Alignment of DNA mass-spectral profiles using network flows

Pavel Skums[1], Olga Glebova[2], Alex Zelikovsky[2], Zoya Dimitrova[1], David S. Campo[1],Lilia Ganova-Raeva[1] and Yury Khudyakov[1]

[1] Laboratory of Molecular Epidemiology and Bioinformatics, Division of Viral Hepatitis, Centers for Disease Control and Prevention, 1600 Clifton Road NE, 30333 Atlanta, GA, USA

[2] Department of Computer Science, Georgia State University, 34 Peachtree str., 30303, Atlanta, GA, USA

Abstract. Mass spectrometry (MS) of DNA fragments generated by base-specific cleavage of PCR products emerges as a cost-effective and robust alternative to DNA sequencing. MS has been successfully applied to SNP discovery using reference sequences, genotyping and detection of viral transmissions. Although MS is yet to be adapted for reconstruction of genetic composition of complex intra-host viral populations on the scale comparable to the next-generation DNA sequencing technologies, the MS profiles are rich sources of data reflecting the structure of viral populations and completely suitable for accurate assessment of genetic relatedness among viral strains. However, owing to a data structure, which is significantly different from sequences, application of MS profiles to genetic analyses remains a challenging task. Here, we develop a novel approach to aligning DNA MS profiles and assessment of genetic relatedness among DNA species using spectral alignments (MSA). MSA was formulated and solved as a network flow problem. It enables an accurate comparison of MS profiles and provides a direct evaluation of genetic distances between DNA molecules without invoking sequences. MSA may serve as accurately as sequence alignments to facilitate phylogenetic analysis and, as such, has numerous applications in basic research, clinical and public health settings.

1 Introduction

Mass spectrometry (MS) of DNA fragments generated by base-specific cleavage of PCR products is a cost-effective and robust alternative to DNA sequencing. MS is cheaper and less labor-intensive than most of the next-generation sequencing technologies [7][8], and also is not prone to the errors characteristic for these technologies. MS has been successfully applied to the reference-guided single nucleotide polymorphism (SNP) discovery [1][17][13], genotyping [7][10], viral transmission detection [8], identification of pathogens and disease susceptibility genes [15][19], DNA sequence analysis [9], analysis of DNA methylation [18], simultaneous detection of bacteria [14] and viruses [16][20].

MS technology is based on matrix-assisted laser desorption/ ionization time-of-flight (MALDI-TOF) analysis of complete base-specific cleavage reactions of a target RNA obtained from PCR fragments [10][17]. RNA transcripts generated from both strands of PCR fragment are cleaved by RNaseA at either U or C, thus querying for every of the 4 nucleotides (A, C, U and G) in separate reactions. Cleavage at any one nucleotide; e.g. U, generates a number of short fragments corresponding to the number of U's in the transcript. The mass and size of the fragments differ based on the number of A, C and G nucleotides residing between the U's that flank each short fragment. The fragments are resolved by MALDI-TOF-MS, resulting in mass spectral profiles, where each peak defines a specific mass measured in Daltons and has intensity that corresponds to the number of molecules of identical masses.

It should be noted that in MALDI-TOF-MS technology all molecules are equally singly charged, so the actual molecular weights could be obtained simply by subtracting the mass of a single hydrogen from every mass from MS profile. Therefore further in the paper we assume that MS profiles contain molecular weights of corresponding DNA molecules.

Unlike sequencing, MS is not readily applicable to reconstruction of the genetic composition of DNA/RNA populations. Algorithms for reconstruction of sequences from MS data were proposed [2]; but, owing to technological and computational limitations, none is widely used.

MS may serve as a rich source of information about the population structure and the genetic relations among populations without sequences reconstruction. One of the most important applications of sequences is to phylogenetic reconstructions. However, construction of phylogenetic trees requires knowledge of genetic distances among species rather than sequences, with sequences being merely used to estimate the distances. Comparison of MS profiles may also accurately approximate genetic distances. The problem of calculating the distance between two MS samples is known as spectral alignment problem [3][11]. It is usually formulated as follows: match the masses from two MS profiles in such a way that some predefined objective function is maximized or minimized. We introduce and discuss the most common objective functions and methods for spectral alignment problem solution in section 2.

Spectral alignment is crucial for the most applications of MS based on the matching of the sample and reference spectra, with the reference MS spectrum generated in silico.

