SYSTEM AND METHOD FOR IDENTIFYING COMPLEX PATTERNS OF AMINO ACIDS

Related Applications

[0001] This application claims priority under 35 U.S.C. § 119(e) from U.S. Provisional Patent Application Ser. No. 60/565,847, filed April 28, 2004 and entitled "SYSTEM AND METHOD FOR IDENTIFYING COMPLEX PATTERNS OF AMINO ACIDS." This application also claims priority under 35 U.S.C. § 119(e) from U.S. Provisional Patent Application Ser. No. 60/653,083, filed February 16, 2005 and entitled "SYSTEM AND METHOD FOR IDENTIFYING COMPLEX PATTERNS OF AMINO ACIDS." Both of these provisional applications are incorporated herein by reference in their entireties and for all purposes.

[0002] Additionally, this application claims priority from and is a Continuation In Part of U.S. Non-provisional Patent Application Ser. No. 10/189,437, entitled "REPLIKIN PEPTIDES AND USES THEREOF," filed July 8, 2002, which is a Continuation In Part of U.S. Non-provisional Patent Application Ser. No. 10/105,232, entitled "REPLIKIN PEPTIDES IN RAPID REPLICATION OF GLIOMA CELLS AND IN INFLUENZA EPIDEMICS," filed March 26, 2002, which is a Continuation In Part of U.S. Non-provisional Patent Application Ser. No. 09/984,057, entitled "REPLIKINS AND METHODS OF IDENTIFYING REPLIKIN-CONTAINING SEQUENCES," filed October 26, 2001. Further, this application claims priority from and is a Continuation In Part of U.S. Non-provisional Patent Application Ser. No. 10/860,050, entitled "REPLIKIN PEPTIDES AND USES THEREOF," filed June 4, 2004. All of these non-provisional applications are incorporated herein by reference in their entireties and for all purposes.

Technical Field

[0003] This invention relates generally to the field of bioinformatics. More particularly, the invention relates to techniques for facilitating the identification of complex patterns of nucleotide or amino acid sequences.

Background of the Invention

[0004] As is well-known, amino acids are the building blocks of proteins. Proteins make up the bulk of cellular structures, and some proteins serve as

enzymes for facilitating cellular reactions. Twenty different amino acids are known to occur in proteins. The properties of each protein are dictated in part by the precise sequence of component amino acids.

[0005] Databases of amino acids and proteins are maintained by a variety of research organizations, including, for example, the National Center for Biotechnology Information (NCBI) at the U.S. National Library of Medicine, and the Influenza Sequence Database at the Los Alamos National Laboratory. These databases are typically accessible via the Internet through web pages that provide a researcher with capabilities to search for and retrieve specific proteins. These databases may also be accessible to researchers via local-area and wide-area networks. Additionally, researchers may directly access amino acid and protein databases stored on peripheral devices, such as magnetic disks, optical disks, static memory devices, and a variety of other digital storage media known in the art.

[0006] In amino acid and protein databases, amino acids are typically encoded as alphabetic characters. FIG. 1 lists each amino acid known to occur in proteins and provides a typical 3-letter abbreviation and single-letter code by which the amino acids may be represented in databases, according to a standard supplied by the International Union of Pure and Applied Chemistry (IUPAC).

[0007] A given protein may be described by its sequence of amino acids. For example, using the single-letter code given in FIG. 1, the character string "crvpsgvdla" corresponds to the protein defined by the following sequence of amino acids: cysteine, arginine, valine, proline, serine, glycine, valine, aspartic acid, leucine, and alanine.

[0008] When a protein database is searched for proteins that satisfy certain criteria (for example, those proteins relating to cancer in humans), the protein database search engine may respond by identifying hundreds or thousands of matching proteins. This set of matching proteins may be narrowed by supplying additional search criteria. At any point during the search process, specific proteins may be selected and reviewed. In FIG. 2, a printout describes a specific protein identified from an NCBI search for proteins relating to human cancer.

[0009] As can be seen in FIG. 2, a protein description may include detailed information describing, among other identifying factors, such information as the title of the protein ("Differential expression of a novel serine protease homologue in squamous cell carcinoma of the head and neck"), the authors of the protein

description ("Lang,J.C. and Schuller,D.E."), and the organism from which the protein was isolated ("Homo sapiens").

[0010] Protein descriptions may include a specific sequence of amino acids that define the protein. For example, in FIG. 2, amino acid sequence data is found at the end of the description of the protein, in a section of the printout prefaced by the word "ORIGIN." In this example, the first few amino acids are "myrpdvvrar," (SEQ ID NO: 1) which correspond to methionine, tyrosine, arginine, proline, aspartic acid, valine, valine, arginine, alanine, and arginine.

