Available online at www.sciencedirect.com
ScienceDirect
Physica A 370 (2006) 651–662
PHYSICA A
www.elsevier.com/locate/physa

Probabilistic analysis of the frequencies of amino acid pairs within characterized protein sequences

Shiyi Shen[a], Bo Kai[a], Jishou Ruan[a,*,1], J. Torin Huzil[b], Eric Carpenter[b], Jack A. Tuszynski[b]

[a]*College of Mathematical Science and LPMC, Nankai University, Tianjin 300071, PR China*
[b]*Department of Oncology, Division of Experimental Oncology, Cross Cancer Institute, University of Alberta, 11560 University Avenue, Edmonton, Canada AB T6G 1Z2*

Received 3 January 2006; received in revised form 22 February 2006
Available online 3 April 2006

Abstract

Here, we describe a unique probabilistic evaluation of the 20, naturally occurring, amino acids and their distributions within the Swiss-Prot and Complete Human Genebank databases. We have developed a computational technique that imparts both directionality and length constraints into searches for unique combinations of amino acids within protein sequences. Using statistical approaches, we have carried out searches of all possible two- and three-residue motifs contained within these databases. This technique is based on the unusually high occurrence of a small number of these motifs when compared to the expected probability of finding a specific residue grouping within a given database. Subsequent filtering of this search to identify such unique combinations has provided several examples that can be used as markers to identify particular proteins within or across databases. We focus on three of these motifs, which were found to be of greatest interest to us. The CC, CM and a combination of the two, CCM motifs all occur either more or less frequently than would be predicted based on standard amino acid distributions within the entire human proteome.

Keywords: Amino acid motif; Fast detection method; Significance test

1. Introduction

The primary sequence of a protein refers to an ordered string of amino acids that belong to a finite set. It is therefore natural to expect that the distribution of amino acids, across all protein sequences, would appear to be fixed, given no mutations over a specific period of time. Of the 20 naturally occurring amino acids, each seems to appear randomly located at a particular position within the sequence of a protein. However, when the sequences of a group of proteins are compared to one another, specific patterns begin to emerge. These patterns are commonly referred to as motifs and frequently produce similar structural elements and in many

*Corresponding author.
E-mail address: jtus@phys.ualberta.ca (J.A. Tuszynski).
[1]Supported by Liuhui Center for Applied Mathematics, and the China–Canada exchange program administered by MITACS (The Mathematics of Information Technology and Complex Systems).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.03.004

cases, similar functions between the proteins that contain them. In recent years, significant efforts have been made to reduce sequences into discrete motifs, so that each can be linked to some specific biological function based on particular rules [1,2]. Databases that consist of identifiable amino acid motifs, such as PROSITE, Pratt and EMOTIF, have been constructed to allow researchers an easy classification of unknown proteins into families based on their common motifs [3–7]. All of these methods share one property, they depend on specific alignment algorithms, such as the basic local alignment search tool (BLAST or PSI-BLAST) to perform the sequence comparison between diverse families of proteins being analyzed [5,8]. These methods have already led to a number of key discoveries and offer the promise of determining biologically meaningful information about function from a protein's primary sequence.

It is generally accepted that a strong correlation exists between a protein's amino acid sequence and its resulting structure and function [9–11]. Techniques that can quickly determine significant similarities between protein sequences are valuable in predicting an unknown protein's function within the cell. Here we describe a procedure that, by computing the frequency of key amino acid pairs, can quickly classify large groups of proteins into functional groups. In order to uncover any relevant amino acid strings, we have utilized statistical methods applicable to datasets with large fluctuations in the search of human protein sequences within both the Swiss-Prot and Complete Human Genebank databases.

The goal of this manuscript is to demonstrate how these two residue motifs can be used to develop rapid methods for searching protein or genomic databases in order to locate potentially interesting or aberrant protein sequences. As an example, we describe the distribution of select amino acid pairs within the genome of the Coronavirus, Tor2, which is responsible for severe acute respiratory syndrome (SARS) [12]. We have examined additional sets, but the limited format of this paper precludes a large presentation of potential applications.

2. Methods

2.1. Breakdown of sequences

The set of all 20, naturally occurring, amino acids are described by their standard single letter representations, and are taken as the complete set contained within all protein sequences as shown below:

$$Z_q = (A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V). \tag{2.1}$$

Residues that are arranged in a certain order can be referred to as vectors and are represented symbolically as

$$b^{(n)} = (b_1, b_2, \ldots, b_n), \quad b_j \in Z_q. \tag{2.2}$$

The coordinates of each vector represent the amino acid symbols contained within any given protein, where n is the total length of the vector. The collection of all vectors $b^{(n)}$ in the proteome can therefore be denoted simply by $Z_q^{(n)}$.

The underlying structure of a protein sequence, which is the order of the amino acids, has unique and identifiable characteristics. As an analogy, in written English, the letter q is almost always followed by u. Also, the frequency of the word "and" is much higher than that of the related three-letter strings: "nad", "nda", "dan" and "dna", which are simply permutations of the same set of individual letters to construct a word with a different meaning. Utilizing similar characteristics, we can intuitively scan for unique sequence vectors contained within any given protein. We expand on these characteristics, by utilizing the concepts of affinity and order stability in Section 2.2.

