Bioinformatics Analysis of Physicochemical Properties of Protein Sorting Signals

Fan Zhang
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC, 29036, USA
zhangf@cse.sc.edu

Jianjun Hu
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC, 29036, USA
jianjunh@cse.sc.edu

Abstract: Subcellular localization of proteins is usually guided by their sorting signals encoded by subsequences of amino acids at the N-terminal or C-terminal ends. These signals are usually composed of a set of physichemically conserved amino acid groups such as the hydrophoblic cores of secretory signal peptides. Using experimentally determined sorting signals, biologists have identified the physichemical models of several categories of sorting motifs. Here we proposed a bioinformatics algorithm for de novo identification of physichemical groups enriched within a set of proteins without knowing the sorting signals themselves. Experiments showed that our binominal distribution based enrichment test algorithm have successfully identified the known sorting signal properties of secretory signal peptides, nuclear localization signals, and mitochondrial sorting signals.

1 INTRODUCTION

Protein sorting is the process by which a cell transports proteins to their destination in the cell or out of it. Precise sorting is critical for the cell as mis-targeting will lead to diseases. Protein sorting is well controlled by a variety of sorting signals encoded by the protein sequences either as polypeptide chain (so-called signal peptides) or as features of the folded protein [1,2]. Throughout years, biologists have identified a variety of signal peptides sorting to different organelles such as mitochondria, chloroplast, and peroxisomes using costly and labor-intensive experimental methods. It is found that sorting signals for different locations have different physichemical properties [1]. For example, mitochondrial targeting signals are rich in positively charged amino acids and hydroxylated ones. Sorting signals of secretory proteins are usually composed of a tri-party structure with a positively charged n-region, a hydrophobic h-region, and a polar c-region leading up to the signal peptidase cleavage site. However, there are many other possible unknown signal models and the lack of experimentally identified sorting signals makes it difficult to analyze their physichemical structures of these sorting signals. On the other hand, recent high-throughput experiments have identified the subcellular location of genome-wide proteins. It remains unclear how such localization datasets can be used to help to discover new sorting signals and their physichemical structures.

Here we proposed an enrichment test algorithm to de novo identify physichemical components of sorting signals enriched for a given set of proteins targeting to the same subcellular location. The method is developed based on calculating the significance of a subsequence of amino acids –amino acid groups (AAGs) with a special physichemical property using binominal test and then using a clustering algorithm to merge smaller amino acid groups into larger ones. By detecting the enrichment score of these AAGs for all input protein set compared to the background, we are able to identify over-represented AAGs that may characterize the sorting signals within these proteins. Experiments showed that our algorithm can rediscover the physichemical properties of secretory sorting signals, mitochondrial sorting signals and some novel AAGs that are not known to biologists.

2 METHODS

1.1 Framework of enrichment analysis of physichemical AAGs

Protein sorting motifs are usually composed of a set of well-conserved physichemical amino acid groups with or without some highly conserved amino acids. These AAGs can have different lengths and can still target to correct location. As these sorting signals are not conserved at the amino acid level, the widely used PWM model for DNA binding sites/motifs is not suitable for sorting motif representation, which cannot model the conservation at the physichemical level. Recently [3], we proposed the AAG concepts to model sorting signals, which group consecutive amino acids with similar physicochemical properties into amino acid groups. Protein sequences can then be represented as AAG sequences using the physicochemical encoding. To identify protein sorting motifs from a given set of protein and a set of background proteins, we first convert protein sequences into AAG sequences and then apply the frequency based enrichment test to AAG sequences to identify the most differentiating/enriched motifs. An overview framework of AAG motif finding algorithm is shown in Fig 1.