[0011] Some protein descriptions may include a sequence of nucleic acid bases, rather than amino acid sequences, that define the protein. As is known, a sequence of three nucleic acid bases (*i.e.*, a nucleic acid base triplet) may correspond to an amino acid according to a mapping provided by the table found in FIG. 3. Each nucleic acid base triplet identified in the table represents or corresponds to a specific amino acid. For example, the nucleic acid triplet GCT (guanine-cytosine-thymine) corresponds to the amino acid Alanine. Similarly, the nucleic acid triplet GCA (guanine-cytosine-adenine) also corresponds to the amino acid Alanine. As another example, the nucleic acid triplets AAA and AAG (adenine-adenine-adenine and adenine-adenine-guanine, respectively) each corresponds to the amino acid Lysine.

The Replikin Pattern

[0012] In previous patent applications, the inventors have identified and described a pattern of amino acids that has been designated a "Replikin pattern" or simply a "Replikin." A Replikin pattern comprises a sequence of about 7 to about 50 contiguous amino acids that includes the following three (3) characteristics:

[0013] (1) the sequence has at least one lysine residue located six to ten amino acid residues from a second lysine residue;

[0014] (2) the sequence has at least one histidine residue; and

[0015] (3) the sequence has at least 6% lysine residues.

[0016] Replikins have been shown to be associated with rapid replication in fungi, yeast, viruses, bacteria, algae, and cancer cells. Based on this association, it is believed that Replikins may be an indicator of disease. Additionally, an increase in concentration of Replikins over time may be an indicator of the imminent onset of disease. For example, before each of the three influenza pandemics of the last

century (identified as H1N1, H2N2 and H3N2), there was a significant increase in the concentration of Replikins in the corresponding influenza virus. With respect to the H5N1 influenza, FIG. 4 illustrates a rapid increase in the concentration of Replikins per 100 amino acids just prior to epidemics in 1997 (indicated as E1), 2001 (indicated as E2) and 2004 (indicated as E3). Replikin patterns have been found in a variety of disease-related proteins, including cancers of the lung, brain, liver, soft-tissue, salivary gland, nasopharynx, esophagus, stomach, colon, rectum, gallbladder, breast, prostate, uterus, cervix, bladder, eye, forms of melanoma, lymphoma, leukemia, and kidney. Importantly, Replikin patterns appear to be absent from the normal healthy human genome. FIG. 5 lists selected examples of Replikin patterns that have been found in various organisms.

[0017] For example, the 13-residue pattern "**h**yp**pk**pgcivpa**k**," (SEQ ID NO: 18) occurring in Hepatitis C (which is the last entry in the Tumor Virus Category of FIG. 5) is a Replikin pattern because: (1) it contains two lysine residues that are 8 positions apart; (2) it contains a histidine residue; and (3) the percentage of lysine residues is 2/13, which is 15.4%.

Amino Acid Search Tools

[0018] As is known in the art, databases of proteins and amino acids may be searched using a variety of database tools and search engines. Using these publicly available tools, patterns of amino acids may be described and located in many different proteins corresponding to many different organisms. Several methods and techniques are available by which patterns of amino acids may be described. One popular format is the PROSITE pattern. A PROSITE pattern description may be assembled according to the following rules:

[0019] (1) The standard International Union of Pure and Applied Chemistry (IUPAC) one-letter codes for the amino acids are used (see FIG. 1).

[0020] (2) The symbol 'x' is used for a position where any amino acid is accepted.

[0021] (3) Ambiguities are indicated by listing the acceptable amino acids for a given position, between square parentheses '[ ]'. For example: [ALT] would stand for Alanine or Leucine or Threonine.

[0022] (4) Ambiguities are also indicated by listing between a pair of curly brackets '{ }' the amino acids that are not accepted at a given position. For example: {AM} stands for any amino acid except Alanine and Methionine.

[0023] (5) Each element in a pattern is separated from its neighbor by a '-'.

[0024] (6) Repetition of an element of the pattern can be indicated by following that element with a numerical value or a numerical range between parenthesis. Examples: x(3) corresponds to x-x-x, x(2,4) corresponds to x-x or x-x-x or x-x-x-x.

[0025] (7) When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern either starts with a '<' symbol or respectively ends with a '>' symbol.

[0026] (8) A period ends the pattern.