Using the entire Swiss-Prot database (release number 43.0) as the basis for our calculations, we have constructed a probability table for the occurrence of the 20 naturally occurring amino acids (see Table 1) [13]. In addition to the Swiss-Prot database, we have also included data from Complete Human Genebank (NCBI Build Number 34 March, 2004) as a source for all characterized and hypothetical human protein sequences.

Table 1
Probability distribution of all amino acids contained in the Swiss-Prot and the frequency distribution in the Complete Human Genebank (square brackets) databases

$p(A) = 0.0777$	$p(C) = 0.0157$	$p(D) = 0.053$	$p(E) = 0.0656$	$p(F) = 0.0405$
[0.0704]	[0.0231]	[0.0484]	[0.0692]	[0.0378]
$p(G) = 0.0691$	$p(H) = 0.0227$	$p(I) = 0.0591$	$p(K) = 0.0595$	$p(L) = 0.096$
[0.0675]	[0.0256]	[0.0450]	[0.0565]	[0.0984]
$p(M) = 0.0238$	$p(N) = 0.0427$	$p(P) = 0.0469$	$p(Q) = 0.0393$	$p(R) = 0.0526$
[0.0237]	[0.0368]	[0.0610]	[0.0465]	[0.0552]
$p(S) = 0.0694$	$p(T) = 0.055$	$p(V) = 0.0667$	$p(W) = 0.0118$	$p(Y) = 0.0311$
[0.0799]	[0.0534]	[0.0613]	[0.0121]	[0.0282]

The occurrence of each individual amino acid has been averaged over the total number of amino acids found within each of the databases.

2.2. Mathematical background

We use the symbol Ω to denote the Swiss-Prot database, and the letter M to denote the number of proteins in the database (133,723). Hence we can state that $\Omega = \{A_s, s = 1, 2, \ldots, M\}$ where $A_s = (a_1, 1, a_2, 2, \ldots, a_s, N_s)$ is the sth protein, a_i with $i \in Z_q$ is the ith amino acid in the sequence, and N_s is the length of the protein sequence A_s. For any vector $b^{(n)}$, if there is a sequence A_s in Ω such that $b^{(n)}$ is a segment of A_s, then we can say $b^{(n)}$ occurs in Ω. We can then use $N_\Omega(b^{(n)})$ to denote the total number of outcomes that $b^{(n)}$ has occurred in Ω, and let $N_\Omega(n)$ be the total number of all vectors occurring in Ω with length n. Therefore,

$$p(b^{(n)}) = \frac{N_\Omega(b^{(n)})}{N_\Omega(n)} \tag{2.3}$$

is the probability that the vector $b^{(n)}$ occurs in Ω. We then introduce

$$k(p, q; b^{(n)}) = \log_2 \frac{p(b^{(n)})}{q(b^{(n)})}, \tag{2.4}$$

where $p(b^{(n)})$ and $q(b^{(n)})$ are any two probability distributions on $Z_q^{(n)}$, and $b^{(n)}$ belongs to $Z_q^{(n)}$. The quantity $k(p, q; b^{(n)})$ in Eq. (2.4) is defined for any two probability distributions $p(b^{(n)})$ and $q(b^{(n)})$ on $Z_q^{(n)}$ where the distribution q arises from independent letter (i.e., amino acid) probabilities. Then, $k(p, q; b^{(n)})$ can be viewed as a random variable about $b^{(n)}$ and its probability distribution is defined as follows:

$$F_{(p,q)}(x) = P_I\{k(p, q; b^{(n)}) \leqslant x\} = \sum_{b^{(n)}: k(p,q,b^{(n)}) \leqslant x} p(b^{(n)}). \tag{2.5}$$

The expectation value $\mu(p, q)$ and variance $\sigma^2(p, q)$ of $p(b^{(n)})$ are expressed mathematically as

$$\mu(p, q) = \sum_{b^{(n)} \in Z_q^{(n)}} p(b^{(n)}) k(p, q; b^{(n)}), \tag{2.6}$$

$$\sigma^2(p, q) = \sum_{b^{(n)} \in Z_q^{(n)}} p(b^{(n)}) [k(p, q; b^{(n)}) - \mu(p, q)]^2. \tag{2.7}$$

Hence we see that $\mu(p, q)$ is simply the Kullback–Leibler entropy [14]. As a consequence, $\mu(p, q)$ is always non-negative and $\mu(p, q) = 0$ if and only if $p(b^{(n)}) = q(b^{(n)})$. As a result, the amino acid sequence of $b^{(n)}$ can be represented by $k(p, q; b^{(n)})$.

In Eq. (2.4), the probability distributions p, q in the $k(p, q; b^{(n)})$ have several possible values. For example, if we let p be the probability of $b^{(n)}$, and only change the value of q in $k(p, q; b^{(n)})$, we obtain the following expression:

$$q(b^{(n)}) = \prod_{j=1}^{n} p(b_j) = p(b_1) p(b_2) \ldots p(b_n). \tag{2.8}$$

Based on Eqs. (2.4) and (2.8), the value of $k(p,q;b^{(n)})$ gives us the affinity for all amino acids within vector $b^{(n)}$. Here two situations are identified as especially interesting:

(i) When $k(p,q;b^{(n)}) \gg 0$ implies that among the amino acids in vector $b^{(n)}$, there is an increased affinity. That is, the vector $b^{(n)}$ occurs more frequently than expected.
(ii) When $k(p,q;b^{(n)}) \ll 0$ implies that among the amino acids in vector $b^{(n)}$ there is a decreased affinity. That is, the vector $b^{(n)}$ occurs less frequently than expected.