Spectral alignments are also used for MS data of proteins [12], but the protein technology and, therefore, the problem formulation and algorithm for its solution are completely different.

In this paper we propose a new formulation of the problem of aligning of the base-specific cleavage MS profiles (MS-Al) and present a method for its finding. The method is based on the reduction of the problem to the network flow problem. MS-Al allows *de novo* comparison of sampled populations and may be used for phylogenetic analysis and viral transmissions detection. For

conservative genomes (such as human genome) it allows accurate estimation of actual genetic distance between DNA sequences.

2 Problem formulation

MS profile $P = \{p_1, ..., p_n\}$ consists of n peaks, where each peak $p_i = (m(p_i), f(p_i))$ is represented by a mass $m(p_i)$ and intensity $f(p_i)$. Further without loss of generality we assume that $f(p_i)$ is an integer proportional to the number of occurrences of the mass $m(p_i)$ in the sample. In the simplest version, the spectral alignment problem could be formulated as follows [3]:

Problem 1.

Input: Two MS profiles $P^1 = \{p_1^1, ..., p_{n_1}^1\}$ and $P^2 = \{p_1^2, ..., p_{n_2}^2\}$

Find: Two subsets $P_*^1 \subseteq P^1$ and $P_*^2 \subseteq P^2$ of matched peaks and a bijection $\pi : P_*^1 \to P_*^2$ such that the following objective function is maximized:

$$score(P_*^1, P_*^2, \pi) - \sum_{p_i^1 \in P^1 \setminus P_*^1} pen(p_i^1) - \sum_{p_i^2 \in P^2 \setminus P_*^2} pen(p_i^2) \tag{1}$$

Here *score* is a matching score function and *pen* is a mismatch penalty function. Usually it is assumed [3] that the function score is additive, which means that matches between different peaks are independent:

$$score(P_*^1, P_*^2, \pi) = \sum_{p_i^1 \in P_*^1} score(p_i^1, \pi(p_i^1)) \tag{2}$$

Most of known score functions are based on matches of peaks with close masses. In the simplest case we can put $pen \equiv 0$ and

$$score(p_i^1, p_j^2) = \begin{cases} 1, & |m_i^1 - m_j^2| < \epsilon; \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Using these functions and a greedy algorithm for solving Problem 1, authors of [4][8] accurately identified HCV transmission clusters.

In general, Problem 1 with a score function (2) could be efficiently solved using dynamic programming [3][11]. However, it assumes that matches between different peaks are independent. In some cases this is not true, and taking into account dependencies between peak matches may significantly improve the quality of an alignment. One such case is MS based on a complete base-specific cleavage. Further we formulate spectral alignment problem in that case.

Let $\Sigma = \{\sigma_1, ..., \sigma_4\} = \{C, A, G, T\}$ be an alphabet, and let Σ^* be the set of strings over Σ. We assume that Σ^* contains the empty string o. Let $s = (s_1, ..., s_n) \in \Sigma^*$ and let $\Sigma_k = \Sigma \setminus \{\sigma_k\}$, $k = 1, ..., 4$. For each $\sigma_k \in \Sigma$ define $s(\sigma_k) = s(k)$ as

$$s(k) = \begin{cases} \{s\}, \ s_i \neq \sigma_k \text{ for every } i = 1, ..., n; \\ \{x \in \Sigma_k^* : s \in \{x\sigma_k y, z\sigma_k x, z\sigma_k x\sigma_k y\} \text{ for some } y, z \in \Sigma^*\}, \ \text{otherwise.} \end{cases} \tag{4}$$

(see [2]). In other words, $s(k)$ is the set of all maximal substrings of s, which does not contain σ_k. For $s^1, s^2 \in \Sigma^*$ denote by $r_{s^1}(s^2)$ the number of substrings of s^1 equal to s^2.

Let $m(\sigma_k)$, $k = 1, ..., 4$ be the mass of the nucleotide σ_k and $m(s) = \sum_{i=1}^{n} m(s_i)$ be the mass of molecule represented by a sequence s.

