On Reconstructing a String from its Substring Compositions

Jayadev Acharya
ECE, UCSD
jacharya@ucsd.edu

Hirakendu Das
ECE, UCSD
hdas@ucsd.edu

Olgica Milenkovic
EE, UIUC
milenkov@uiuc.edu

Alon Orlitsky
ECE & CSE, UCSD
alon@ucsd.edu

Shengjun Pan
CSE, UCSD
s1pan@ucsd.edu

***Abstract*—Motivated by protein sequencing, we consider the problem of reconstructing a string from the compositions of its substrings. We provide several results, including the following. General classes of strings that cannot be distinguished from their substring compositions. An almost complete characterization of the lengths for which reconstruction is possible. Bounds on the number of strings with the same substring compositions in terms of the number of divisors of the string length plus one. A relation to the turnpike problem and a bivariate polynomial formulation of string reconstruction.**

I. Introduction

A protein is a long sequence of amino acids whose composition and order determine the protein's properties and functionality. A common tool for finding the amino-acid sequence is Mass Spectrometry [1, 2]. It takes a large number of identical proteins and passes them through an ion source which randomly breaks them into substrings. The resultant mixture is then analyzed by a mass spectrometer that gives the mass to charge ratio. This in conjunction with other methods yield the weights of the substrings generated. The weight information is then used to infer the amino-acid sequence.

In this paper we make two simplifying assumptions that reduce protein reconstruction to a combinatorial problem, which we analyze. The *composition* of a string is the multiset of its elements, namely the number of times each of its elements appears, regardless of the order. For example, the composition of the sequence $BABCAA$ is the multiset $\{A, A, A, B, B, C\}$, which we often denote A^3B^2C, indicating that the sequence has three A's, two B's, and one C.

Assumption 1: The composition of every protein substring can be deduced from its weight.

For example, let A, B, and C be three amino acids with respective weights 13, 7, and 4. A sequence of weight 11 clearly consists of one B and one C. Similarly, a weight of 18 implies two B's and one C in any order possible. However, a weight of 20 could arise from one A and one B or from 5 C's, in which case we are not able to deduce the composition from weights. The assumption is that such confusions never arise. While clearly idealized, the assumption could be valid if all amino-acid weights are sufficiently large and different.

Assumption 2: If the protein sequence is of length n, then all $\binom{n+1}{2}$ substrings appear roughly equal number of times in the mixture.

For example, the string AAB has $6 = \binom{3+1}{2}$ substrings: A, A, B, AA, AB, and AAB. The assumption implies that in the mixture, each substring will appear the same number of times, hence for example A will occur twice as many times as B, AB, AA and AAB.

Under the first assumption, mass-spectrometry protein sequencing reduces to reconstructing a sequence from a collection of its substring compositions where each substring composition is given an unknown number of times. With the second assumption, the problem is further reduced to reconstructing a sequence from the collection of its substring composition, each given exactly once. We study the extent to which that can be done.

II. Definitions

The *composition multiset* of a sequence $\overline{s} = s_1 s_2 \dots s_n$ is the multiset

$$\mathcal{S}_{\overline{s}} \stackrel{\text{def}}{=} \{\{s_i s_{i+1} \dots s_j\} : 1 \le i \le j \le n\}$$

of compositions of all $\binom{n+1}{2}$ substrings of $\overline{s}$. For example,

$$S_{AAB} = \{A, A, B, A^2, AB, A^2B\},$$
$$S_{ABA} = \{A, A, B, AB, AB, A^2B\}.$$

Note that each composition is given with the number of times it appears. We study the extent to which a sequence can be reconstructed from its decomposition multiset.

Two sequences $\overline{s}$ and $\overline{t}$ are *confusable*, denoted $\overline{s} \sim \overline{t}$, if they have the same decomposition multisets.

The *reversal* of a sequence $\overline{s} = s_1 s_2 \dots s_n$ is the sequence $\overline{s}' \stackrel{\text{def}}{=} s_n s_{n-1} \dots s_1$. A sequence and its reversal clearly have the same composition multiset, and are hence confusable. We call this a *trivial confusion*, and are interested in non-trivial confusions.