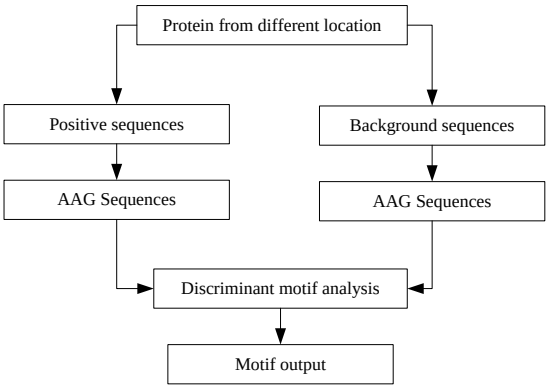

Fig. 1 Framework of AAG based protein sorting motif analysis

1.2 Physicochemical property of amino acids

Functions of proteins usually strongly depend on the physichemical properties of its component amino acids, which can be characterized by their physicochemical properties. 20 amino acids can then be classified into different physicochemical property classes [4] including Hydrophobic (H), Polar (P), Charged (C), Positive Charged (O), negative charged (N), Aliphatic (A), Aromatic (R), Tiny (T), and Small (S). Each class includes a set of amino acids as shown in Fig 2. For example, hydrophobic amino acid class includes [ACGTIVLMFYWHK], and charged amino acid class includes [DEKHR].

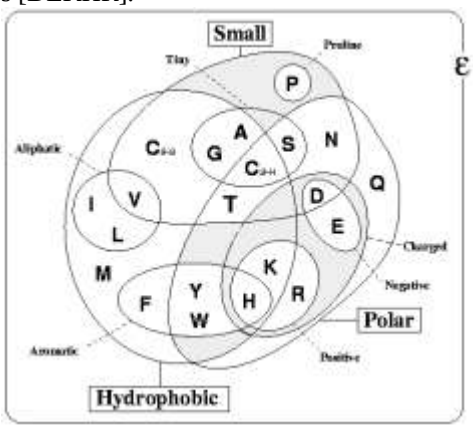

Fig.2. Venn diagram of physicochemical properties of amino acids

1.3 Physicochemical encoding of Protein sequence:

As shown in Fig2, we have totally 9 physicochemical property classes. Given a protein sequence, for each class we scan the sequence from left to right, if the amino acid is in this class, we replace it by "1" otherwise replace it by "0". In this way, we can convert an amino acid sequence into multiple physichemical property sequences (Fig.3)

	M	E	K	I	P	V	S	A	F	L	L	L	V	A	L
[ACGTIVLMFYWHK]	1	0	1	1	0	1	0	1	1	1	1	1	1	1	1
[DEKHR]	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
[YWHKRDEQNSCT]	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 3 Physicochemical encoding protein sequences. Three physicochemical classes are shown here Hydrophobicity (row 2), Charged (row 3) and Polarity (row 4)

After generating physicochemical encodings of protein sequences, we use the following clustering algorithm to identify significant AAGs enriched with the same physicochemical property, as shown in Fig 2.

1.4 Clustering method for AAG identification:

As shown in Fig 4, a physichemical sequence can be easily grouped into small segments with homogenous physichemical property. The question is how to merge those segments to generate meaningful AAGs that tolerate some gaps. We proposed a hierarchical clustering algorithm to solve this problem.

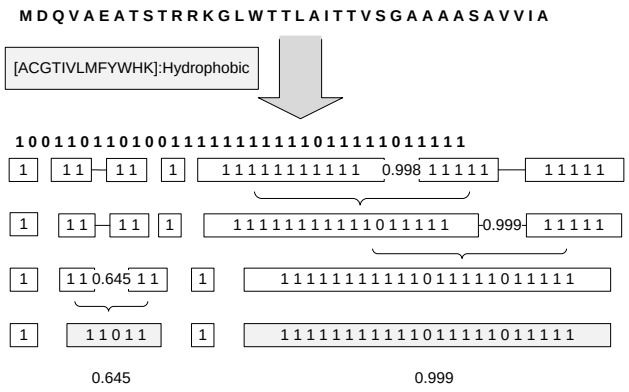

Fig 4 An example of AAG clustering. Firstly, a protein sequence is converted to physicochemical encoding sequence. Basic AAGs are detected. Then in each iteration, the AAG pair with maximum similarity score will be grouped together until the clustering can no longer increase the AAG significance score. The AAGs will then be tested by an enrichment score threshold and only those AAGs with enrichment scores larger than the threshold will be reported.