In a similar approach to the condition described above, we may choose $q(b^{(n)}) = p(\pi_n(b^{(n)}))$, where $\pi_n(b^{(n)})$ is a permutation of $b^{(n)}$, and π_n denotes a permutation transform. In this case, the order sensitivity of $b^{(n)}$ can also be reflected by the value of $k(p,q;b^{(n)})$ and produces the following outcomes:

(i) when $k(p,q;b^{(n)}) \gg 0$ implies that $b^{(n)}$ occurs much more often than $\pi_n(b^{(n)})$;
(ii) when $k(p,q;b^{(n)}) \ll 0$ implies that $\pi_n(b^{(n)})$ occurs much more often than $b^{(n)}$.

In addition to the concepts of affinity and order sensitivity, we can also design other analysis techniques by choosing alternate representations of $q(b^{(n)})$. Quantifying the terms $k(p,q;b^{(n)})$, we can represent vector frequency numerically. In cases where $k(p,q;b^{(n)}) - \mu(p,q) \geqslant \Delta\sigma(p,q)$ the corresponding vector $b^{(n)}$ is called Δ-positive, while in those cases where the opposite takes place: $k(p,q;b^{(n)}) - \mu(p,q) \leqslant -\Delta\sigma(p,q)$ the vector $b^{(n)}$ is called Δ-negative. By judiciously choosing a Δ equal to 2, we obtained 16, two-residue, positive-pairs, and 8 two-residue negative-pairs (see Table 2) out of a total of 400 possibilities of amino acid pairs.

Table 2
Based on the (Δ,σ)-criterion, we obtained the following amino acid motifs that are statistically significant (calculated frequency) when compared to the frequency of amino acids described in Table 1

Amino acid pair	Calculated frequency (%)	Expected frequency (%)
AA	0.78	0.60
RR	0.36	0.28
NN	0.22	0.18
CC	0.04	0.02
CH	0.04	0.03
HC	0.04	0.03
QQ	0.24	0.15
EE	0.58	0.43
EK	0.48	0.39
HH	0.08	0.05
HP	0.14	0.11
KK	0.47	0.31
PP	0.31	0.22
SS	0.63	0.48
WW	0.02	0.01
YY	0.12	0.10
CM	0.03	0.04
EP	0.24	0.26
ES	0.36	0.46
HE	0.12	0.15
HK	0.11	0.14
IM	0.12	0.14
PM	0.09	0.11
WP	0.04	0.08

The first 16 residue pairs represent "positive" motifs and occur more frequently within the database than expected. The remaining eight pairs represent "negative" motifs (grayed boxes) and occur less frequently than expected. Data are taken from the Swiss-Prot database and it includes all the Δ-positive and Δ-negative pairs.

2.3. Developing detection methods

Using any one of the positive motifs from Table 2, we can develop a test method to detect changes in their distribution through any protein database. If $b^{(2)} = (b_1, b_2)$ is Δ-positive, then $p(b_1b_2) > 2^{\mu+\Delta\sigma}p(b_1)p(b_2)$, where $p(\cdot)$ is the probability of the amino acids or vectors $\cdot$ in the database. Since

$$N_\Omega(2) = \sum_{\mu=1}^{M}(N_\mu - 1), \quad p(b_j) = N_\Omega(b_j) \Big/ \sum_{\mu=1}^{M} N_\mu, \tag{2.9}$$

we have

$$N_\Omega(b^{(2)}) > \left(\frac{\sum_{\mu=1}^{M}(N_\mu - 1)}{\sum_{\mu-1}^{M} N_\mu}\right)(2^{\mu+\Delta\sigma}p(b_1))N_\Omega(b_2) \tag{2.10}$$

or

$$N_\Omega(b^{(2)}) > \left(\frac{\sum_{\mu=1}^{M}(N_\mu - 1)}{\sum_{\mu-1}^{M} N_\mu}\right)(2^{\mu+\Delta\sigma}p(b_2))N_\Omega(b_1), \tag{2.11}$$

and since

$$\sum_{\mu=1}^{M}(N_\mu - 1) \text{ is very close to } \sum_{\mu=1}^{M} N_\mu \text{ for Swiss-Prot database,} \tag{2.12}$$

we have

$$N_\Omega(b^{(2)}) > (2^{\mu+\Delta\sigma}p(b_1))N_\Omega(b_2) \tag{2.13}$$

and

$$N_\Omega(b^{(2)}) > (2^{\mu+\Delta\sigma}p(b_2))N_\Omega(b_1).$$

As a consequence of Eq. (2.13), the detection of a deleterious mutation based on the increased appearance of an arbitrary string of these residues becomes possible. For example, according to Table 2, CC is an attractive motif with $\Delta = 2$, then

$$N_\Omega(CC) > (2^{\mu+2\sigma}p(C))\left(\frac{\sum_{\mu=1}^{M}(N_\mu - 1)}{\sum_{\mu-1}^{M} N_\mu}\right)N_\Omega(C) \approx (2^{\mu+2\sigma}p(C))N_\Omega(C). \tag{2.14}$$

Let $N_s(CC)$ and $N_s(C)$ be the number of CC and the number of C in the sth protein. Then we have $N_\Omega(CC) > 0.013376N_\Omega(C)$ because $\mu + +2\sigma = 0.28245$ and $p(C) = 0.0157$ since

$$N_\Omega(CC) = \sum_{s=1}^{M} N_s(CC), \quad N_\Omega(C) = \sum_{s=1}^{M} N_s(C). \tag{2.15}$$

Based upon the result that the CC pair represents a rare event, proteins satisfying $N_s(CC) > 0.013376N_s(C)$, as defined in Eq. (2.15), suddenly become interesting. We can therefore consider the number $N_s(CC)$ as an index to develop detection methods to test for the presence of a CC containing protein within a database. Using this same reasoning, one can also use any other one of the remaining motifs in Table 2 to develop detection methods for additional proteins.

