String Reconstruction from Substring Compositions

Olgica Milenkovic

A joint work with J. Acharya, H. Das, A. Orlitsky, and S. Pan (UCSD)

University of Illinois at Urbana-Champaign

KTH-SPE, February 2011

Reconstructing Strings from "Traces": The History of Bioinformatics

- Most (all?) computational challenges in bioinformatics pertain to DNA, RNA, or protein string analysis.

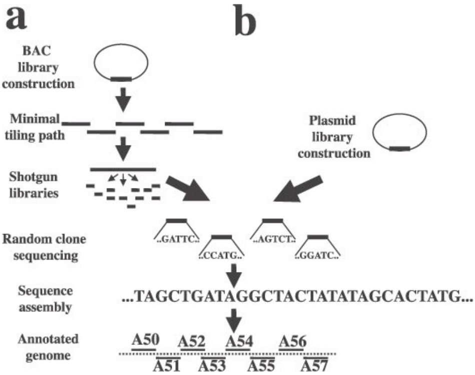
Reconstructing Strings from "Traces": The History of Bioinformatics

- Most (all?) computational challenges in bioinformatics pertain to DNA, RNA, or protein string analysis.
- What are traces? Most frequently, substrings or subsequences of a string.

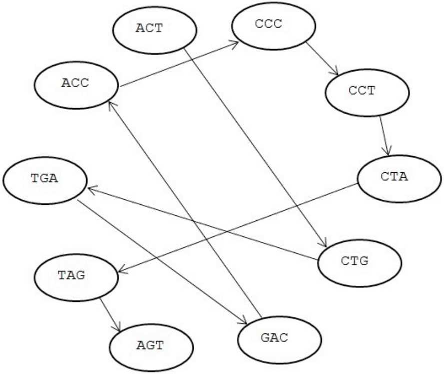
Reconstructing Strings from "Traces": The History of Bioinformatics

- Most (all?) computational challenges in bioinformatics pertain to DNA, RNA, or protein string analysis.
- What are traces? Most frequently, substrings or subsequences of a string.
- Subsequences: obtained by deleting some elements in original sequence, without changing order of elements. Arise due to mutations in genomic codes.
- Substrings: consecutive strings of elements. Mostly arise during sequencing/identification of genomic/proteomic sequences.
- Given: Traces or properties of traces; Task: Reconstructing the sequences(s).

Example I: Shotgun Sequencing (a)



Example I: Shotgun Sequencing (b)



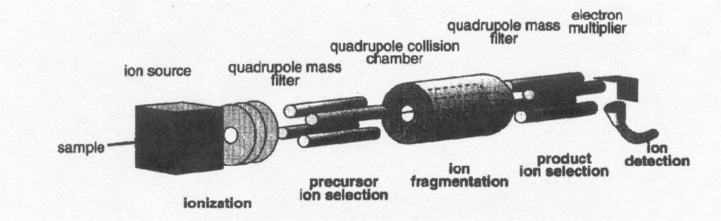
Example II: TMS (a)

- A common tool for sequencing a protein (sequence of amino acids): tandem mass spectrometry.
- A large number of identical proteins broken down into substrings.
- Substrings are ''analyzed'' in order to reconstruct amino-acid sequence.

Example II: TMS (b)

TANDEM Mass Spectrometry

- Tandem mass spectrometry uses two stages of mass analysis, one to preselect an ion and the second to analyze fragments induced by collision with an inert gas like argon or helium.



Triple Quadrupole Mass Spectrometer

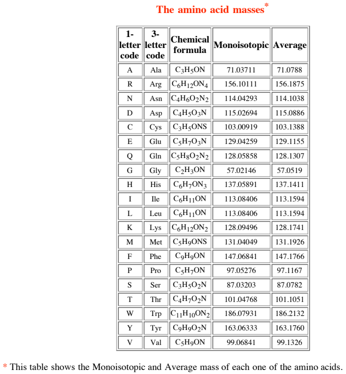
Taken fromHoffman, Edmond, "Tandem Mass Spectrometry: A Primer", Journal of Mass Psectrometry, Vol 31, 129-137, 1996.

TMS: Simplifications?