To apply hierarchical clustering to AAG discovery, we scan the physicochemical encoding sequence from left to right, group consecutive "1"s in the sequence together, and use those all "1"s subsequences as the data points,

because AAG may have gaps and noise in sequence, so the all "1" data points are actually the fragments of significant AAG we want to find, to group those fragments together to form longer and more informative AAG, clustering method is applied to those small groups. The main idea is that each time we group 2 best candidate fragments as defined by a distance function F mapping from a group pair to a real number in [0,1],

Distance function F is calculated by binomial p-value of AAG, binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p, let the number of experiments be N, probability of each experiment yields to be success be p, the probability of the event that totally k experiment yields to be success out of N can be represented as:

$$b(k, N, p) = \Pr\{X = k \mid N, p\} = \binom{k}{N} p^k (1-p)^{(N-k)}$$

So the probability to observe an AAG with length greater than or equal to K having a specific physichemical is equal to:

$$es(s) = 1 - B(k, N, p) = 1 - \Pr\{X < k \mid N, p\}$$

$$= 1 - \sum_{i=0}^{k-1} \binom{i}{N} p^i (1-p)^{(N-i)}$$

Given a AAG sequence, k can be retrieved by counting number of "1" in the sequence, N is the length of the AAG, p can be calculated by prior knowledge: using a large background sequence set and count the number of amino acid with a given physichemical property and divided by the total number of amino acid in the sequence set. We define this as the enrichment score for an AAG. The enrichment score measures how rare the AAG is enriched. The lower the score, the less likely this AAG occurs by chance.

We define the similarity function of two AAGs in a physicochemical encoding sequence by calculating the enrichment score of their merged AAG. If two AAGs are not adjacent in sequence without other AAGs between them, the similarity function of them are set to be undefined, or infinite large, which means the clustering can only happens between adjacent AAGs. If two AAGs are adjacent, the distance function will be the enrichment score of the sequence by concatenating them together. Each time, the clustering algorithm merges the pair of AAGs which are most similar among all neighboring AAGs. The iterative merging continues until the score of merged AAG is less than the score of any of them.

Similarity function: $f(g_1, g_2) = s(g_1 \cup g_2)$ if g1 and g2 are adjacent, else equals to infinity, an example of clustering method is shown in Fig4. When the clustering procedure stops, the remaining AAGs will be further tested by a user specified threshold enrichment score. For example in Fig4, if the threshold is set to 0.6, then only the highlighted AAG in the result will be reported as the significant AAG.

For each protein sequence, we extract all AAGs for different physicochemical properties. We then put these AAGs together and sort them by the starting position in the sequence and represent them using the a single character symbol such as H for hydrophobicity. This will convert the protein sequence to a AAG sequence,. Each element of AAG sequence represents an significant AAG. The order of the elements represents the order of the AAGs in protein sequence as shown in Fig5.