Suppose that $S = \{s^1, ...s^m\}$, $s^j \in \Sigma^*$, is a sample tested using MS with base-specific cleavage. Let $S(k) = \bigcup_{j=1}^{m} s^j(k)$. MS profile P of S is partitioned into four subprofiles: $P = P(A) \cup P(G) \cup P(C) \cup P(T)$, where

$$P(\sigma_k) = \{p_i^{\sigma_k} = (m, f) : m \in \{m(s) : s \in S(k)\}, f = \sum_{\substack{s \in S(k): \\ m(s)=m}} \sum_{j=1}^{m} r_{s^j}(s)\} \quad (5)$$

Example 1. Let $S = \{s\}$ and $R = \{r\}$ be two samples each containing one sequence, s =AAGCTAGTTCA, r =AAGCTCGTTCA. Then

$$s(C) = \{\text{AAG,TAGTT,A}\}, s(A) = \{\text{GCT,GTTC}\},$$
$$s(G) = \{\text{AA,CTA,TTCA}\}, s(T) = \{\text{AAGC,AG,CA}\}$$
$$r(C) = \{\text{AAG,T,GTT,A}\}, r(A) = \{\text{GCTCGTTC}\},$$
$$r(G) = \{\text{AA,CTC,TTCA}\}, r(T) = \{\text{AAGC,CG,CA}\}$$

If $P_S = P_S(C) \cup P_S(A) \cup P_S(G) \cup P_S(T)$ and $Q_R = Q_R(C) \cup Q_R(A) \cup Q_R(G) \cup Q_R(T)$ are MS profiles of S and R, respectively, then they have the following form:

$P_S(C)$	$Q_R(C)$
p_1^C=(2m(A)+m(G),1)	q_1^C=(2m(A)+m(G),1)
p_2^C=(3m(T)+m(A)+m(G),1)	q_2^C=(m(T),1)
p_3^C=(m(A),1)	q_3^C=(2m(T)+m(G),1)
	q_4^C=(m(A),1)
$P_S(A)$	$Q_R(A)$
p_1^A=(m(G)+m(C)+m(T),1)	q_1^A=(3m(T)+3m(C)+2m(G),1)
p_2^A=(2m(T)+m(G)+m(C),1)	
$P_S(G)$	$Q_R(G)$
p_1^G=(2m(A),1)	q_1^G=(2m(A),1)
p_2^G=(m(C)+m(T)+m(A),1)	q_2^G=(2m(C)+m(T),1)
p_3^G=(2m(T)+m(C)+m(A),1)	q_3^G=(2m(T)+m(C)+m(A),1)
$P_S(T)$	$Q_R(T)$
p_1^T=(2m(A)+m(G)+m(C),1)	q_1^T=(2m(A)+m(G)+m(C),1)
p_2^T = (m(A)+m(G),1)	q_2^T = (m(C)+m(G),1)
p_3^T=(m(C)+m(A),1)	q_3^T=(m(C)+m(A),1)

6 of 11 peaks from P_S could be matched by the equal masses and the cleavage base with peaks from Q_R (p_1^C and q_1^C, p_3^C and q_4^C, p_1^G and q_1^G, p_3^G and q_3^G, p_1^T and q_1^T, p_3^T and q_3^T). However, it is easy to see that a single A-C SNP at position

6 between s and r causes the following relations between masses of remaining peaks:

$$m(p_2^C) = m(q_2^C) + m(q_3^C) + m(A) \tag{6}$$

$$m(p_1^A) + m(p_2^A) + m(C) = m(q_1^A) \tag{7}$$

$$m(p_2^G) - m(A) = m(q_2^G) - m(C) \tag{8}$$

$$m(p_2^T) - m(A) = m(q_2^T) - m(C) \tag{9}$$

If peaks and pairs of peaks are matched according to the relations (6)-(9) (p_2^C and (q_2^C, q_3^C), (p_1^A, p_2^A) and q_1^A, p_2^G and q_2^G, p_2^T and q_2^T), then all peaks from P_S and Q_R will be matched. Moreover, masses of single nucleotides and subprofiles involved in (6)-(9) allow to guess the corresponding SNP between s and r and in some cases the number of such type of matches allows to estimate the number of SNP's (in this example 1 SNP).