- De Novo Sequencing: One cannot make use of genomic data in analysis (post-translational modifications very high).
- "Unique mass sums": Composition of protein can be inferred from its weight.
- All bonds may be broken independently and with equal probability.

TMS: Simplifications?

- De Novo Sequencing: One cannot make use of genomic data in analysis (post-translational modifications very high).
- "Unique mass sums": Composition of protein can be inferred from its weight.
- All bonds may be broken independently and with equal probability.



The amino acid masses*

1-letter code	3-letter code	Chemical formula	Monoisotopic	Average
A	Ala	C_3H_5ON	71.03711	71.0788
R	Arg	$C_6H_{12}ON_4$	156.10111	156.1875
N	Asn	$C_4H_6O_2N_2$	114.04293	114.1038
D	Asp	$C_4H_5O_3N$	115.02694	115.0886
C	Cys	C_3H_5ONS	103.00919	103.1388
E	Glu	$C_5H_7O_3N$	129.04259	129.1155
Q	Gln	$C_5H_8O_2N_2$	128.05858	128.1307
G	Gly	C_2H_3ON	57.02146	57.0519
H	His	$C_6H_7ON_3$	137.05891	137.1411
I	Ile	$C_6H_{11}ON$	113.08406	113.1594
L	Leu	$C_6H_{11}ON$	113.08406	113.1594
K	Lys	$C_6H_{12}ON_2$	128.09496	128.1741
M	Met	C_5H_9ONS	131.04049	131.1926
F	Phe	C_9H_9ON	147.06841	147.1766
P	Pro	C_5H_7ON	97.05276	97.1167
S	Ser	$C_3H_5O_2N$	87.03203	87.0782
T	Thr	$C_4H_7O_2N$	101.04768	101.1051
W	Trp	$C_{11}H_{10}ON_2$	186.07931	186.2132
Y	Tyr	$C_9H_9O_2N$	163.06333	163.1760
V	Val	C_5H_9ON	99.06841	99.1326

* This table shows the Monoisotopic and Average mass of each one of the amino acids.

TMS: Simplifications?

- Mass of each fragment is known.
- If mass is known, composition of each fragment is known as well .
- All fragments are used in analysis.
- Example: ILLINOIS gives rise to $I, L, L, I, N, O, I, S, IL, LL, LI, \dots ILL, LLI, \dots, ILLINOIS$ and $\{I, I, I, L, L, N, O, S, IL, IL, L^2, IN, NO, IO, IS, \dots, IL^2, IL^2, \dots I^3L^2NOS\}$.

Outline

- Related work on sequence reconstruction using traces.
- Problem statement.
- Main results.
- The Turnpike problem.
- Polynomial representation of strings.
- Work in progress.

Related Work

- Reconstruction from substrings of given length [Dudik et al, Levenshtein]
- Reconstruction from substrings with random deletions [MacGregor et.al., Viswanathan et.al., Mitzenmacher et.al.]
- The turnpike problem [Patterson, Dix et.al., Skiena et.al.]

Problem Statement

- The composition multiset of a string $s = s_1 s_2 \cdots s_n \in \{0, 1\}^n$ is the multiset

$$\mathcal{S}_s = \{\{s_i, s_{i+1}, \cdots, s_j\} : 1 \leq i \leq j \leq n\}.$$

- Example:

$$\mathcal{S}_{001} = \{0, 0, 1, 0^2, 01, 0^2 1\}$$

Problem Statement

- The composition multiset of a string $s = s_1 s_2 \cdots s_n \in \{0,1\}^n$ is the multiset

$$\mathcal{S}_s = \{\{s_i, s_{i+1}, \cdots, s_j\} : 1 \leq i \leq j \leq n\}.$$

- Example:

$$\mathcal{S}_{001} = \{0, 0, 1, 0^2, 01, 0^2 1\}$$

Problems:

Given $\mathcal{S}_s$, can you reconstruct s uniquely?

- Why binary? Pick one AA, call it "0", and all other AAs "1". Reconstruct location of "0"s. Repeat for AAs denoted by "1".