Protein sequences are over alphabet of size 20, corresponding to the number of amino acids. Yet reconstruction of binary sequences can be extended to reconstruct sequences over any finite alphabet. This can be done by substituting a subset of symbols with a 1 and others with 0. We then use the binary reconstruction algorithm to get the position of 1's. The process can be repeated to get positions of all symbols one after another. We therefore consider only the problem of reconstructing binary sequences from their composition multiset.

Let

$$\mathcal{C}_{\overline{s}} = \{\overline{t} : \mathcal{S}_{\overline{s}} = \mathcal{S}_{\overline{t}}\}$$

be the set of all sequences which are confusable with $\overline{s}$, and let

$$H_n = \max_{\overline{s} \in \{0,1\}^n} |\mathcal{C}_{\overline{s}}|$$

be the largest number of mutually confusable n-bit sequences. From the above, $H_n \geq 2$ for every $n \geq 2$, and we would like to see when it is strictly larger.

III. Results

We provide a general construction of non-trivially confusable sequences. While we don't know if this construction covers all confusable sequences, we show that it comes close to determining which sequence lengths n have non-trivial confusions.

Specifically, the construction shows that if $n+1$ is a product of two integers, each ≥ 3, then $H_n > 2$, namely there are non-trivial confusions. Conversely, if $n+1$ is not a product of two integers ≥ 3, then it is either 8, a prime or twice a prime. We show that if $n+1$ is a prime then $H_n = 2$, namely there are no non-trivial confusions, and that if $n+1$ is twice a prime then $H_n \leq 4$, namely any sequence is non-trivially confused with at most one sequence and its reversal.

We also provide general bounds on H_n. We show that H_n can be arbitrarily large as n increases by proving that $H_{p^k} = 2^k$ for p prime ≥ 3. We also prove that for all n, $H_n \leq 2^{d(n+1)-1}$, where $d(n)$ is the number of divisors of n.

n	Confusable Sequences	Construction
8	{01001101, 01101001}	01∘01
11	{00100011001, 00110010001}	001∘01
	{00100011001, 00110010001}	01∘001
14	{01001001001101, 01101001001001}	01∘0001
	{01001001101001, 01001101001001}	01∘0010
	{01001001101101, 01101101001001}	01∘0011
	{01001101001101, 01101001101001}	01∘0101
	{01010010110101, 01011010100101}	0101∘01
	{00110001110011, 00111001100011}	0011∘01
	{00010000110001, 00011000100001}	0001∘01

TABLE I
CONFUSABLE SEQUENCES OF LENGTH $\leq$14

To derive some of these results, we relate the sequence reconstruction problem to the *turnpike problem* and its formulation as polynomial factorization [3, 4]. We show that sequence reconstruction can be formulated as a bivariate polynomial factorization problem, and use some turnpike-problem results as well as some new ones specific to sequence reconstruction to derive the bounds mentioned above. [5] looks at the computational aspect of the problem and mention an algorithm for reconstruction of sequences from their multisets.

IV. Confusable strings

Table I outlines all confusable sequences of length at most 14. The *complement* of a sequence is obtained by complementing each of its bits. If two sequences are confusable, so are their complements. For parsimony, the table therefore lists only sequences starting with a 0, and omits reversals.

Note that for length $n \leq 7$ there are only trivial confusions, and the shortest non-trivially confusable sequences are the 8-bit sequences 01001101 and 01101001. Then again there are no non-trivial confusions until length 11, etc. All of these are explained by the following constructions.

The confusable 8-bit sequences in the table can be represented as **01** 0 **01** 1 **01** and **01** 1 **01** 0 **01**. These sequences are formed by interleaving 01 with the bits 01 and with the bits 10 respectively. This patterns extends to yield other confusable sequences.