In general, the relations analogous to (6)-(9) have the following form:

$$m(p_i^{\sigma_{k_1}}) = m(q_{i_1}^{\sigma_{k_1}}) + m(q_{i_2}^{\sigma_{k_1}}) + m(\sigma_{k_2}) \tag{10}$$

$$m(p_{j_1}^{\sigma_{k_2}}) + m(p_{j_2}^{\sigma_{k_2}}) + m(\sigma_{k_1}) = m(q_j^{\sigma_{k_2}}) \tag{11}$$

$$m(p_{h_1}^{\sigma_{k_3}}) - m(\sigma_{k_2}) = m(q_{h_2}^{\sigma_{k_3}}) - m(\sigma_{k_1}) \tag{12}$$

$$m(p_{l_1}^{\sigma_{k_4}}) - m(\sigma_{k_2}) = m(q_{l_2}^{\sigma_{k_4}}) - m(\sigma_{k_1}) \tag{13}$$

Usually there are many possible alternative matches between peaks according to (10)-(13). The goal is to choose the optimal assignments such that the alignment score is maximized. Therefore the problem could be formulated as follows. Let $P_{(2)}$ be a set of all 2-element subsets of a set P. For $p \in P$ denote by $P_{(2)}(p)$ the set of all 2-subsets containing p. If P is a MS-profile, add to P an auxiliary empty peak $p_\epsilon = (0, \infty)$ with 0 mass and unbounded intensity. We will call such profile an extended MS profile. We assume without loss of generality that all other peaks have intensity 1 (otherwise if peak p_i has intensity $f(p_i) > 1$ replace it with $f(p_i)$ peaks of intensity 1). Further, extend an alphabet Σ by addition of an auxiliary empty symbol ϵ with $m(\epsilon) = 0$. Those additional objects are needed to include insertions, deletions and mutations in homopolymers (i.e. sequences of identical nucleotides) in the model.

Problem 2.

Input: Two extended MS profiles $P^1 = \{p_1^1, ..., p_{n_1}^1\} = P^1(C) \cup P^1(A) \cup P^1(G) \cup P^1(T) \cup \{p_\epsilon\}$ and $P^2 = \{p_1^2, ..., p_{n_2}^2\} = P^2(C) \cup P^2(A) \cup P^2(G) \cup P^2(T) \cup \{p_\epsilon\}$

Find: Two subsets $P^1_* \subseteq P^1 \cup P^1_{(2)}$ and $P^2_* \subseteq P^2 \cup P^2_{(2)}$ of matched peaks and pairs of peaks and a bijection $\pi : P^1_* \to P^2_*$ such that the following conditions hold:

(i) $|P^j_* \cap (P_{(2)}(p^j_l) \cup \{p^j_l\})| \leq 1$ for every $p^j_l \in P^j \setminus \{p^j_\epsilon\}$, $j = 1, 2$ (every peak is matched at most once either as a singleton or as a member of a pair)
(ii) $\pi(\{p^1_i, p^1_j\}) \in P^2$ for every pair $\{p^1_i, p^1_j\} \in P^1_{(2)}$ (pair of peaks should be matched to a single peak);
(iii) there exists a bijection $\psi : P^1_* \cap P^1_{(2)} \to P^2_* \cap P^2_{(2)}$ (matchings of pairs of peaks go in pairs)

and the objective function (1) is maximized. The objective function should be defined in such a way that

a) a pair of peaks is matched to a peak and vise versa only if (10) and (11) holds for them; the bijection ψ maps pairs which are conjugate by (10) and (11);
b) the number of matches involving pairs is as small as possible. Each such match potentially corresponds to an insertion, deletion or replacement and we are trying to align MS profiles with the smallest number of involved mismatches as possible - analogously to alignment of sequences using edit distance.

In the next section we will show how to define such a function and present an algorithm for its calculation. This is a new approach, which, as Example 1 shows, is more accurate than the approaches with direct matchings of peaks.

Note that (10)-(13) holds for a certain SNP, if it is isolated, which means that substrings between it and the closest SNPs contain all four nucleotides. For the conservative genomes this is a reasonable assumption: it was shown in [1] that the overwhelming majority of SNPs in human genome are isolated (for the data analyzed in [1] the average and minimal distance between two neighbor SNPs is 231bp and 14bp, respectively). Therefore for such genomes a solution of Problem 2 provides a reliable estimation for the number and types of SNPs. If two mutations happen in close proximity, then the relation between peaks caused by them is more complex than (10)-(13). Moreover, if sample contains more than one unknown sequence, it is usually impossible to separate peaks between sequences. Therefore for a highly mutable genomes, such as viral genomes, solution of Problem 2 provides a distance, which specifies and generalizes the most commonly used distance with the score function (3), instead of direct estimation of the number of mismatches.