2.4. Methods based on the negative two-residue motifs

In an identical fashion to the previous example using the positive two-residue motifs in Table 2, we can also use any of the "negative" two-residue motifs to develop search methods. For example, the CM pair is a "negative" two-residue "word" with $\Delta = 2$. That is, $p(CM) < 2^{\mu-2\sigma}p(C)p(M)$. Using identical arguments to those described in Section 2.3, we obtain $N_\Omega(CM) < 0.013376N_\Omega(M)$ and $N_\Omega(CM) < 0.008512N_\Omega(C)$. Since

both $N_s(M)$ and $N_s(C)$ are small numbers for fixed proteins, the overall number of CM pairs in most proteins should be close to zero. Therefore, we can also use any of the remaining seven "negative" motifs in Table 2 to develop additional detection methods.

3. Results

3.1. The distribution of "positive" CC and "negative" CM pairs within Swiss-Prot

In our analysis of the positive and negative two-residue pairs, we have arbitrarily chosen CC and CM as candidates with which to search the Swiss-Prot database. The total number of proteins in Swiss-Prot is 133,723, amongst which the number satisfying $N_s(CC) = 0$ is 128,922, while the total number satisfying $N_s(CC) \geqslant 1$ is 14,801. Since the total number of amino acids in the Swiss-Prot is 46,120,418, we would expect to find only 11,368 of CC pairs in Swiss-Prot database based upon the amino acid frequencies given in Table 1, thus we would expect to find only at most 11,368 proteins having the CC pairs. Not surprisingly, there are only a small number of proteins that contain more than two of these CC pairs. For example, the total number of proteins in Swiss-Prot that satisfy $N_s(CC) \geqslant 7$ is only 104, amongst which, there are only a total of 23 found in humans (Tables 3 and 4).

3.1.1. Frequency of CM

In contrast to the CC motif, the total number of proteins in the Swiss-Prot database having more than one CM pair is quite small. For example, among the 133,723 proteins contained within this database, the number of proteins such that $N_s(CM) \geqslant 1$ is 1595, and the number of proteins satisfying the condition $N_s(CM) \geqslant 4$ is only 46. For human proteins in Swiss-Prot, the total number of proteins containing CM pairs is 1596, and there are only 13 proteins that satisfy $N_s(CM) \geqslant 4$ (Table 3). However, we would expect to find about 17,233 proteins having CM pairs in the Swiss-Prot database, based upon the amino acid frequencies given in Table 1, which is greater than the number observed.

3.1.2. C-rich proteins do not necessarily contain numerous CC pairs

Cysteine rich proteins and the presence of CC pairs in proteins are both concepts that have been studied previously [15,16]. If we combine these two concepts and describe them in mathematical terms, we observe that the rates $q_1 = N_s(C)/N_s$ and $q_2 = N_s(CC)/N_s(C)$ describe these two conditions, where N_s, $N_s(C)$ and $N_s(CC)$ are the numbers of all amino acids, cysteine, and CC, respectively. We can then derive

$$\left\{ proteins \middle| \frac{N_s(C)}{N_s} > p(C) + 3\sigma_1, \frac{N_s(CC)}{N_s(C)} > p(\xi_{\tau+1} = C | \xi_\tau = C) + 3\sigma_2 \right\}, \tag{3.1}$$

where σ_1 and σ_2 are the standard deviations about

$$\left\{ \frac{N_s(C)}{N_s} \middle| s = 1, 2, \ldots, M \right\} \tag{3.2}$$

and

$$\left\{ \frac{N_s(CC)}{N_s(C)} \middle| s = 1, 2, \ldots, M \right\},$$

respectively.

Table 3
Frequency of CC and CM pairs within all human protein sequences found within the Swiss-Prot database

# pairs	1	2	3	**4**	5	**6**	7	8	9	>9
CC	1678	385	136	35	14	9	7	5	4	7
CM	1308	244	31	8	3	1	0	0	1	0

Statistical thresholds for CC (6) and CM (4) pairs are shown in bold case.

Table 4
Relationship between the total number of CC pairs within human proteins and disease