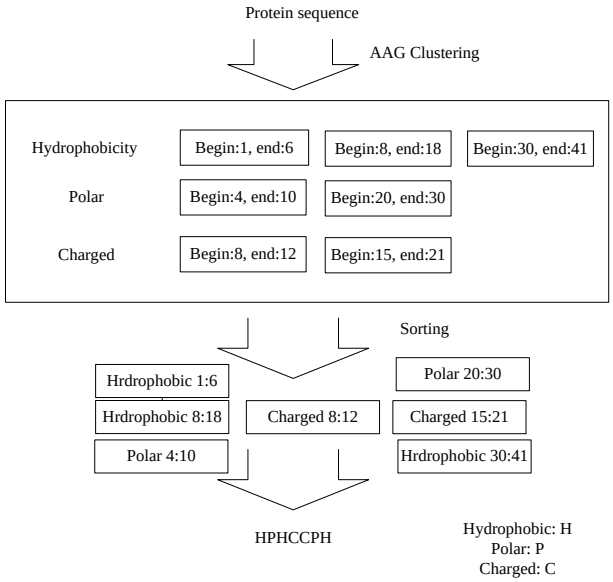

Figure 4: Converting a protein sequence into AAG sequence. Proteins are firstly converted to AAGs by the clustering algorithm. And then all AAGs from different property classes will be put together and sorted by their beginning position in protein sequence. Finally, the sorted AAGs in list will be represented by a single letter (in this example are "H","P" and "C") and form the final AAG sequence, in figure is "HPHCCPH"

1.5 Identifying discriminating AAG motifs

To identify physichemical AAGs enriched for a location, we use the frequent anchor analysis algorithm [site] to find the discriminating motifs in the AAG sequences generated from protein sequences. Frequent anchor analysis works like this: given a set of AAG sequences as the positive set, a set of AAG sequences as the background, and a motif length N, the algorithm enumerates all possible combinations of N-length combination of AAGs and find the most significant combination in the positive set vs. the background. Significant score of an AAG combination is defined as follows:

Let positive sequence set be P[1..Np], background sequence set be N[1…Nn], for a given k-AAG

combination C[1..k], let supP(C) be the support of C in P (the number of sequences containing C in P), sup_rP(C) be the support ration of C in P (equals supP(C) divided by the number of sequences in P), sup_rN(C) be the support of C in N (the number of sequence in N containing C divided by number of sequences in N), the binomial p-value of the motif can be calculated by:

$$B(\sup_P(C), N_P, \sup\_r_N(C)) = \Pr\{X < \sup_P(C) \mid N_P, \sup\_r_N(C)\} = \sum_{i=0}^{\sup_P(C)-1} \binom{i}{N_P} \sup\_r_N{}^{i} (1-\sup\_r_N)^{(N_P-i)}$$

For all combination of AAG in given length, we calculate the p-value of them and sort them in descendent order, the AAG combination on the top of the list are the most differentiating motifs between positive sequence set and background. Also the confidence score are output with the top-ranking motifs, providing a quantified measurement for the discovered motifs.

3 RESULTS

3.1 Datasets preparation

To evaluate the capability of our algorithm to identify physichemical motifs for sorting signals, we collected the following datasets from known sources (1) Secreted signal peptide from SPdb (Signal Peptide Database) [5], (2) Mammalian secreted protein sequences from LOCATE database [6], (3) Bacterial secreted proteins from SignalP [7] experiment dataset, (4) Nuclear localization signal motif and nuclear localized proteins from NLSdb Database [8], (5) chloroplast proteins from BaCello dataset [9], and we use cytoplasmic proteins from LOCATE database as background for AAG algorithm, detailed information of experimental data are included in Table1.

For all these sequences we extracted a set of protein sequences targeting to a specific location and then we extracted the N-terminal and C-terminal 50 amino acids and run our motif detection algorithm and then we report the single AAG and two AAG combination physichemical motifs

3.2 De novo identification of physichemical AAGs for secretory sorting signals

We applied our physichemical AAG motif detection algorithm to the N-terminal 50 amino acids of all secretory proteins in the SPDb dataset. Table 2 shows the top two single-AAG and double-AAG motifs and their enrichment scores. It is found that the widely known

Table 1. Protein sequence datasets for experiments

#	Positive dataset	Background dataset	Num of sequences in positive/ background dataset
1	Secreted signal peptide form SPDB	Cytoclasmic proteins from LOCATE database	158/2000
2	SignalP training dataset (secreted)	SignalP training dataset (cytoplasmic)	168/150
3	Mammalian secreted proteins from LOCATE database	Cytoclasmic proteins from LOCATE database	2025/2000
4	Nuclear translocation signal from NLS database	Cytoclasmic proteins from LOCATE database	4142/2000

hydrophobic AAG motif is highly enriched among these secretory motifs with a frequency of 0.97. This hydrophobic core is necessary for signal peptides as suggested by previous biological experiments. We also found an interesting aliphatic AAG motif, A(7,7) with average position at 7th amino acid from the N-terminal and average length of 7. This motif is embedded within the hydrophobic AAG motif. The algorithm also identified the typical tiny amino acids T(15,6) occurring after the hydrophobic cores of signal peptides.