[0027] Examples of PROSITE patterns include:

[0028] PA [AC]-x-V-x(4)-{ED}. This pattern is translated as: [Alanine or Cysteine]-any- Valine -any-any-any-any-{any but Glutamic Acid or Aspartic Acid}

[0029] PA <A-x-[ST](2)-x(0,1)-V. This pattern, which must be in the N-terminal of the sequence ('<'), is translated as: Alanine -any-[Serine or Threonine]-[Serine or Threonine]-(any or none)-Valine.

[0030] Another popular format for describing amino acid sequence patterns is the regular expression format that is familiar to computer scientists. In computer science, regular expressions are typically used to describe patterns of characters for which finite automata can be automatically constructed to recognize tokens in a language. Possibly the most notable regular expression search tool is the Unix utility *grep*.

[0031] In the context of describing amino acid sequence patterns, a simplified set of regular expression capabilities is typically employed. Amino acid sequence patterns defined by these simple regular expression rules end up looking quite similar to PROSITE patterns, both in appearance and in result. A regular expression description for an amino acid sequence may be created according to the following rules:

[0032] (1) Use capital letters for amino acid residues and put a "-" between two amino acids (not required).

[0033] (2) Use "[...]" for a choice of multiple amino acids in a particular position. [LIVM] means that any one of the amino acids L, I, V, or M can be in that position.

[0034] (3) Use "{...}" to exclude amino acids. Thus, {CF} means C and F should not be in that particular position. In some systems, the exclusion capability can be specified with a "^" character. For example, ^G would represent all amino acids except Glycine, and [^ILMV] would represents all amino acids except I, L, M, and V.

[0035] (4) Use "x" or "X" for a position that can be any amino acid.

[0036] (5) Use "(n)", where n is a number, for multiple positions. For example, x(3) is the same as "xxx".

[0037] (6) Use "(n1,n2)" for multiple or variable positions. Thus, x(1,4) represents "x" or "xx" or "xxx" or "xxxx".

[0038] (7) Use the symbol ">" at the beginning or end of the pattern to require the pattern to match the N or C terminus. For example, ">MDEL" (SEQ ID NO: 108) finds only sequences that start with MDEL (SEQ ID NO: 108). "DEL>" finds only sequences that end with DEL.

[0039] The regular expression, "[LIVM]-[VIC]-x (2)-G-[DENQTA]-x-[GAC]-x (2)-[LIVMFY](4)-x (2)-G" illustrates a 17 amino acid peptide that has: an L, I, V, or M at position 1; a V, I, or C at position 2; any residue at positions 3 and 4; a G at position 5 and so on ....

[0040] Other similar formats are in use as well. For example, the Basic Local Alignment Search Tool (BLAST) is a well-known system available on the Internet, which provides tools for rapid searching of nucleotide and protein databases. BLAST accepts input sequences in three formats: FASTA sequence format, NCBI Accession numbers, or GenBank sequence numbers. However, these formats are even more simple in structure than regular expressions or PROSITE patterns. An example sequence in FASTA format is (SEQ ID NO: 3):

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

[0041] Features of the BLAST system include sequence comparison algorithms that are used to search sequence databases for regions of local alignments in order to detect relationships among sequences which share regions of similarity. However, the BLAST tools are limited in terms of the structure of amino acid

sequences that can be discovered and located. For example, BLAST is not capable of searching for a sequence that has "at least one lysine residue located six to ten amino acid residues from a second lysine residue," as required by a Replikin pattern, for example. Nor is BLAST capable of searching for amino acid sequences that contain a specified percentage or concentration of a particular amino acid, such as a sequence that has "at least 6% lysine residues."

Need for Replikin Search Tools

[0042] As can be seen from its definition, a Replikin pattern description cannot be represented as a single linear sequence of amino acids. Thus, PROSITE patterns and regular expressions, both of which are well suited to describing ordered strings obtained by following logical set-constructive operations such as negation, union and concatenation, are inadequate for describing Replikin patterns.

[0043] In contrast to linear sequences of amino acids, a Replikin pattern is characterized by attributes of amino acids that transcend simple contiguous ordering. In particular, the requirement that a Replikin pattern contain at least 6% lysine residues, without more, means that the actual placement of lysine residues in a Replikin pattern is relatively unrestricted. Thus, in general, it is not possible to represent a Replikin pattern description using a single PROSITE pattern or a single regular expression.