Protein name	Total length	Number of C	Number of CC	Function
NIC1 (NICE-1 protein)	99	17	9	(Uncharacterized) Involved in epidermal differentiation
VTDB (Vitamin D-binding protein)	474	28	7	(Secreted-plasma) Prevents polymerization of actin
AFAM (Afamin)	599	34	8	(Secreted) Contains albumin domains
MCS (Sperm mitochondrial associated cysteine-rich protein)	116	20	9	(Cytoplasmic) Associated male infertility
ALBU (Serum albumin)	609	35	8	(Secreted-plasma) Familial dysalbuminemic hyperthyroxinemia
DJC5 (DnaJ homolog)	198	14	8	(Membrane bound) Involved in presynaptic function
DJCX (DnaJ homolog)	199	14	7	(Membrane)
KRUA (Keratin, ultra high-sulfur matrix protein A)	169	60	19	(Extracellular) Cuticle layers of differentiating hair follicles
KRUB (Keratin, ultra high-sulfur matrix protein B)	194	67	22	(Extracellular) Cuticle layers of differentiating hair follicles
CIWC (Potassium channel subfamily K member 12)	430	19	9	(Membrane protein) Probable potassium channel subunit
MDFI (Myogenic repressor I-mf)	246	29	7	(Cytoplasmic)
GRN (Granulins)	593	88	28	(Secreted) May play a role in inflammation, wound repair, and tissue remodeling
ECM1 (Extracellular matrix protein 1 [Precursor])	540	28	7	(Extracellular matrix) Lipoid proteinosis
MU5A (Mucin 5AC)	1233	95	7	(Extracellular matrix)
LTBS (Latent transforming growth factor beta binding protein, isoform 1S)	1394	139	7	(Secreted)
LTBL (Latent transforming growth factor beta binding protein)	1595	138	8	(Secreted) Involved in the assembly, secretion and targeting of TGFβ1
LYST (Lysosomal trafficking regulator)	3801	93	9	(Cytoplasmic) Chediak–Higashi syndrome (hypopigmentation)
FBN1 (Fibrillin 1)	2871	361	17	(Extracellular matrix) Ongenital contractural arachnodactyly
FBN2 (Fibrillin 2)	2911	363	17	(Extracellular matrix) Ongenital contractural arachnodactyly
VWF (Von Willebrand factor)	2813	234	8	(Secreted) Von Willebrand disease

Shown here are the top 20 of 28 human proteins from the Swiss-Prot database, which contain seven or more CC pairs. Proteins are ordered based upon their total CC content in relation to the total number of cysteine residues and the overall length of the protein.

We have found the following computational results based on Swiss-Prot database

$$\sigma_1 \approx 0.024318, \quad \sigma_2 \approx 0.073867 \quad \text{and} \quad p(\xi_{\tau+1} = C|\xi_\tau = C) \geqslant 2^{0.28245} p(C).$$

For $\Delta = 3$, we have a much simpler method to determine that a protein sequence is both C- and CC-rich if

$$\left\{proteins\left|\frac{N_s(C)}{N_s} > 0.088654\right.\right\} \cap \left\{proteins\left|\frac{N_s(CC)}{N_s(C)}\right| > 0.2413\right\}. \tag{3.3}$$

So, a much simpler method to test whether a protein is a both C- and CC-rich for $\Delta = 3$ is

$$\text{Number of } CC \times 4.14 > \text{ Number of } C,$$
$$\text{Number of } C \times 11.28 > \text{ Length of protein.}$$

Using this principle, we found a total of eight proteins that are both C- and CC-rich within the Swiss-Prot database.

For $\Delta = 2$, there is an easier method to determine whether a protein sequence is both C- and CC-rich if

$$\left\{proteins\left|\frac{N_s(C)}{N_s} > 0.064336\right.\right\} \cap \left\{proteins\left|\frac{N_s(CC)}{N_s(C)} > 0.1675\right.\right\}. \tag{3.4}$$

So, a much simpler method to test whether a protein is both C- and CC-rich for $\Delta = 2$ is based on the inequalities

$$\text{Number of } CC \times 6 > \text{ Number of } C,$$
$$\text{Number of } C \times 16 > \text{ Length of protein.}$$

Using this principle, we can get a total of 65 proteins that are both C- and CC-rich within the entire Swiss-Prot database.

For $\Delta = 1$, there is also an easier method to determine if a protein sequence is both C- and CC-rich sequence when

$$\left\{proteins\left|\frac{N_s(C)}{N_s} > 0.04\right.\right\} \cap \left\{proteins\left|\frac{N_s(CC)}{N_s(C)} > 0.093667\right.\right\}. \tag{3.5}$$

So, a much simpler method to test whether a protein is both C- and CC-rich for $\Delta = 1$ is based on the inequalities

$$\text{Number of } CC \times 11 > \text{ Number of } C,$$
$$\text{Number of } C \times 25 > \text{Length of protein.}$$

Using this principle, we have identified a total of 200 proteins that are both C- and CC-rich within the Swiss-Prot database. From these results, it becomes clear that there are many C-rich proteins and many CC-rich proteins that are not mutually inclusive. The sets of both C- and CC-rich proteins are only subsets of the set containing all C-rich proteins or that containing all CC-rich proteins.

3.2. Distribution of CC and CM in human protein sequences in Genebank

Because the Swiss-Prot database is somewhat biased to those proteins that are of interest to researchers, we have decided to examine a database that was, in essence, complete. For this, we chose the human subset sequences contained within the Complete Human Genebank database. This was the most complete and readily available source of human genetic sequences and should therefore provide us with an accurate representation of both CC and CM pair distributions. We observed that the total number of CC pairs within the human subset of sequences contained in Genebank (a total of 27,178 proteins) was 9631. These pairs were contained within a total of 6295 proteins (Table 4), a number that is substantially larger than the 3198 we obtained from Swiss-Prot data (Table 3). Since the total number of amino acids in The Complete Human Genebank is 5,777,674, we can easily get the expected number of CC in Complete Human Genebank is 1424, which is much less than this actual value 9631. The same also holds true for the CM pair, where we would expect to see 2159 CM pairs, but actually observe 5007 (Table 5). This result is surprising but true and shows that there are fundamental difference between the Human and other species in terms of the frequency of CC and CM pairs.