Physichemical Properties	Abbreviation
Hydrophobic	H
Charged	C
Polar	P
Aliphatic	A
Aromatic	R
Small	S
Tiny	T
Proline	L
Positive Charged	O
Negative Charged	N

Figure 5 Amino acid physiochemical properties and corresponding abbreviations

For simplicity, we use abbreviation of AAG

physiochemical properties in following tables, full names of the physichemical properties are listed in Fig.5

Table 2. Enriched AAGs in secretory proteins in SPdb database. H(5,13) means hydrophobic AAG motif with average position of 5 amino acids from N-termional and average length of 13.

Top AAG	AAG frequency in positive dataset	AAG frequency ratio in background	Enrichment score or p value
H (5,13)	0.97	0.47	$<10^{-10}$
A (7,7)	0.98	0.76	$<10^{-10}$
H(4,13)A(7,7)	0.66	0.20	$<10^{-10}$
A(9,7) T(15,6)	0.22	0.07	$<10^{-10}$

According to previous studies, SPs generally consist of three regions: a positively charged n-region, a hydrophobic h-region, and a polar c-region leading up to the signal peptidase cleavage site. However, our algorithm does not find the n-region and the c-region motifs. We ran another test on the secretory proteins from signalP website and tested the physichemical AAG motifs for the first 100 amino acids rather than 50 amino acids. The result is shown in Table 3. It again identified the hydrophobic AAG and the Tiny/small AAGs. Especially it also identified the OH motif which is positive charge and hydrophobic AAGs corresponding to the canonical model of signal peptides. It also identified the P(34,12) motif which usually appears after the hydrophobic AAG as shown by their average distance from the N-terminal. This means that our algorithm has successfully identified all the major physichemical features of secretory signal peptides without knowing the exact sorting signals.

Table 3. Over-represented physichemical AAGs of secretory proteins within N-terminal 50 and 100 amino acids. All AAG has a significance score/p-value of less than 10^{-10}

N50		N100	
AAG	Freq diff	AAG	Freq diff
H(8,16)	0.33	T(27,12)	0.39
T(17,9)	0.31	P(34,12)	0.23
S(21,9)	0.19	S(25,11)	0.19
O(5,4)H(9,16)	0.29	S(41,18)T(42,15)	0.51
S(20,10)T(22,8)	0.22	P(58,23)S(60,24)	0.50
A(16,6)S(24,10)	0.21	T(46,15)P(52,17)	0.33

We have also applied the physichemical motif enrichment test to the mammal LOCATE database for secretory proteins and also identified the hydrophobic AAGs and Charged AAGs. A new aromatic AAG was also found to be ranked high in the result list which may have special functions.

3.2 De novo identification of physichemical AAGs for nuclear localization signals

We collected all the nuclear targeting proteins related to the nuclear localization signals in the NLSdb. Table 4 shows that N-terminal of nuclear proteins are enriched with Polar, Charged and Positive charged AAGs, which however are not as significant as AAGs of secretory signals. This is because most of the sorting signals for nuclear proteins are located at the C-terminal. Table 5 shows that C terminal of nuclear localized proteins are enriched with Polar, positive Charged and Charged AAGs, which matches the classical model of nuclear localization signals. We also found that the C-terminal of nuclear localization signals is also enriched with hydrophobic AAGs.