[0044] Accordingly, there is a need in the art for a system and method to scan a given amino acid sequence and identify all instances of a Replikin pattern. Similarly, there is a need in the art for a system and method to search protein databases and amino acid databases for amino acid sequences that match a Replikin pattern. Additionally, there is a need in the art for a generalized search tool that permits researchers to locate amino acid sequences of arbitrary specified length that includes any desired combination of the following characteristics: (1) a first amino acid residue located more than N positions and less than M positions away from a second amino acid residue; (2) a third amino acid residue located anywhere in the sequence; and (3) the sequence contains at least R percent of a fourth amino acid residue. Finally, the shortcomings of the prior art are even more evident in research areas relating to disease prediction and treatment. There is a significant need in the art for a system to predict in advance the occurrence of disease (for example, to predict strain-specific influenza epidemics) and similarly to enable synthetic vaccines to be designed based on amino acid sequences or amino acid motifs that are discovered to be conserved over time and which have not been

previously detectable by prior art methods of searching proteins and amino acid sequences.

Brief Description Of The Drawings

[0045] FIG. 1 is a conversion table that enables amino acids to be encoded as single alphabetic characters according to a standard supplied by the International Union of Pure and Applied Chemistry (IUPAC).

[0046] FIG. 2 is a printout of a human cancer protein (SEQ ID NO: 4) obtained by searching a protein database maintained by the National Center for Biotechnology Information (NCBI).

[0047] FIG. 3 is a conversion table illustrating a correspondence between nucleic acid base triplets and amino acids.

[0048] FIG. 4 is a graph illustrating a rapid increase in the concentration of Replikin patterns in a selected strain of Hemagglutinin prior to the outbreak of three "Bird Flu" epidemics.

[0049] FIG. 5 is a table illustrating selected examples of Replikin patterns that have been found in various organisms (SEQ ID NOS 10, 18, 43, and 48-89).

[0050] FIG. 6 is a high-level block diagram of a computer system incorporating a system and method for identifying Replikin patterns in amino acid sequences, in accordance with an embodiment of the present invention.

[0051] FIG. 7 is a simple flow chart illustrating a general method for locating a Replikin pattern in a sequence of amino acids, according to an embodiment of the present invention.

[0052] FIG. 8 is a flow chart illustrating a generalized method for locating a plurality of Replikin-like patterns in a sequence of amino acids, according to an embodiment of the present invention.

[0053] FIG. 9 is a source code listing containing a procedure for discovering Replikin patterns in a sequence of amino acids, in accordance with an embodiment of the present invention.

[0054] FIG. 10. is a table illustrating Replikin scaffolds occurring in substantially fixed amino acid positions in different proteins (SEQ ID NOS 5-9, 11-17, 19-42, 44-47, 90-107, and 2 are disclosed respectively from top to bottom).

[0055] FIG. 11 is a simplified block diagram of a computer system platform useful with the present invention.

Detailed Description

[0056] Embodiments of the present invention are directed to a system and method for identifying and/or locating complex patterns in an amino acid sequence. According to an aspect of the present invention, techniques are provided to facilitate queries of protein databases. For protein descriptions received in response to the queries, embodiments of the present invention may scan the received protein descriptions to identify and locate Replikin patterns. According to an embodiment, a Replikin pattern is a sequence of from 7 to about 50 amino acids that include the following three (3) characteristics, each of which may be recognized by an embodiment of the present invention: (1) the sequence has at least one lysine residue located six to ten amino acid residues from a second lysine residue; (2) the sequence has at least one histidine residue; and (3) at least 6% of the amino acids in the sequence are lysine residues. Another embodiment of the present invention may identify and/or locate a complex amino acid sequence having specified length constraints, which further includes any combination of the following characteristics: (1) a first amino acid residue located more than N positions and less than M positions away from a second amino acid residue; (2) a third amino acid residue located anywhere in the sequence; and (3) at least R percent of a fourth amino acid residue. According to yet another embodiment, the present invention may count occurrences of the identified amino acid sequences and may report the counted occurrences, either as raw absolute values or as ratios of the number of identified amino acid sequences per N amino acids in the protein. Still another embodiment of the present invention may analyze the evolution of identified amino acid sequence patterns in variants of a given protein over time, and may also analyze the similarities and differences between instances of identified amino acid sequence patterns across a plurality of different proteins over time. As a result of the analysis, yet another embodiment of the present invention may identify potential amino acid scaffolding structures that appear to be preserved over time and across different proteins, as component elements of the identified amino acid sequence patterns mutate and/or evolve.

[0057] Embodiments of the present invention will be described with reference to the accompanying drawings, wherein like parts are designated by like reference

numerals throughout, and wherein the leftmost digit of each reference number refers to the drawing number of the figure in which the referenced part first appears.