Table 5
Frequency of CC and CM pairs within human proteins in the Complete Human Genebank database

# pair	1	2	3	4	5	6	7	8	9	>9
CC	4651	1081	330	87	37	29	24	8	7	41
CM	3488	520	94	36	5	2	1	0	1	0

3.3. The distribution of CC and CM in proteins with at least four pairs

To create a single CM pair, it is obvious that one needs only a single C. If a protein sequence also contains a CC pair, the total number of cysteines available to create a CM pair has been reduced by a factor of 2. Using the cutoff value of four CC pairs as statistically significant, we determined the probability of finding a CM pair within proteins that already contained a CC pair (Table 5). Even though both CC and CM can occur in the same protein, the total number of proteins in Genebank that contain both CC and CM pairs is 1400 and these pairs almost always occur at different locations in the sequence. However, in some rare cases the second C within the CC pair is shared with the first C within the CM pair, giving us a new triplet CCM.

3.4. Frequency of the triplet CCM

If we filter the set of proteins described above for redundant entries, we obtain a set of 120 unique proteins that contain the CCM triplet in the human proteome. Using the probability distributions obtained from Swiss-Prot, we obtained the probability of CCM as 0.000005866. Since the total of all protein lengths in the Swiss-Prot database is 49,190,847, there should be a total of 289 CCM triplets. This value is almost identical to the predicted number of CCM of 290. This result suggests that the CCM triplet at least in Swiss-Prot occurs randomly. However, we can also analyze the CCM probability using a different approach, putting $q(CCM) = p(CC)p(M)$ into Eq. (2.4). If CCM occurs randomly, then we have $p(CCM) = p(C)p(C)p(M)$, and since $p(CC) > 2^{\mu+2\sigma}p(C)p(C)$, we will get

$$k(p, q; CCM) = \log_2 \frac{p(CCM)}{q(CCM)} = \log_2 \frac{p(C)^2}{p(CC)} \leqslant -(\mu + 2\sigma) < 0. \tag{3.6}$$

Based on Eq. (3.6), it is clear that the presence of a CC pair strongly excludes M. Therefore, the probability of the CCM triplet should occur less frequently than that of CC or M individually.

4. Discussion

We have demonstrated that certain pairs of amino acids occur within the sequences of proteins with either a greater, or lower, than expected frequency, as calculated from the observed frequencies of the individual amino acids. We expect that this can be used to rapidly classify proteins that carry these statistically uncommon pairs, into large families. As an example, we focused on the double-cysteine (CC) pair, a pair that occurred more frequently, within the Swiss-Prot and Complete Human Genebank databases. We observed that within the human proteome, the actual frequency of CC was twice that of the calculated mean frequency.

The study of C-rich proteins has had a long history and many of these proteins have been identified as extracellular components [15]. While the fact that extracellular proteins contain numerous cysteine residues is not new, the distribution of these cysteines into distinct CC pairs is interesting. Proteins containing the CC motifs have previously been mentioned in the literature; however, their surprising prevalence within protein databases has not been mentioned [15–17]. To demonstrate the statistical relevance of the CC motif, we analyzed their distribution within both the Swiss-Prot and the Complete Human Genebank databases.

If we examine those proteins within the Swiss-Prot database that contain more CC pairs than the threshold we can see that most of them are extracellular, either being secreted into the blood stream or are a part of the extracellular matrix (Table 4). It is understood that proteins with high numbers of cysteine residues are

generally stabilized by disulfide bridges and these proteins are usually found outside the cell. This, however, does not explain the subset of proteins that contain numerous CC pairs, with respect to the calculated probabilities. Further analysis is required to determine if these pairs have arisen due to some type of duplication event through evolution, leading to the increased stabilization of extracellular matrix proteins.

While the CC motif occurs more frequently than expected and the frequency of the CM pair occurs less frequently, as predicted by the amino acid frequencies within the Swiss-Prot database, the occurrence of a combination of the two, CCM, occurs more frequently. Using the genebank database, we determined the total number of human proteins containing the CCM motif is 119 and the total number of proteins is 27,178. In the Swiss-Prot database, the total number of the proteins having CCM is about 290 but the total number of proteins in this database is 128,494. In other words, the rate of all CCM proteins in human cell is twice as much as the rate in Swiss-Prot, implying that the rate of CCM proteins in human proteins is statistically significant. Using the probability distributions obtained from Swiss-Prot, the probability of a CCM triplet is 0.000005866, and total length of all human protein sequences in genebank is 5,777,674, so the number of CCM should be approximately 35, much less than 119. In other words, even though the occurrence of the CCM triplet seems to be random within the Swiss-Prot database, its frequency seems significant in human proteins.

