refinement algorithms. The most significant motifs resulting from this are then output to a file.

qPMS [24] stands for quorum planted motif search. The program searches for motifs in either DNA or protein sequences. It uses the (l, d) motif search algorithm known as the planted motif search. qPMS takes as input a set of sequences and two values, l and d. It returns all sequences M of length l, which appear in at least q% of the sequences.

SLiMFinder [13] identifies novel short linear motifs (SLiMs) in a set of sequences. SLiMs are microdomains that have important functions in many diverse biological pathways. SLiM-mediated functions include posttranslational modification, subcellular localization, and ligand binding. SLiMs are generally less than 10 amino acids long, many of which will be "flexible" in terms of the conserved amino acid. SLiMFinder constructs such motifs by grouping dimers into longer patterns: motifs with fixed amino acid positions are identified and then grouped to include amino acid ambiguity and variable-length wildcards. Finally, motifs that are overrepresented in a set of unrelated proteins are identified.

Dilimot [25] proceeds as follows: in the first step, a user provided set of protein sequences is filtered to eliminate repetitive sequences as well as the regions least likely to contain linear motifs. In the second step, overrepresented motifs are identified in the nonfiltered sequences and ranked according to scores that take into account the background probability of the motif,

the number of sequences containing the motif, the size of the sequence set, and the degree to which the motif is conserved in other orthologous proteins.

Program	Description	Advantages	Disadvantages
Teiresias	Finds motifs that are frequent in a set of related sequences	Does not need background sequences; Very fast	Too many redundant motifs discovered
MEME	Finds motifs in related sequences using Gibbs sampling and expectation maximization	Does not need background sequences; Fast, Multi-thread version available; User friendly output	User defines the number of motifs to discover
Pratt	Discovers flexible motifs in related sequences	Does not need background sequences	Unable to discover effectively exact motifs
qPMS	Finds overrepresented motifs in a set of sequences based on Quorum Planted Motif Search	Fast; Low memory consumption	Limited to 20 protein sequences
SlimFinder	Finds overrepresented motifs in a set of unrelated sequences relative to background sequences	Well documented; Can use filters	Needs background sequences
MotifHound	Exhaustively finds motifs overrepresented in a set of unrelated sequences relative to background sequences	Exhaustive exploration of motifs; Can use filters Fast; Multi-thread version available	Needs background sequences
Dilimot	Finds overrepresented motifs in a set of unrelated sequences relative to a background sequences	Integrates several types of sequence information on motifs	Needs background sequences; Source code not available
FirePro	Correlates overrepresented motifs in a set of sequences with specific functions or behaviors	User friendly output	Needs background sequences

Table 2. Advantages and limitations of the most used motif discovery programs.

MotifHound [26] is suitable for the discovery of small and degenerate linear motifs. The method needs two input datasets: a background set of protein sequences and a subset of this background set that represents the query sequences. MotifHound first enumerates all possible motifs present in the query sequences, and then calculates the frequency of each motif in both the query and the background sets.

FIRE-pro [27] stands for finding informative regulatory elements in proteins. Its main goal is to discover protein motifs that correlate with the biological behavior of the corresponding proteins. FIRE-pro calculates a mutual information measure between frequent *k*-mer motifs and a "protein behavior profile" containing experimental data about the function of the proteins.

Most of these programs need prior knowledge about either the input sequences or the motif structure. Furthermore, they are generally designed to discover frequent motifs that occur in all or most of the sequences. The subfamily-specific motifs, which differentiate a specific subset of sequences, pose a greater challenge due to the statistical nature of these programs or the default choice of parameters used. Nevertheless, these "rare" motifs are often characteristic of important biological functions or context-specific modifications, including substrate binding sites, protein-protein interactions, or posttranslational modification sites.

In the final section of this chapter, we will discuss the use of "intelligent algorithms" that should be more reliable for the discovery of significant rare motifs in addition to the conserved and known ones.