[0058] FIG. 6 is a high-level block diagram of a computer system incorporating a system and method for identifying Replikin patterns in amino acid sequences, in accordance with an embodiment of the present invention. As shown in FIG. 6, computer workstation 610 may be a computer having a processor and a memory configured to permit a researcher to search protein databases and to scan protein descriptions for selected amino acid patterns. To accomplish these functions, computer workstation 610 may include protein and amino acid research system 630, which may receive instructions from a user/researcher to conduct protein searching and amino acid scanning operations. According to an embodiment, protein and amino acid research system 630 may further include amino acid sequence scanner 640 that scans and searches retrieved protein and amino acid sequences for specific patterns of amino acids, including Replikin patterns. Protein and amino acid research system 630 may communicate with network interface 620 to obtain protein sequences and amino acid sequences from resources on network 660, which may include the Internet. Alternatively, protein and amino acid research system 630 may obtain protein sequences and amino acid sequences from a local protein database 650. In addition, protein and amino acid research system 630 may obtain protein sequences and amino acid sequences directly from other input means, such as keyboard input. Protein and amino acid research system 630 may also communicate with network interface 620 to transmit results to other computers on network 660.

Scanning for Replikin Patterns

[0059] Embodiments of the present invention may include a generalized method and system for identifying complex patterns of amino acids within proteins. For any protein definition identified or selected by protein and amino acid research system 630, the user may direct embodiments of the invention to search for a variety of complex patterns of amino acids. As an example of one pattern of amino acids, the present invention provides a method for identifying nucleotide or amino acid sequences that include a Replikin pattern. FIG. 7 is a simple flow chart illustrating a general method for locating a Replikin pattern in a sequence of amino acids, according to an embodiment of the present invention. The method 700 may begin after a sequence of amino acids has been obtained. Typically, the sequence of amino acids may be represented by alphabetic characters according to the code

supplied in FIG. 1. However, other encodings are envisioned by the present invention as well.

[0060] Referring to FIG. 7, once a sequence of amino acids has been obtained, the sequence is searched for a Replikin pattern (710), which comprises a subsequence (or string) of amino acids that includes the following characteristics:

[0061] (1) the string contains from 7 to about 50 amino acids;

[0062] (2) the string contains at least one lysine residue located 6 to 10 positions from a second lysine residue;

[0063] (3) the string contains at least one histidine residue; and

[0064] (4) the string contains at least 6% lysine residues.

[0065] Once a string of amino acids is found to match the Replikin pattern, the string may be identified or marked (720) accordingly.

[0066] A given sequence of amino acids may contain many subsequences or strings that match the Replikin pattern. Additionally, Replikin patterns may overlap each other. Thus, to locate and identify all possible Replikin patterns in a sequence of amino acids, method 700 may be invoked iteratively for each subsequence of amino acids contained within the original sequence of amino acids.

[0067] When method 700 is invoked iteratively to identify and locate all possible Replikin patterns in an amino acid sequence, an embodiment of the present invention may count the number of resulting Replikin patterns. A Replikin count may be reported as an absolute number. Additionally, embodiments of the invention may also determine a ratio of the number of Replikins per N amino acids in the sequence. For example, an embodiment may determine that a given protein contains a ratio of 6 Replikins for every 100 amino acids. Replikin ratios have been shown by laboratory experiment and by epidemiological evidence to correlate directly to the rate that a given protein replicates. Rapid replication of proteins may be an indication of disease. For example, the presence of relatively high ratios of Replikin patterns has been correlated to epidemics of influenza. Similarly, an increase in the count of Replikin patterns observed in a protein over time may also be an indication of future disease caused by the organism from which the protein was obtained (*see, e.g.*, FIG. 4). Thus, the ability to detect and count Replikin patterns within sequences of amino acids is a significant advantage of the present invention.

[0068] Still referring to FIG. 7, embodiments of the present invention may utilize method 700 to identify and locate other complex patterns of amino acids, which exhibit characteristics similar to Replikin patterns. That is, although some embodiments of the present invention may specify exact values for: (1) distances between amino acids, (2) acceptable lengths of recognized amino acid sequences, and (3) the percentage or concentration of specific amino acids, these exact values may also be expressed as variables. Thus a researcher may employ an embodiment of the present invention to identify sequences of amino acids in a protein that have the following characteristics:

[0069] (1) the sequence contains from *rmin* to *rmax* amino acids;

[0070] (2) the sequence contains at least one lysine residue located *kmin* to *kmax* amino acid residues from a second lysine residue;

[0071] (3) the sequence contains at least one histidine residue; and

[0072] (4) the sequence contains at least *kpercent* lysine residues.