Problem Statement

- The composition multiset of a string $s = s_1 s_2 \cdots s_n \in \{0,1\}^n$ is the multiset

$$\mathcal{S}_s = \{\{s_i, s_{i+1}, \cdots, s_j\} : 1 \leq i \leq j \leq n\}.$$

- Example:

$$\mathcal{S}_{001} = \{0, 0, 1, 0^2, 01, 0^2 1\}$$

Problem:

Given $\mathcal{S}_s$, can you reconstruct s uniquely?

- Strings s and t are confusable if they have the same composition multiset. Each string is confusable with its *reversal*.
- A string is reconstructible if it is confusable only with its reversal. Set of strings confusable with sequence s: C_s

Main results

Theorem

A string is reconstructable iff its length n satisfies:

- $n \leq 7$.
- $n \geq 8$ and $n+1$ is a prime or twice a prime.

If $n \geq 8$ and $n+1$ is a product of two integers each ≥ 3, the string is confusable.

Theorem

If the prime factorization of $n+1$ is $2^{e_0} p_1^{e_1} \cdots p_k^{e_k}$, then

$$2^{\lfloor \frac{e_0}{2} \rfloor + e_1 + \cdots + e_k} \leq C_n \leq \min\left\{2^{(e_0+1)(e_1+1)\cdots(e_k+1)-1}, (n+1)^{1.23}\right\}$$

where C_n denotes the maximum size of a confusable set for strings of length n.

Main results

Theorem

If the prime factorization of $n+1$ is $2^{e_0} p_1^{e_1} \cdots p_k^{e_k}$, then

$$2^{\lfloor \frac{e_0}{2} \rfloor + e_1 + \cdots + e_k} \leq C_n \leq \min \left\{ 2^{(e_0+1)(e_1+1)\cdots(e_k+1)-1}, (n+1)^{1.23} \right\}$$

where C_n is the maximum size of a confusable set for strings of length n.

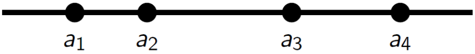
Conjecture: Lower bound is tight.

$$C_{p^m-1} = 2^m, \quad C_{2^m-1} = 2^{\lfloor m/2 \rfloor}.$$

The Turnpike Problem

The Turnpike problem

Given the multiset $\mathcal{D}$ of $\binom{n}{2}$ distances between n points $a_1, \ldots, a_n$, find the relative position of the points on a line.



Example: Exits at 0,1,5,7 produce the multiset $\{1, 2, 4, 5, 6, 7\}$.

The Turnpike problem

- Originated in DNA sequencing and X-ray Crystallography.
- Polynomial time algorithm in largest inter-exit distance known.
- In general, no polynomial time reconstruction algorithm is known.
- Generating polynomial of a point set A:

$$P_A(x) = \sum_{i=1}^{n} x^{a_i}$$

- Generating polynomial of the distance set:

$$D_A(x) = \sum_{i<j} x^{|a_j - a_i|}$$

- $$\begin{aligned} P_A(x)P_A\left(\frac{1}{x}\right) &= n + \sum_{i<j} x^{|a_j-a_i|} + \sum_{i<j} x^{-|a_j-a_i|} \\ &= n + D_A(x) + D_A\left(\frac{1}{x}\right) \end{aligned}$$

The Turnpike problem

- Originated in DNA sequencing and X-ray Crystallography.
- Polynomial time algorithm in largest inter-exit distance known.
- In general, no polynomial time reconstruction algorithm is known.

The Turnpike problem

- $$\begin{aligned} P_A(x)P_A\left(\frac{1}{x}\right) =& n+\sum_{i<j} x^{|a_j-a_i|}+\sum_{i<j} x^{-|a_j-a_i|} \\ =& n+D_A(x)+D_A\left(\frac{1}{x}\right) \end{aligned}$$

  The solution to the turnpike problem is unique iff no two different generating polynomials P_A and P_B satisfy,

  $$P_A(x)P_A\left(\frac{1}{x}\right) \neq P_B(x)P_B\left(\frac{1}{x}\right).$$

Bivariate Generating Polynomials:

Let z_i and w_i be the number of zeros and ones in the first i bit of the string s. The generating polynomial of s is

$$P_s(x,y) = \sum_{i=0}^{n} x^{z_i} y^{w_i}$$

Example:

$$P_{0100} = 1 + x + xy + x^2 y + x^3 y$$

The composition multiset can be written as a composition polynomial:

$$\begin{aligned} \mathcal{S}_{0100} =& \{0,0,0,1,0^2,01,01,0^2 1,0^2 1,0^3 1\} \\ =& 3x + y + x^2 + 2xy + 2x^2 y + x^3 y \end{aligned}$$

Properties of bivariate (composition) generating polynomials $P_s(x, y)$:

- All coefficients equal to one;
- Each monomial degree appears exactly once, and all degrees in $[0, n]$ are present;
- The ratio of two unique monomials of degree $k + 1$ and k is either x or y.