The *interleaving* of sequence $\overline{s}$ with the bits of a sequence $\overline{t} = t_1 \dots t_m$ is the sequence $\overline{s} \circ \overline{t} \stackrel{\text{def}}{=} \overline{s} t_1 \overline{s} t_2 \dots t_m \overline{s}$. The sequences above are $01 \circ 01$ and $01 \circ 10$. We prove that $\overline{s}_1 \circ \overline{s}_2$ is always confusable with $\overline{s}_1 \circ \overline{s}_2'$, where $\overline{s}_2'$ was defined to be the reversal of $\overline{s}_2$.

A sequence $\overline{s}$ is *factorizable* if it is of the form $\overline{s}_1 \circ \overline{s}_2$ with both $\overline{s}_1$ and $\overline{s}_2$ of length at least 2. Following are some useful properties of $\circ$.

Lemma 1:

1) $(\overline{s}_1 \circ \overline{s}_2)' = \overline{s}'_1 \circ \overline{s}'_2$.
2) '$\circ$' is associative, i.e., $\overline{s}_1 \circ (\overline{s}_2 \circ \overline{s}_3) = (\overline{s}_1 \circ \overline{s}_2) \circ \overline{s}_3 \stackrel{\text{def}}{=} \overline{s}_1 \circ \overline{s}_2 \circ \overline{s}_3$.
3) Every sequence $\overline{s}$ has a unique maximal factorization as $\overline{s} = \overline{s}_1 \circ \overline{s}_2 \circ \cdots \circ \overline{s}_m$. ■

The next theorem characterizes a class of mutually-confusable sequences. The confusable sequences obtained by this method can be related to the equivalence condition for Ribbon Schur functions [6]. Another class of confusable sequences is provided in Corollary 3.

Theorem 2: Let $\overline{s}_1 \circ \overline{s}_2 \circ \cdots \circ \overline{s}_m$ be the maximal factorization of $\overline{s}$. Then $\overline{s}$ is confusable with all sequences of the form $\overline{t}_1 \circ \overline{t}_2 \circ \cdots \circ \overline{t}_m$, where each $\overline{t}_i$ is either $\overline{s}_i$ or $\overline{s}'_i$.

Proof: $\overline{s} = \overline{s}_1 \circ \overline{s}_2 \circ \cdots \circ \overline{s}_m$. Let $\overline{s}^*$ be obtained by replacing $\overline{s}_j$ by $\overline{s}'_j$ in $\overline{s}$ for some j. It suffices to show that $\overline{s} \sim \overline{s}^*$. This is because any configuration of $\overline{t}$ can be reached from $\overline{s}$ by replacing one factor of the factorization with its reversal at each step.

We first show that $\overline{s} = \overline{s}_1 \circ \overline{s}_2 \sim \overline{s}_1 \circ \overline{s}'_2 = \overline{t}$. Let $\overline{s}_1 = s_1^n$ and $\overline{s}_2 = t_1^k$. Consider any substring $\overline{s}_a^b$ of $\overline{s}$. Let $a = p \cdot (n+1) - q$, and $b = r \cdot (n+1) + s$, with $q, s < n+1$ and $a \leq b$. Let $a' = (k+1-r) \cdot (n+1) - q$ and $b' = (k+1-p) \cdot (n+1) + s$. The substrings $\overline{t}_{a'}^{b'}$ and $\overline{s}_a^b$ can be seen to have the same composition. This yields a bijection from the substrings of $\overline{s}$ to the substrings of $\overline{t}$ such that each substring and its image have same composition. This proves that the two have same composition multiset. This and Lemma 1 prove that $\overline{s}^* \sim \overline{s}$.

$$\begin{aligned}\overline{s}^* &= \overline{s}_1 \circ \cdots \overline{s}_{j-1} \circ \overline{s}'_j \circ \overline{s}_{j+1} \circ \cdots \overline{s}_m \\ &\sim (\overline{s}_1 \circ \cdots \overline{s}_{j-1})' \circ \overline{s}'_j \circ \overline{s}_{j+1} \circ \cdots \circ \overline{s}_m \\ &\sim (\overline{s}_1 \circ \cdots \overline{s}_{j-1})' \circ \overline{s}'_j \circ (\overline{s}_{j+1} \circ \cdots \circ \overline{s}_m)' \\ &= \overline{s}'_1 \circ \cdots \circ \overline{s}'_m = \overline{s}' \sim \overline{s}. \end{aligned}$$ ■

All confusable sequences of length less than 23 have the form described in the above theorem. The 23-bit sequences 01000101010000100011001 and 01010100010000110010001 are confusable but not of the above form. Rather, they have the more general structure described in the next corollary.