[0073] FIG. 8 is a flow chart illustrating a generalized method 800 for locating a plurality of Replikin-like patterns in a given sequence of amino acids, according to an embodiment of the present invention. The method 800 begins by locating a first lysine residue in the given sequence (810). Then, the method 800 may determine whether a second lysine residue resides within *kmin* to *kmax* positions of the first lysine residue (820). As indicated in FIG. 8, *kmin* and *kmax* define the limits on the distance between the first and second lysine residues. For a typical Replikin pattern, *kmin* will equal 6 and *kmax* will equal 10. However, these values may be varied by a researcher interested in discovering other similar patterns.

[0074] Once method 800 has identified two lysine residues that are close enough to each other (820), the method 800 may examine every histidine residue that resides within *rmax* positions of both the first and second lysine residues (830). When method 800 is employed to identify and locate typical Replikin patterns, *rmax* will usually be set to equal 50. For every histidine residue that resides within *rmax* positions of the two lysine residues identified in steps (810) and (820), method 800 will construct the shortest string of amino acid residues that includes the first lysine residue, the second lysine residue, and the identified histidine residue (840). Then, method 800 will determine whether the length of that shortest string is within the desired range – that is, whether it contains at least *rmin* amino acid residues and no more than *rmax* amino acid residues (850). Finally, if the identified string of amino

acids also contains at least *kpercent* of lysine residues (860), the string will be identified as matching the desired Replikin-like pattern (870).

[0075] Still referring to FIG. 8, it is apparent that method 800 may identify several Replikin-like patterns from a single given amino acid sequence. This may happen because method 800 may examine more than one histidine residue that resides within *rmax* positions of the two identified lysine residues. Each identified histidine residue may, in combination with the two lysine residues, match the desired Replikin-like pattern.

[0076] One embodiment of the method illustrated by FIG. 8 is shown in FIG. 9, which is a source code listing containing a procedure for discovering all Replikin patterns present in a given sequence of amino acids, in accordance with an embodiment of the present invention. The "match" procedure shown in FIG. 9 is programmed in an interpreted shell language called "Tcl" and recognizes Replikins in a straightforward fashion. As known in the art, the "Tool Command Language" or Tcl (pronounced "tickle") is a simple interpreted scripting language that has its roots in the Unix command shells, but which has additional capabilities that are well-suited to network communication, Internet functionality and the rapid development of graphical user interfaces.

[0077] Alternative methods of recognizing Replikin patterns are also covered by the teachings of the present invention. For example, the match procedure shown in FIG. 9 could be implemented in other programming languages such as Java or C or C++. Additionally, alternative embodiments of the Replikin recognizing algorithm may identify the characteristics of a Replikin pattern in any order, and may also traverse component amino acid sequences and subsequences using recursive techniques, iterative techniques, parallel processing techniques, divide-and-conquer techniques or any combination thereof.

Protein Search Engine

[0078] Returning to FIG. 6, the present invention may include a search engine to access and interact with amino acid and protein databases, either locally or over a network such as the Internet, to retrieve protein definitions. For example, protein and amino acid research system 630 may accept protein search criteria from a user, and may then access a plurality of on-line amino acid and protein database search engines to retrieve protein definitions that match the supplied search criteria. Protein database search criteria may comprise any text string that may form a valid

search term in any of the on-line protein or amino acid search engines. Typically, these search criteria relate to text that may be found in the printout that describes each specific protein. For example, if the user supplied the search criteria "influenza type A," embodiments of the present invention may forward this text string to a plurality of Internet protein and amino acid search engines, each of which may then return any protein descriptions found in their databases that contained the terms "influenza type A." Employing amino acid sequence scanner 640, each of the returned protein descriptions may be scanned for the presence of Replikin patterns.

[0079] Additional embodiments of the present invention may permit a user to select or de-select a plurality of Internet protein search engines and to customize the search criteria and protein retrieval capabilities of the present invention for each of the selected on-line protein search engines. Moreover, embodiments of the invention may also permit a user to access a local protein database 650 or to supply a specific protein definition directly, for example, by supplying a local file name containing the protein definition, or by other methods known in the art for supplying parameters to computer software.