One possible explanation as to the increased frequency of the CCM triplet within the human proteome is an increased rate of mutations at either of the two C positions (XCM or CYM). Our results suggest that the CC pair strongly excludes M because the conditional probability of M under the condition that the first two positions are occupied by CC is much less than the unconditional probability of M. So, mathematically, not only is CM a negative motif, but so is CCM. The strong competitiveness of CC and strong repulsion between C and M, and between CC and M, lead us to conclude that the mutational event leading to CCM would result from either XCM or CXM. If we let X and Y be the amino acids that correspond to the CC pair, then in addition to X-Cys-Met, we may find that the triplet Cys-Y-Met will also have a chance to change into CCM. In fact, since that the repulsion between CC and M, as well as between C and M are strong, it would imply that the number of Cys-Y-Met is greater than the number of X-Cys-Met in total protein sequences. Thus, the probability of changing Cys-Y-Met into CCM would be greater if the mutation at Y occurred at the same rate as that at X. Mutations to cysteine can arise from a single base substitution from phenylalanine, serine, tyrosine, arginine and glycine. The most relevant mutation would seem to come from serine to cysteine, where there are a total of four possible substitutions TCT, TCC, AGT, AGC to either TGT or TGC. A more detailed analysis of those residues and mutation frequencies at these positions is still required before any concrete conclusions can be made regarding the evolution of the CCM triplet in the human proteome and the discrepancy between it and the Swiss-Prot database. However, as an interesting aside, recent research has identified several proteins that contain the CCM triplet; however, the function of CCM, if any, has not been characterized [18–22].

4.1. Example of the CC and CM detection method for SARS

Having developed a formal approach to the search for interesting amino acid pairs in protein sequences, we now discuss its application in the detection of specific diseases. In our first example, we will focus on developing detection methods for the Coronavirus responsible for causing severe acute respiratory syndrome (SARS) [23].

4.1.1. Statistical results for CC and CM in the Swiss-Prot database

We have obtained the following statistical results with regard to the total number of CC and CM pairs contained in the Swiss-Prot database:

(i) The total number of these proteins satisfying $N_s(CC) \geqslant 1$ is 14,801, and 2280 of these are found in humans.
(ii) The total number of human proteins in Ω that satisfy $N_s(CC) \geqslant 6$ is 32.
(iii) The total number of the proteins in Ω that satisfy $N_s(CM) \geqslant 1$ is 12,438, 1596 of these are found in humans.
(iv) The total number of human proteins in Ω that satisfy $N_s(CM) \geqslant 4$ is only 13.

4.1.2. Statistical results of CC and CM in the SARS proteins

The genome of the SARS-associated Coronavirus contains a total of 11 open reading frames, which code for six characterized and five uncharacterized proteins [12]. After examining the complete proteome the SARS-associated Coronavirus we observed a statistically significant increase in the occurrence of the CC and CM pairs in the SARS-associated Coronavirus' genome.

Of the 11 putative proteins, four contain either CC or CM pairs. Two large poly-proteins, GI:29836505 and GI:29836495 contain 12 CC and 4 CM pairs, as well as 6 CC and 4 CM pairs, respectively. The putative spike glycoprotein, GI: 29836496 contains 3 CC and 1 CM, while the small envelope protein, GI: 29836499, contains only 1 CC pair. The remaining 6 proteins, of which two have been identified as a nucleocapsid and matrix protein, contain no CC or CM pairs. When compared to the occurrence of CC and CM pairs within all human proteins found in the Swiss-Prot database, it becomes clear that there is an increased occurrence of these pairs within the SARS-associated Coronavirus proteome.

It is clear that the frequency of CC and CM pairs is significantly higher in the SARS genome than it is in the human genome. It is our hope that these unique two residue motifs could be used as a means of identifying a virus such as SARS. While currently impossible to determine, one can see a possible clinical use for determining the frequency of CC or CM pairs in an individual. If a person has been infected by SARS, we would expect the numbers of CC and CM pairs to be much higher than in a healthy individual. If we were then able to develop a clinically feasible technique that could identify the CC or CM pair content in total cellular protein, without damaging the cell, it could be possible to implement a detection method simply by counting the numbers of CC or CM pairs in all proteins examined. If the presence of increased CC or CM pairs was indicative of infection by SARS we could develop new drugs to inhibit the translation of proteins that contain these pairs within infected cells. As a very specific example, this could be accomplished through the use of specific tRNA mimics that could block translating ribosome when these pairs appear at the P and A site simultaneously [24,25]. While tRNA mimics could normally be seen as non-specific protein synthesis inhibitors, the lower frequency of the CC or CM motifs in host proteins as compared to the relatively high occurrence in viral protein provide a starting point for the analysis of treatments using methods such as these.

5. Summary and conclusions

In this paper we have performed a detailed search of the Swiss-Prot and Complete Human Genebank database in relation to the frequencies of each of the 20 natural amino acids and motifs made up of two residues. This has provided us with a new insight into the distribution of proteins within these databases and the information that may be contained within a protein's primary amino acid sequence. Our focus in this quest has been to examine specific pairs or triplets of residues and use probability theory to assess whether or not particular strings of residues occur more frequently than their probability would indicate. By analyzing affinity and anti-affinity, we developed a procedure to search out all two- and three-residue clusters within these databases. Through this technique, we can rapidly classify large groups of proteins into families based on the occurrence of these motifs. Combining this concept with a specific knowledge of protein function, we propose novel ideas aimed at developing fast detection methods for specific diseases based on the unusually high occurrence of some of these motifs when compared with the expected probabilities. A timely example of the application of this method is given, namely the analysis of the SARS virus. This method will hopefully lead to the discovery of new diagnostic techniques and possible treatment of diseases, by recognizing the location of target domains in the expressed sequence.