Corollary 3: Let $\overline{s}_1, \ldots, \overline{s}_m$ be sequences of the same type (hence same length). Let $\overline{s}_0$ be any string and $s_1 s_2 \ldots s_{m-1}$ any sequence. Then

$$\begin{aligned}&(\overline{s}_1 \circ \overline{s}_0) s_1 (\overline{s}_2 \circ \overline{s}_0) s_2 \ldots s_{m-1} (\overline{s}_m \circ \overline{s}_0) \sim \\ &(\overline{s}_1 \circ \overline{s}'_0) s_1 (\overline{s}_2 \circ \overline{s}'_0) s_2 \ldots s_{m-1} (\overline{s}_m \circ \overline{s}'_0)\end{aligned}$$

Note that if the theorems above provide complete characterization of confusable sequences then sequences which are confusable with sequences other than their reversals, must have length one less than product of two numbers larger than 3. In particular, n+1 cannot be a prime.

V. Polynomial Formulation

$\overline{s} = s_1 s_2 \cdots s_n$ is a binary sequence. Consider the following representation of the sequence as a bivariate polynomial.

$$P(\overline{s}) = P(x, y) = \sum_{i=0}^{n} x^{i - w_i} y^{w_i},$$

where $w_i = s_1 + s_2 + \cdots + s_i$ is the weight (number of one's) of the first i bits of $\overline{s}$. For example

$$P(1011) = 1 + y + xy + xy^2 + xy^3.$$

This is called *generating polynomial* of the sequence. We can represent a sequence with its generating polynomial and vice versa.

A polynomial is generating polynomial of length n if and only if 1) tt has $n+1$ terms, exactly one each of degree 0, 1, 2, …, n, 2) each term has coefficient 1 and 3) the ratio of the term of degree $i+1$ to the term of degree i is either x or y for i = 0, 1, …, $n-1$.

Let $\mathcal{S}_{\overline{s}}(x, y)$ be the polynomial obtained by first replacing in $\mathcal{S}_{\overline{s}}$ each 0 and 1 with x and y respectively and then summing up all the terms.

For $P(\overline{s}) = P(x, y)$ the following is easy to see

$$P(x, y) P(\frac{1}{x}, \frac{1}{y}) = n + 1 + \mathcal{S}_{\overline{s}}(x, y) + \mathcal{S}_{\overline{s}}^{-1}(\frac{1}{x}, \frac{1}{y}).$$

Sequence reconstruction problem is thus equivalent to the following:

$$\text{Given } P(x, y) P(\frac{1}{x}, \frac{1}{y}) \text{ find } P(x, y).$$

Two sequences $P(\overline{s}) = P(x, y)$ and $P(\overline{t}) = Q(x, y)$ have identical composition multiset if and only if

$$P(x, y) P(\frac{1}{x}, \frac{1}{y}) = Q(x, y)\, Q(\frac{1}{x}, \frac{1}{y}). \qquad (1)$$

A similar formulation was provided for the general turnpike problem in [3, 4]. Lemmas 4,5 are from these sources. We follow their arguments and prove some results for the bivariate polynomial formulation of the sequence reconstruction problem we defined above. Using them we prove some of the results for the reconstruction of sequences from their composition multisets.

If $P(x,y)$ is $s_1 s_2 \dots s_n$, $P(\frac{1}{x},\frac{1}{y})$ (normalized) is $s_n s_{n-1} \dots s_1$. A polynomial $P(x,y)$ is called *reciprocal* if there exist $\nu, \mu \in \mathbb{Z}$, such that

$$P(x,y) = x^{\mu} y^{\nu} P(\frac{1}{x},\frac{1}{y}).$$

A polynomial which is not reciprocal is called non-reciprocal.