Replikin Analysis

[0080] Embodiments of the present invention may be employed not only to identify and locate Replikin patterns in amino acid sequences. Embodiments may also be used to discover and analyze similarities in the structure of Replikin patterns occurring in different proteins, or to analyze different Replikin patterns occurring in the same protein over time. FIG. 10. for example, is a table illustrating a Replikin "fixed scaffold" structure that was preserved in a "Bird Flu" influenza virus over an 87 year period from 1917 to 2004. Embodiments of the present invention may assemble a number of discovered Replikin patterns in proteins, including Replikin patterns discovered in variants of the same protein. Along with each Replikin pattern, embodiments of the present invention may also associate a date when each protein was first identified. When directed by a researcher, an embodiment may sort and display a plurality of selected Replikin patterns according to content, date or other criteria, in order to reveal substantially fixed amino acid structures that have been preserved in Replikin patterns over time and which may be present in different proteins as well as variants of the same protein. Further, when directed by a researcher, an embodiment may employ known methods of pattern analysis to compare a plurality of selected Replikin patterns in order to identify such fixed amino acid structures automatically. As an example, in FIG. 10,

the illustrated Replikin patterns appear to demonstrate – in this case – a relatively fixed scaffold structure of (usually) 29 amino acids that begins with a pair of lysine residues (kk) at the amino terminal, ends with a pair of histidine residues (hh) at the carboxyl terminal, and contains a lysine residue in either position 8, 10 or 11. This conservation of scaffold structure over decades permits synthetic vaccines to be prepared rapidly and inexpensively. To synthesize such vaccines after a Replikin scaffolding structure has been identified, a researcher may select elements of that scaffolding structure that are conserved over time and which are also present in a current variant of a protein. A vaccine may then be prepared based on the selected elements from the scaffolding structure. Because such vaccines are based on conserved scaffolding structures, they may be effective for multiple years and may also be developed well in advance of an anticipated outbreak.

[0081] The discovery of Replikins themselves, as well as embodiments of the present invention for identifying and locating Replikin patterns, provides targets for the identification of pathogens, as well as facilitates the development of anti-pathogen therapies, including vaccines. In general, knowledge of and identification of the Replikin family of peptides enables development of effective therapies and vaccines for any organism that harbors Replikins. Specifically, identification of Replikins provides for the detection of viruses and virus vaccine development, including the influenza virus. Further, identification of Replikins also provides for the detection of other pathogens, such as malaria, anthrax and small pox virus, in addition to enabling the development of therapies and vaccines that target Replikin structures. Additional examples provided by the identification of Replikins include the detection of infectious disease Replikins, cancer immune Replikins and structural protein Replikins.

[0082] Embodiments of the present invention enable important Replikin patterns of amino acids to be recognized, located and analyzed in manners that are not found in the prior art. Using prior art capabilities, researchers have been limited in by existing techniques for describing sequences of amino acids. Indeed, limitations of the prior art have in some ways dampened research in this field, since heretofore it has not been possible to specify sequences of amino acids that comprise non-linear attributes. Until the development of the methods and embodiments of the present invention, descriptions of amino acid sequences were limited to linear sequences containing, at most, repetitive substrings and logical constraints on substring content. Embodiments of the present invention enable a

new class of amino acid sequences to be discovered, located and analyzed using tools not found in the prior art. This new class of amino acids is characterized by attributes such as specific amino acid concentration and distance relationships between specific amino acids. These attributes transcend simple contiguous ordering and thus are not easily described, discovered or located by existing methods known in the art.

[0083] The functionality of the foregoing embodiments may be provided on various computer platforms executing program instructions. One such platform 1100 is illustrated in the simplified block diagram of FIG. 11. There, the platform 1100 is shown as being populated by a processor 1160, which communicates with a number of peripheral devices via a bus subsystem 1150. These peripheral devices typically include a memory subsystem 1110, a network interface subsystem 1170, and an input/output (I/O) unit 1180. The processor 1160 may be any of a plurality of conventional processing systems, including microprocessors, digital signal processors and field programmable logic arrays. In some applications, it may be advantageous to provide multiple processors (not shown) in the platform 1100. The processor(s) 1160 execute program instructions stored in the memory subsystem 1110. The memory subsystem 1110 may include any combination of conventional memory circuits, including electrical, magnetic or optical memory systems. As shown in FIG. 11, the memory system may include read only memories 1120, random access memories 1130 and bulk storage 1140. Memory subsystem 1110 not only stores program instructions representing the various methods described herein but also may store the data items on which these methods operate. Network interface subsystem 1170 may provide an interface to outside networks, including an interface to communications network 1190 comprising, for example, the Internet. I/O unit 1180 would permit communication with external devices, which are not shown.

[0084] Several embodiments of the present invention are specifically illustrated and described herein. However, it will be appreciated that modifications and variations of the present invention are covered by the teachings of the present invention without departing from the spirit and intended scope of the invention. Additionally, the teachings of the present invention may be adaptable to other sequence-recognizing problems that have heretofore been addressed using sequential linear analyses limited to the identification of specific sequences of component elements.