Example:

$$P_{0100} = 1 + x + xy + x^2y + x^3y$$

Generating polynomials

A simple expression for composition multisets

$$P_s(x,y)P_s(\frac{1}{x},\frac{1}{y}) = n+1+\mathcal{S}_s(x,y)+\mathcal{S}_s(\frac{1}{x},\frac{1}{y})$$

- A string s is confusable with all strings f whose generating polynomials satisfy

$$P_s(x,y)P_s(\frac{1}{x},\frac{1}{y}) = P_f(x,y)P_f(\frac{1}{x},\frac{1}{y})$$

- To reconstruct a string from its composition multiset $\mathcal{S}$, find a generating polynomial satisfying

$$P(x,y)P(\frac{1}{x},\frac{1}{y}) = n+1+\mathcal{S}(x,y)+\mathcal{S}(\frac{1}{x},\frac{1}{y})$$

Reciprocals of polynomials

- Let z and w be the largest degrees of x and y in $P(x, y)$
- The reciprocal of $P(x, y)$ is

$$P^*(x,y) = x^z y^w P(\frac{1}{x}, \frac{1}{y})$$

- Example:

$$(y + 3xy - 2y^2)^* = xy^2 \left(\frac{1}{y} + 3\frac{1}{xy} - 2\frac{1}{y^2} \right) = xy + 3y - 2x$$

Reciprocals of generating polynomial

- Let s^* be the reversal of s. Then $P_s^* = P_{s^*}$.
- Example:

$$\begin{aligned} P_{0100^*} &= \left(1 + x + xy + x^2y + x^3y\right)^* \\ &= 1 + x + x^2 + x^2 + x^2y + x^3y = P_{0010} \end{aligned}$$

- String s and t are confusable if and only if $P_s P_s^* = P_t P_t^*$.
- Follows from

$$P(x,y)P(\frac{1}{x},\frac{1}{y}) = n + 1 + \mathcal{S}(x,y) + \mathcal{S}(\frac{1}{x},\frac{1}{y})$$

Generating polynomials: Turnpikes and Compositions

- Composition problem: Generalization of Turnpike Problems for bivariate polynomials.
  Some problems possible to answer for composition problems, but not the turnpike problem.

Generating polynomials: Turnpikes and Compositions

- Composition problem: Generalization of Turnpike Problems for bivariate polynomials.
  Some problems possible to answer for composition problems, but not the turnpike problem.
- Composition problem: Special Instant of the Turnpike problem!
  Map 0 to 1, 1 to $n+1$. Then 1100 corresponds to exits $0, 5, 10, 11, 12$.
  Distances are in $\{1, 1, 2, 5, 5, 6, 7, 10, 11, 12, \}$ which translates into $\{0, 0, 1, 1, 01, 0^2 1, 1^2, 01^2, 0^2 1^2\}$.
  Use turnpike algorithms (say, backtracking) for string reconstruction.

Proof of main theorem

Key observations:

- If two strings are confusable, with generating polynomials $P_1(x, y)$ and $P_2(x, y)$, then there exist polynomials $Q(x, y)$ and $R(x, y)$ such that $P_1(x, y) = Q(x, y)R(x, y)$ and $P_2(x, y) = Q(x, y)R^*(x, y)$.

Proof of main theorem

Key observations:

- If two strings are confusable, with generating polynomials $P_1(x, y)$ and $P_2(x, y)$, then there exist polynomials $Q(x, y)$ and $R(x, y)$ such that $P_1(x, y) = Q(x, y)R(x, y)$ and $P_2(x, y) = Q(x, y)R^*(x, y)$.
- If $P_1(x, y)$ is a generating polynomial, and $P_1(x, y)P_1^*(x, y) = P_2(x, y)P_2^*(x, y)$, then $P_2(x, y)$ has to be a generating polynomial or the negative of a generating polynomial.