References

[1] D.W. Mount, Bioinformatics: Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, New York, 2001.
[2] P. Baldi, S. Brunak, P. Frasconi, G. Soda, G. Pollastri, Exploiting the past and the future in protein secondary structure prediction, Bioinformatics 15 (1999) 937–946.
[3] B. Rost, Review: protein secondary structure prediction continues to rise, J. Struct. Biol. 134 (2001) 204–218.
[4] R.K. Hart, A.K. Royyuru, G. Stolovitzky, A. Califano, Systematic and fully automated identification of protein sequence patterns, J. Comput. Biol. 7 (2000) 585–600.

[5] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[6] A. Bairoch, Prosite: a dictionary of sites and patterns in proteins, Nucleic Acids Res. 20 (Suppl) (1992) 2013–2018.
[7] P. Bork, E.V. Koonin, Protein sequence motifs, Curr. Opin. Struct. Biol. 6 (1996) 366–376.
[8] S.Y. Shen, J. Yang, A. Yao, P.I. Hwang, Super pairwise alignment: an efficient approach to global alignment for homologous sequences, J. Comput. Biol. 9 (2002) 477–486.
[9] B. Rost, R. Casadio, P. Fariselli, C. Sander, Transmembrane helices predicted at 95% accuracy, Protein Sci. 4 (1995) 521–533.
[10] B. Rost, C. Sander, Bridging the protein sequence-structure gap by structure predictions, Annu. Rev. Biophys. Biomol. Struct. 25 (1996) 113–136.
[11] H.H. Gan, R.A. Perlow, S. Roy, J. Ko, M. Wu, J. Huang, S. Yan, A. Nicoletta, J. Vafai, D. Sun, L. Wang, J.E. Noah, S. Pasquali, T. Schlick, Analysis of protein sequence/structure similarity relationships, Biophys. J. 83 (2002) 2781–2791.
[12] M.A. Marra, S.J. Jones, C.R. Astell, R.A. Holt, A. Brooks-Wilson, Y.S. Butterfield, J. Khattra, J.K. Asano, S.A. Barber, S.Y. Chan, A. Cloutier, S.M. Coughlin, D. Freeman, N. Girn, O.L. Griffith, S.R. Leach, M. Mayo, H. McDonald, S.B. Montgomery, P.K. Pandoh, A.S. Petrescu, A.G. Robertson, J.E. Schein, A. Siddiqui, D.E. Smailus, J.M. Stott, G.S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T.F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G.A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R.C. Brunham, M. Krajden, M. Petric, D.M. Skowronski, C. Upton, R.L. Roper, The genome sequence of the SARS-associated Coronavirus, Science 300 (2003) 1399–1404.
[13] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The Swiss-Prot protein knowledgebase and its supplement tremble in 2003, Nucleic Acids Res. 31 (2003) 365–370.
[14] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[15] J.H. Buchanan, A cystine-rich protein fraction from oxidized alpha-keratin, Biochem. J. 167 (1977) 489–491.
[16] H. Zhang, G. Serrero, Inhibition of tumorigenicity of the teratoma Pc cell line by transfection with antisense Cdna for Pc cell-derived growth factor (Pcdgf, epithelin/granulin precursor), Proc. Natl. Acad. Sci. USA 95 (1998) 14202–14207.
[17] C. Combadiere, S.K. Ahuja, P.M. Murphy, Cloning and functional expression of a human eosinophil Cc chemokine receptor, J. Biol. Chem. 271 (1996) 11034.
[18] S.M. Strittmatter, M. Igarashi, M.C. Fishman, Gap-43 amino terminal peptides modulate growth cone morphology and neurite outgrowth, J. Neurosci. 14 (1994) 5503–5513.
[19] Y. Liu, D.A. Fisher, D.R. Storm, Intracellular sorting of neuromodulin (Gap-43) mutants modified in the membrane targeting domain, J. Neurosci. 14 (1994) 5807–5817.
[20] E. Reinhard, E. Nedivi, J. Wegner, J.H. Skene, M. Westerfield, Neural selective activation and temporal regulation of a mammalian Gap-43 promoter in zebrafish, Development 120 (1994) 1767–1775.
[21] V.A. Neel, M.W. Young, Igloo, a Gap-43-related gene expressed in the developing nervous system of drosophila, Development 120 (1994) 2235–2243.
[22] J.P. Kapfhammer, M.E. Schwab, Increased expression of the growth-associated protein Gap-43 in the myelin-free rat spinal cord, Eur. J. Neurosci. 6 (1994) 403–411.
[23] C. Drosten, S. Gunther, W. Preiser, S. van der Werf, H.R. Brodt, S. Becker, H. Rabenau, M. Panning, L. Kolesnikova, R.A. Fouchier, A. Berger, A.M. Burguiere, J. Cinatl, M. Eickmann, N. Escriou, K. Grywna, S. Kramme, J.C. Manuguerra, S. Muller, V. Rickerts, M. Sturmer, S. Vieth, H.D. Klenk, A.D. Osterhaus, H. Schmitz, H.W. Doerr, Identification of a novel Coronavirus in patients with severe acute respiratory syndrome, N. Engl. J. Med. 348 (2003) 1967–1976.
[24] T.M. Schmeing, P.B. Moore, T.A. Steitz, Structures of deacylated tRNA mimics bound to the E site of the large ribosomal subunit, RNA 9 (2003) 1345–1352.
[25] M. Selmer, S. Al-Karadaghi, G. Hirokawa, A. Kaji, A. Liljas, Crystal structure of thermotoga maritima ribosome recycling factor: a tRNA mimic, Science 286 (1999) 2349–2352.