5. Intelligent algorithms for protein motif discovery

Intelligent algorithms include optimization and nature inspired algorithms. Among these, artificial immune systems are especially adapted to pattern discovery, and have been used recently for motif discovery in DNA sequences. The high complexity and dimensionality of the problems in bioinformatics are an interesting challenge for testing and validating new computational intelligence techniques. Similarly, the application of AIS to bioinformatics may bring important contributions to the biological sciences, providing an alternative form of analyzing and interpreting the huge volume of data from molecular biology and genomics [28].

Artificial immune systems are a class of computationally intelligent systems inspired by the principles and processes of the vertebrate immune system. The algorithms typically apply the structure and function of the immune system to solving hard computational problems. Since their introduction in the 1990s, a number of common techniques have been developed, including:

- Clonal selection algorithms model how antibodies of the immune system adaptively learn the features of the intruding antigen and defend the body from it. The algorithms are most commonly applied to optimization and pattern recognition domains.
- Negative selection refers to the identification and deletion of self-reacting cells, i.e., cells that may attack self-tissues. The algorithms are typically used for classification and pattern recognition problems, especially in the anomaly detection domain.
- Immune network algorithms focus on the network graph structures involved where antibodies represent the nodes and the training algorithm involves growing or pruning edges between the nodes based on affinity. The algorithms have been used to solve clustering, data visualization, control, and optimization problems.
- Dendritic cell algorithms are inspired by the danger theory algorithm of the mammalian immune system, and particularly the role and function of dendritic cells, from the molecular networks present within the cell to the behavior exhibited by a population of cells as a whole.

Although a number of these different AIS can be used for pattern recognition, the clonal selection algorithm seems to be particularly well suited for protein motif discovery in large sets of sequences. In particular, the capabilities for self-organization of huge numbers of immune cells mean that no prior information is needed. In addition, the system does not require outside intervention and so it can automatically classify pathogens (motifs) and it can react to pathogens that the body has never seen before. Another advantage of AIS is the fact that there are varying types of elements that protect the body from invaders, and there are different lines of defense, such as innate and adaptive immunity. These features can be abstracted to model the diverse types of motifs found in protein molecules (see Section 1). These different mechanisms are organized in multiple layers that act cooperatively to provide high noise tolerance and high overall security.

The use of such intelligent algorithmic approaches should improve the whole motif discovery process: from the selection of suitable sets of sequences, via data cleaning and preprocessing, motif identification and evaluation, to the final presentation and visualization of the results. Nevertheless, a number of issues remain to be addressed before such systems can be applied to the very large datasets produced by NGS technologies. In particular, the substantial time and memory requirements of AIS are a limiting factor, although these can be significantly reduced thanks to the inherently parallel nature of the algorithms.

Acknowledgements

We would like to thank the members of the BICS and BISTRO Bioinformatics Platforms in Strasbourg for their support. This work was supported by Institute funds from the CNRS, the Université de Strasbourg and the Faculté de Médecine de Strasbourg.

Author details

Salma Aouled El Haj Mohamed[1,2,3], Mourad Elloumi[2] and Julie D. Thompson[3*]

*Address all correspondence to: thompson@unistra.fr

1 Faculty of Science, Doctoral School of Mathematics, Computer Science and Material Science and Technology, University of Tunis El Manar, Tunis, Tunisia

2 Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE), University of Tunis, Tunis, Tunisia

3 Department of Computer Science, ICube, UMR 7357, University of Strasbourg, CNRS, Strasbourg Federation of Translational Medicine (FMTS), Strasbourg, France

References

[1] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006 34:D108–D110.

[2] Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, Zhang AW, Parcy F, Lenhard B, Sandelin A, Wasserman WW. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016 44:D110-–D115.

[3] Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, Alquicira-Hernández K, Martínez-Flores I, Pannier L, Castro-Mondragón JA, Medina-Rivera A, Solano-Lira H, Bonavides-Martínez C, Pérez-Rueda E, Alquicira-Hernández S, Porrón-Sotelo L, López-Fuentes A, Hernández-Koutoucheva A, Moral-Chávez VD, Rinaldi F, Collado-Vides J. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2016 44:D133-143.