Table 4. Over-represented physichemical AAGs of nuclear proteins within N-terminal 50 and 100 amino acids. All AAG has a significance score/p-value of less than 10^{-10}

N50		N100	
AAG	p-value	AAG	p-value
R(20,3)	0.00	R(34,3)	0.03
H(14,12)	0.03	H(27,12)	0.52
O(15,5)	0.11	P(21,8)	0.69
C(18,9)P(18,9)	$<10^{-10}$	R(48,3)P(53,8)	$<10^{-10}$
C(18,7)O(19,5)	$<10^{-10}$	A(30,7)H(34,12)	$<10^{-10}$
A(14,8)H(20,12)	$<10^{-10}$	O(39,5)R(43,3)	$<10^{-10}$

Table 5. Over-represented physichemical AAGs of nuclear proteins within C-terminal 50 and 100 amino acids. All AAG has a significance score/p-value of less than 10^{-10}

C50		C100	
AAG	p-value	AAG	p-value
O(13,6)	$<10^{-10}$	C(17,10)	$<10^{-10}$
R(20,3)	$<10^{-10}$	O(17,6)	$<10^{-10}$
H(13,12)	$<10^{-10}$	H(26,12)	$<10^{-10}$
C(19,9)O(19,7)	$<10^{-10}$	O(32,6)C(36,9)	$<10^{-10}$
C(13,11)P(13,12)	$<10^{-10}$	C(36,9)O(37,6)	$<10^{-10}$
H(15,11)C(19,12)	$<10^{-10}$	O(34,7)P(38,10)	$<10^{-10}$

4 DISCUSSIONS

Experimentally identifying and dissecting protein sorting signals are costly and labor-intensive. Current DNA motif algorithms such as AlignACE [10] and MEME [11] cannot be effectively used to identify such physichemically conserved amino acid motifs which are not conserved at amino acid sequence level. Here we proposed a bioinformatics algorithm using enrichment test to de novo identify the physichemical build block- amino acid groups with shared physichemical properties using the protein amino acid sequences only that targets to the same location. The enrichment analysis has successfully identified the variety known physichemical structures of known motifs such as the tri-part structure of secretory signal peptides as well as the hydrophobic and positive charged amino acid groups of the nuclear localization signals. With such identified physichemical AAGs, we will then be able to develop more precise prediction algorithms for both localization signals as well as protein subcellular localization prediction.

5 ACKNOWLEDGEMENTS

This work is supported by NSF Career Award under grant DBI-0845381. Acknowledgment is also made to the University of South Carolina's High Performance Computing Group for the computing time used in this research.

References

1. Emanuelsson O: **Predicting protein subcellular localisation from amino acid sequence information.** *Brief Bioinform* 2002, **3:** 361-376.

2. Emanuelsson O, von Heijne G: **Prediction of organellar targeting signals.** *Biochimica et Biophysica Acta-Molecular Cell Research* 2001, **1541:** 114-119.

3. J.Hu, F.Zhang: **Improving Protein Localization Prediction Using Amino Acid Group Based Physichemical Encoding.** In: *Lecture Notes in Bioinformatics*.

4. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M: **AAindex: amino acid index database, progress report 2008.** *Nucleic Acids Research* 2008, **36:** D202-D205.

5. Choo KH, Tan TW, Ranganathan S: **SPdb - a signal peptide database.** *Bmc Bioinformatics* 2005, **6**.

6. Fink JL, Aturaliya RN, Davis MJ, Zhang FS, Hanson K, Teasdale MS *et al*.: **LOCATE: a mouse protein subcellular localization database.** *Nucleic Acids Research* 2006, **34:** D213-D217.

7. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *Journal of Molecular Biology* 2004, **340:** 783-795.

8. Nair R, Carter P, Rost B: **NLSdb: database of nuclear localization signals.** *Nucleic Acids Res* 2003, **31:** 397-399.

9. Pierleoni A, Martelli PL, Fariselli P, Casadio R: **BaCelLo: a balanced subcellular localization predictor.** *Bioinformatics* 2006, **22:** E408-E416.

10. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16:** 939-945.

11. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:** 28-36.