Proof of main theorem

Key observations:

- If two strings are confusable, with generating polynomials $P_1(x,y)$ and $P_2(x,y)$, then there exist polynomials $Q(x,y)$ and $R(x,y)$ such that $P_1(x,y) = Q(x,y)R(x,y)$ and $P_2(x,y) = Q(x,y)R^*(x,y)$.
- If $P_1(x,y)$ is a generating polynomial, and $P_1(x,y)P_1^*(x,y) = P_2(x,y)P_2^*(x,y)$, then $P_2(x,y)$ has to be a generating polynomial or the negative of a generating polynomial.
- If $P_s(x,y)$ factors into m terms $P_1, P_2, ..., P_m$, then s is confusable with any sting that has a factorization $\bar{P}_1, \bar{P}_2, ..., \bar{P}_m$, with $\bar{P} = P$ or $\bar{P} = P^*$. If P_s is irreducible, s is reconstructable. If ρ_s is number of non-reciprocal factors, the string s can be confused with 2^{ρ_s} strings.

Proof of main theorem

Irreducibility of polynomials in $f \in \mathcal{Z}[x]$ and $f \in \mathcal{Z}[x, y]$:

- If $f(x, y) \in \mathcal{Z}[x, y]$ composite, $f(x, q(x))$ composite as well, for any polynomial $q(x)$.
- If $f(x, q(x))$ irreducible, then $f(x, y)$ irreducible; correspondence $y \rightarrow x$ easy to establish since $f(x, y)$ is a generating polynomial.

Proof of main theorem

Irreducibility of polynomials in $f \in \mathcal{Z}[x]$ and $f \in \mathcal{Z}[x, y]$:

- If $f(x, y) \in \mathcal{Z}[x, y]$ composite, $f(x, q(x))$ composite as well, for any polynomial $q(x)$.
- If $f(x, q(x))$ irreducible, then $f(x, y)$ irreducible; correspondence $y \to x$ easy to establish since $f(x, y)$ is a generating polynomial. Example: $1 + x + x^2$ is irreducible, and so are $1 + x + x^2$, $1 + x + xy$, $1 + y + xy$, and $1 + y + y^2$.
- Converse not true: $f(x, y) = x(x + 1) + y^2$ is irreducible, while $f(x, x) = x(x + 1) + x^2 = x(2x + 1)$ is not.

Proof: $n+1$ is a prime.

- For $n \leq 7$ computer search shows that all strings are reconstructible.
- If $P_s(x, y)$ is irreducible, then s is reconstructible.
- If $P_s(x, x)$ is irreducible, then so is $P_s(x, y)$.
- If $n = p - 1$ for a prime p, then

$$P_s(x, x) = 1 + x + x^2 + \cdots + x^{p-1}.$$

Proof of Theorem: $n + 1$ is a prime.

- If $n = p - 1$ for a prime p, then

$$P_s(x, x) = 1 + x + x^2 + \cdots + x^{p-1}.$$

- Replace x by $x + 1$, to obtain

$$P_s((x + 1), (x + 1)) = 1 + (x + 1) + (x + 1)^2 + \cdots + (x + 1)^{p-1}.$$

- Eisenstein criteria: Sufficient conditions for polynomials with integer coefficients to be irreducible over the integers

$$mod(a_i, p) = 0,\ i \leq n = p - 1, mod(a_{p-1}, p) \neq 0,\ \ mod(a_0, p^2) \neq 0.$$

Proof of Theorem: $n + 1$ is twice a prime.

- If $n = 2p - 1$ for a prime p, then

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

- Three different ways to write $P(x, y)$ as product of $f(x, y)$ and $g(x, y)$ for non-trivial, non-reciprocal f and g.
- In each case, can show that string is reconstructible

When do "confusions" occur?

- Strings 01001101 and 01101001 are confusable.
- These strings can be parsed as **01**0**01**1**01** and **01**1**01**0**01**
- Both have a common substring **01**, interleaved with 01.
- Interleaving s with bits of t: $s \circ t = st_1st_2 \cdots st_ms$
- Strings $s \circ t$ and $s \circ t^*$ are confusable:

When do "confusions" occur?