Lemma 4: If $P(x,y)$, and $Q(x,y)$ are two sequences which are not reversals and have same composition multisets, then there are nonreciprocal polynomials $A, B \in \mathbb{Z}[x,y]$, and integers μ, ν satisfying

$$P(x,y) = A(x,y)B(x,y)$$
$$Q(x,y) = x^{\mu} y^{\nu} A(x,y) B(\frac{1}{x},\frac{1}{y}). \quad \blacksquare$$

A polynomial whose all coefficients are 0 or 1 is called a *0-1 polynomial*. Every generating polynomial is 0-1.

Lemma 5: Let $P(x,y)$ be a 0-1 polynomial. Any $Q(x,y)$ satisfying

$$P(x,y)P(\frac{1}{x},\frac{1}{y}) = Q(x,y)Q(\frac{1}{x},\frac{1}{y})$$

is also a 0-1 polynomial or negative of a 0-1 polynomial. $\blacksquare$

The proofs of these two lemmas can be derived from [3, 4] with slight modification and are omitted here.

Lemma 6: If $P(x,y)$ is a generating polynomial and $Q(x,y)$ satisfies

$$P(x,y)P(\frac{1}{x},\frac{1}{y}) = Q(x,y)Q(\frac{1}{x},\frac{1}{y})$$

then $Q(x,y)$ is a generating polynomial or negative of one.

Proof: It suffices to show that $Q(x,y)$ satisfies the conditions of generating polynomial mentioned before.

By Lemma 5, $Q(x,y)$ can be assumed to be 0-1 and hence satisfies Property 2.

We first show that there is exactly one term of each degree in $Q(x,y)$. It is easy to see that the number of terms in $Q(x,y)$ is exactly $n+1$. The largest degree term is unique, has degree n and coefficient 1. Hence, either there is exactly one term of each degree in $Q(x,y)$ or there are two distinct terms of some degree. All 0 degree terms in $P(x,y)P(\frac{1}{x},\frac{1}{y})$ are obtained by dividing terms of same degree with each other. When $P(x,y)$ is generating polynomial then there is only one term of each degree and thus each such division yields 1 and the constant term in $P(x,y)P(\frac{1}{x},\frac{1}{y})$ is $n+1$. By the condition provided in the lemma, we see that there cannot be more than one distinct term of any degree. If not, we get a term of total degree 0 which is not a constant in $Q(x,y)Q(\frac{1}{x},\frac{1}{y})$. This proves property 1 of generating polynomial.

To prove property 3., consider all the terms of degree 1 in $Q(x,y)Q(\frac{1}{x},\frac{1}{y})$. There are n such terms and each is obtained by dividing a term of degree $i+1$ by a term of degree i. By condition given in the lemma, each such term is equal to either x or y since they correspond to the number of 0's and 1's in the sequence corresponding to $P(x,y)$. $\blacksquare$

The above lemmas imply:

Theorem 7: For a sequence $\overline{s} = P(x,y)$, let

$$P(x,y) = P_0(x,y)P_1(x,y)P_2(x,y)\dots P_k(x,y),$$

where $P_i(x,y) \in \mathbb{Z}[x,y]$, P_0 is reciprocal and $P_1, \dots, P_k$ are non-reciprocal and irreducible, then there are exactly 2^k sequences in $\mathcal{C}_{\overline{s}}$. $\blacksquare$

Theorem 8:

$$H_n \leq \min\{2^{d(n+1)-1}, (n+1)^{1.23}\},$$

where $d(n)$ is the number of divisors of n.