What is claimed is:

1. A machine-readable medium having stored thereon executable instructions that when executed by a processor, cause the processor to recognize a pattern of amino acids, comprising:

locating at least a pair of first amino acid residues within a predetermined first distance of each other in an encoded sequence of amino acid residues;

locating a second amino acid residue within a predetermined second distance of each member of the pair of first amino acid residues;

identifying a string of amino acid residues, said string containing the pair of first amino acid residues and the second amino acid residue;

determining the percentage of first amino acid residues in the string; and

recognizing the string as the pattern of amino acids if the percentage of first amino acid residues in the string is at least a predetermined percentage.

2. A machine-readable medium having stored thereon executable instructions that when executed by a processor, cause the processor to recognize a pattern of amino acids, comprising:

locating a string of amino acids within an encoded representation of a sequence of amino acid residues defining a protein, the string conforming to a regular expression including at least a first amino acid and a second amino acid;

discarding the string of amino acids as not conforming to the pattern when the distance between a pair of occurrences of the second amino acid in the string is outside a first predetermined range;

discarding the string of amino acids as not conforming to the pattern when the distance between an occurrence of the first amino acid and an occurrence of the second amino acid is outside a second predetermined range; and

recognizing the string as the pattern of amino acids if the percentage of second amino acids in the string is at least a predetermined percentage.

3. A machine-readable medium having stored thereon executable instructions that when executed by a processor, cause the processor to recognize a scaffolding structure of amino acids, comprising:

assembling a list of Replikin patterns occurring in a plurality of proteins; and

identifying the scaffolding structure as a pattern of substantially fixed amino acid residues occurring in each Replikin pattern in the list wherein each Replikin pattern in the list is recognized by a method comprising:

determining whether a first lysine residue resides within 6 to 10 positions from a second lysine residue in a subsequence of amino acid residues;

if so, identifying a string of 7 to 50 consecutive amino acid residues in the subsequence of amino acid residues, said string containing the first lysine residue, the second lysine residue and a histidine residue;

calculating a percentage of lysine residues in the string; and

recognizing the string as a Replikin pattern if the percentage of lysine residues is at least 6 percent.

4. The machine-readable medium of claim 3, wherein the plurality of proteins includes variants of the same protein.

5. The machine-readable medium of claim 3, wherein the plurality of proteins includes different proteins.

6. A method of synthesizing a vaccine, comprising:

assembling a list of Replikin patterns occurring in variants of a protein associated with a disease-causing organism;

identifying a scaffolding structure of amino acids as a pattern of substantially fixed amino acid residues occurring in each Replikin pattern in the list;

selecting elements of the scaffolding structure, wherein said elements are conserved over time and wherein said elements are represented in a current variant of the protein; and

synthesizing a vaccine based on the selected elements,

wherein each Replikin pattern in the list is recognized by a method comprising:

determining whether a first lysine residue resides within 6 to 10 positions from a second lysine residue in a subsequence of amino acid residues;

if so, identifying a string of 7 to 50 consecutive amino acid residues in the subsequence of amino acid residues, said string containing the first lysine residue, the second lysine residue and a histidine residue;

calculating a percentage of lysine residues in the string; and

recognizing the string as a Replikin pattern if the percentage of lysine residues is at least 6 percent.

7. The method of claim 6, wherein the disease-causing organism is influenza.

8. The machine-readable medium of claim 1 wherein said pattern of amino acids further comprises a residue of said first amino acid at one terminus of said string and a residue of said first amino acid or a residue of said second amino acid at the other terminus of said string.

9. The machine-readable medium of claim 3 wherein each Replikin pattern in the list is recognized by said method further comprising identifying a lysine residue at one terminus of said string and a lysine residue or histidine residue at the other terminus of said string.

10. The method of synthesizing a vaccine of claim 6 wherein said each Replikin pattern in said list is recognized by said method further comprising identifying a lysine residue at one terminus of said string and a lysine residue or histidine residue at the other terminus of said string.

11. A vaccine synthesized by the method of any one of claims 6, 7 or 10.

12. The machine-readable medium of any one of claims 1 to 3, substantially as herein described with reference to any one of the Examples and/or Figures 4 to 11 thereof.

13. The machine-readable medium of any one of claims 1 to 5, 8 or 9, substantially as herein described.

14. The method of claim 6, substantially as herein described with reference to any one of the Examples and/or Figures 4 to 11 thereof.

15. The method of any one of claims 6, 7 or 10, substantially as herein described.