[4] Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Romá-Mateo C, Theodosiou A, Mitchell AL. The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. Database J Biol Databases Curation. 2012 bas019.

[5] Sigrist CJ, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. New and continuing developments at PROSITE. Nucleic Acids Res. 2013 41:D344–D347.

[6] Dinkel H, Van Roey K, Michael S, Kumar M, Uyar B, Altenberg B, Milchevskaya V, Schneider M, Kühn H, Behrendt A, Dahl SL, Damerell V, Diebel S, Kalman S, Klein S, Knudsen AC, Mäder C, Merrill S, Staudt A, Thiel V, Welti L, Davey NE, Diella F, Gibson TJ. ELM 2016-data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res. 2016 44:D294–D300.

[7] Tompa P, Davey NE, Gibson TJ, Babu MM. A million peptide motifs for the molecular biologist. Mol Cell. 2014 55:161–169.

[8] Boeva V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Front Genet. 2016 7:24.

[9] Lei C, Ruan J. A particle swarm optimization-based algorithm for finding gapped motifs. BioData Mining. 2010 3:9–10.

[10] Seeja KR. AISMOTIF- An artificial immune system for DNA motif discovery. Int J Comput Sci. Issues. 2011 8:143–149.

[11] Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics. 1998 14:55–67.

[12] Bailey TL, Bodén M, Whitington T, Machanick P. The value of position-specific priors in motif discovery using MEME. BMC Bioinformat. 2010 11:179.

[13] Edwards RJ, Davey NE, Shields D. SLiMFinder: A probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PLoS One. 2007 2:e967.

[14] Frades I, Resjö S, Andreasson E. Comparison of phosphorylation patterns across eukaryotes by discriminative N-gram analysis. BMC Bioinformat. 2015 16:239.

[15] Bhowmick P, Guharoy M, Tompa P. Bioinformatics approaches for predicting disordered protein motifs. Adv Exp Med Biol. 2015; 870:291–318.

[16] Brazma A, Jonassen I, Eidhammer I, Gilbert D. Approaches to the automatic discovery of patterns in biosequences. J Comp Biol. 1998 5:279–305.

[17] Henikoff S, Henikoff JG. Position-based sequence weights. J Mol Biol. 1994 243:574–578.

[18] Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990; 18:6097–6100.

[19] Churchill GA. Stochastic models for heterogeneous DNA sequences. Bull Math Biol. 1998; 51:79–94.

[20] Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci. 1993; 9:745–756.

[21] Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics. 2001; 17:700–712.

[22] Rodi DJ, Mandava S, Makowski L. DIVAA: analysis of amino acid diversity in multiple aligned protein sequences. Bioinformatics. 2004; 20:3481–3489.

[23] Jonassen I, Collins JF, Higgins DG. Finding flexible patterns in unaligned protein sequences. Protein Sci. 1995; 4:1587–1595.

[24] Dinh H, Rajasekaran S, Davila J. qPMS7: a fast algorithm for finding (l,d) motifs in DNA and protein sequences. PLoS One. 2012; 7:e41425.

[25] Neduva V, Russell RB. DILIMOT: discovery of linear motifs in proteins. Nucleic Acids Res. 2006; 34:W350–W355.

[26] Lieber DS, Elemento O, Tavazoie S. Large-scale discovery and characterization of protein regulatory motifs in eukaryotes. PLoS One. 2010; 5:e14444.

[27] Kelil A, Dubreuil B, Levy ED, Michnick SW. Fast and accurate discovery of degenerate linear motifs in protein sequences. PLoS One. 2014; 9:e106081.

[28] Al-Enezi A, Abbod MF, Alsharhan Al-Enezi S. Artificial immune systems-models, algorithms and applications. IJRRAS. 2010; 3:118–131.