Proof: The argument $(n+1)^{1.23}$ of the min function in the theorem follows from [4]. Replacing y with x^{n+1} in $P(x,y)$ we obtain $P(x)$. We can reprove all the theorems above for this $P(x)$. It then follows from [4] that

$$H_n \leq (n+1)^{1.232}$$

Substituting $y = x$ in $P(x,y)$ we obtain,

$$P(x,x) = 1 + x + x^2 + x^3 + \dots + x^n$$
$$\Rightarrow P(x,x)(x-1) = x^{n+1} - 1.$$

The following [7, pp. 197] factorizes $x^n - 1$.

$$x^n - 1 = \prod_{d|n} \Phi_d((x),$$

where $\Phi_d(x)$ is the dth cyclotomic polynomial (leading coefficient 1 and degree $\phi(d)$ whose roots are the dth primitive roots of unity). $\phi(d)$ is the euler totient function giving the number of positive integers less than d and relatively prime to it. Since cyclotomic polynomials are irreducible, the number of factors of $x^n - 1$ is $d(n)$. Substituting $y = x$ in $P(x,y)$ does not change its degree, and we can conclude that the number of non-reciprocal factors of $P(x,y)$ is at most $d(n+1) - 1$. The size of a confusable class is upper bounded by $2^{d(n+1)-1}$. $\blacksquare$

When $n+1$ is a prime power, the bound in Theorem 8 is tight.

Corollary 9: If $n+1 = p^k$ for a prime $p \geq 3$, then

$$H_n = 2^k.$$

Proof: Theorem 8 shows that $H_n \leq 2^k$. To prove that there exists a sequence with $|\mathcal{C}_{\overline{s}}| = 2^k$, choose any non-reciprocal sequence of length $p-1$. Consider the set of 2^k sequences obtained by interleaving the sequence or its reciprocal with itself or its reciprocal k times. Each such sequence has the same composition multiset. For example, when $n = 26$, $n+1 = 3^3$. Consider, $\mathcal{C} = \{\overline{s} : \overline{s} = \overline{s}_1 \circ \overline{s}_2 \circ \overline{s}_3$, with $\overline{s}_i = 01$ or $10\}$. It consists of 2^3 confusable sequences. ■

In particular, for length prime minus 1, there are no non-trivial confusions, which also has a simple proof.

Corollary 10: If $n+1$ is prime,

$$H_n = 2.$$

Proof: Since $n+1$ is a prime

$$P(x,x) = 1 + x + x^2 + \ldots + x^n$$

is irreducible over $\mathbb{Z}$, because all $(n+1)$th root of unity except '1' are primitive. ■

Corollary 11: If $n+1$ is twice a prime ≥ 3,

$$H_n \leq 4.$$

Proof: Clearly,

$$\begin{aligned} P(x,x) =& 1 + x + x^2 + \ldots + x^{2p-1} \\ =& (1+x)(1 + x + x^2 + \ldots + x^{p-1}) \\ & (1 - x + x^2 - x^3 + \ldots + x^{p-1}) \end{aligned}$$

$P(x,y)$ has at most three factors. If it indeed has three factors then we can show that one of them must be $1+x$ or $1+y$. Both these polynomials are reciprocal and thus $P(x,y)$ has at most two non-reciprocal factors. ■

References

[1] T. E. Creighton, *Proteins: Structures and Molecular Properties*, 2nd ed. W. H. Freeman, 1992.

[2] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, 2nd ed. Cold Spring Harbor Laboratory Press, 2001.

[3] J. Rosenblatt and P. D. Seymour, "The structure of homometric sets," *SIAM Journal on Algebraic and Discrete Methods*, vol. 3, no. 3, pp. 343–350, 1982.

[4] S. S. Skiena, W. D. Smith, and P. Lemke, "Reconstructing sets from interpoint distances (extended abstract)," in *Symposium on Computational Geometry*, 1990, pp. 332–339.

[5] S. Chen, Z. Huang, and S. Kannan, "Reconstructing numbers from pairwise function values," in *International Symposium on Algorithms and Computation*, 2009, pp. 142–152.

[6] L. J. Billera, H. Thomas, and S. van Willigenburg, "Decomposable compositions, symmetric quasisymmetric functions and equality of ribbon Schur functions," *Adv. Math.*, vol. 204, no. 1, pp. 204–240, 2006.

[7] I. Niven, H. S. Zuckerman, and H. L. Montgomery, *An Introduction to the Theory of Numbers*, 5th ed. Wiley Interscience, 1991.