- Generalization: A string $s = s_1 \circ s_2 \circ s_3 \dots s_k$ is confusable with any string of the form $s = \bar{s}_1 \circ \bar{s}_2 \circ \bar{s}_3 \dots \bar{s}_k$. For k, get that $n+1 = (a+1)(b+1), a, b \geq 2$.
- Hence,

$$2^{\lfloor \frac{e_0}{2} \rfloor + e_1 + \cdots + e_k} \leq C_n \leq \min \left\{ 2^{(e_0+1)(e_1+1)\cdots(e_k+1)-1}, (n+1)^{1.23} \right\}$$

- Upper bound: $P(x,x) = (x^{n+1}-1)/(x-1)$, and $x^n - 1 = \prod_{d|n} \Phi_d(x)$, where Φ's are cyclotomic polynomials (think BCH codes!). Degree of terms - Euler totient function - equals number of integers co-prime to d.

Algorithms I

$f \in \mathcal{Z}[x]$ and $f \in \mathcal{Z}[x, y]$:

- Zassenhaus algorithm (1969);
- LenstraLenstraLovasz (LLL) (1982);
- van Hoeij (2002);
- Kluners et.al. (2010) - bivariate polynomial.
- Key ideas: Relate factorization of univariate polynomials to factorization of polynomial mod p, where p is a prime; relate factorization of bivariate polynomials to factorization of univariate polynomials;
  Use Hensel's lifting: lift factorization mod p to mod $p^k, k > 1$, or mod y^k to $mody^{2k}$ - surveys by Kluner and Sudan.

Algorithms II

- $\mathcal{S}(x, x^{n+1})$ as part of a turnpike problem algorithm.
- Perform factorization (LLL, Hoeij etc).
- Look at exponents in polynomial combinations obtained from factorization. Exponents determine where "0"s and "1"s.
- Algorithm is polynomial in n (complexity $O(n^{2.46})$!.
- May also use the backtracking algorithm!

Algorithms III

- Backtracking algorithm: Start from largest distance, and move down-wards (similar to sequential decoding!)
  Example: $\{1, 3, 3, 6, 7, 9\}$. Start with placing points at 0 and 9. Now look at how to realize distance 7 - shall the point go left or right?
  a) Two choices at each step, sometimes locally indistinguishable.
  b) Try both choices; take one first and then backtrack if you hit a dead end.
- Total time: $O(2^n\, n \log n)$.
- Best known provable worst case solution. But not the case for the special distances induced by composition problem! Know composition problem polynomial, turnpike problem in fewP, unlikely NP hard.

Future Work I: Multiple strings?

- Problem: substrings of two irreversible strings combined. What are the exist non-confusable instances?
- Motivation: simultaneous TMS of a collection of proteins.
- Equivalent to reconstructing one string using only a subset of strings: all substrings of the first half, all substrings of the second half.
- For each length $n > 3$ there exist at least two non-trivial confusable pairs.

  Example: for $n = 4$, $\{1100, 1010\} \sim \{1001, 0110\}$.

  Example: for $n = 8$, 159 confusable pairs.
- Use $P_{s_1}(x, y)P_{s_1}(1/x, 1/y) + P_{s_2}(x, y)P_{s_2}(1/x, 1/y) = P_{t_1}(x, y)P_{t_1}(1/x, 1/y) + P_{t_2}(x, y)P_{t_2}(1/x, 1/y)$ criteria.

Future Work II: How many substrings suffice?

- Information regarding all substrings is redundant: length 1 and length n, for example.
- Using backtracking-based ideas, can show that a large class of strings can be reconstructed using only substrings of length $\geq n/2$.
- There exist examples where one must know at least half the substrings of length 2.

Future Work III: Can one correct errors?

- Error: one (or r) element(s) in one (l) composition(s) is (are) wrong!
- Example: $\mathcal{S}_{001} = \{0, 0, 1, 0^2, 01, 0^21\}$. Instead, you are given $\{0, 0, 1, 01, 01, 0^21\}$.
- Which errors can be corrected?

To the audience:

THANK YOU!