3 Network flow method for spectral alignment

For a directed graph (or network) N with a vertex set V, an arcs set A, pair of source and sink $s, t \in V$, arcs capacities cap and possibly arc costs $cost$ a network flow is a mapping $f : A \to \mathbb{R}_+$ such that $f(a) \leq cap(a)$ for every $a \in A$

(capacity constraints) and $\sum_{uv \in A} f(uv) - \sum_{vw \in A} f(vw) = 0$ for every $v \in V \setminus \{s,t\}$ (flow conservation constraints). The value of flow is $|f| = \sum_{sv \in A} f(sv)$. Classical network flow problem either searches for a flow of maximum value (Maximum Flow Problem) or for a flow with a given value of a minimum cost (Minimum-cost Flow Problem)

It is well-known, that in discrete optimization many matching-related problems (such as Maximum Bipartite Matching Problem, Assignment problem, Minimum Cost Bipartite Perfect Matching Problem, Linear Assignment Problem, etc.) could be solved using either network flows or shortest path - based algorithms. It suggests that a similar approach could be used for Problem 2. However, the formulation of Problem 2 is more complex that of above-mentioned problems, so the reduction of Problem 2 to the network flow-based problem appeared to be rather sophisticated. Below we present that reduction.

Let $P^1 = \{p_1^1, ..., p_{n_1}^1\} = P^1(C) \cup P^1(A) \cup P^1(G) \cup P^1(T) \cup \{p_\epsilon\}$ and $P^2 = \{p_1^2, ..., p_{n_2}^2\} = P^2(C) \cup P^2(A) \cup P^2(G) \cup P^2(T) \cup \{p_\epsilon\}$ be extended MS profiles. Let also $\delta \in \mathbb{R}_+$ be the mass precision, $g \in \mathbb{R}_+$ be the mismatch penalty and $p, q \in \mathbb{R}_+$ be the mutation (i.e. replacement, insertion, deletion) penalties corresponding to pairs of relations (10),(11) and (12),(13), respectively. Construct the network

$$N = (V, A, l, m, cost, cap) \tag{14}$$

where $l : V \to \Sigma^*$ is a vertices labels function, $m : V \to \mathbb{R}_+$ is vertices weights function, $cost : A \to \mathbb{R}_+$ and $cap : A \to \mathbb{R}_+$ are cost and capacity functions of arcs, respectively. Vertex set

$$V = \{s,t\} \cup V_1 \cup V_2 \cup V_{p_1} \cup V_{p_2} \cup V_{a_1} \cup V_{a_2} \cup V_{d_1} \cup V_{d_2}$$

and arc set A are constructed as follows:

1) s and t are the source and sink, respectively.
2) for each peak $p_i^j \in P^j(\sigma)$, $j = 1,2$, $i = 1, ..., n_j$, $\sigma \in \Sigma$ the set V_j contains $f(p_j^i)$ vertices $v_j^i(1), ..., v_j^i(f(p_j^i))$. For each $v_j^i(k)$ $l(v_j^i(k)) = \sigma$, $m(v_j^i(k)) = m(p_i^j)$. For an empty peak $p_\epsilon \in P^j$, $j = 1,2$, the set V_j contain the unique vertex v_ϵ^j with $l(v_\epsilon^j) = o$ and $m(v_\epsilon^j) = 0$.
3) For each $v \in V_1 \setminus \{v_\epsilon^1\}$ the set A contains an arc sv with $cost(sv) = 0$ and $cap(sv) = 1$. For each $v \in V_2 \setminus \{v_\epsilon^2\}$ A contains an arc vt with $cost(vt) = 0$ and $cap(vt) = 1$. There are also arcs sv_ϵ^1 and $v_\epsilon^2 t$ with $cost(sv_\epsilon^1) = cost(v_\epsilon^2 t) = 0$ and $cap(sv_\epsilon^1) = cap(v_\epsilon^2 t) = \infty$.
4) $uv \in A$ for each $u \in V_1$, $v \in V_2$ such that $|m(u) - m(v)| < \delta$ and $l(u) = l(v)$; $cost(uv) = 0$, $cap(uv) = 1$.
5) For every $u, v \in V_1$ and $w \in V_2$ such that
   a) $l(u) = l(v) = l(w)$,
   b) there exists $\sigma \in \Sigma$ such that $|m(u) + m(v) + m(\sigma) - m(w)| < \delta$,

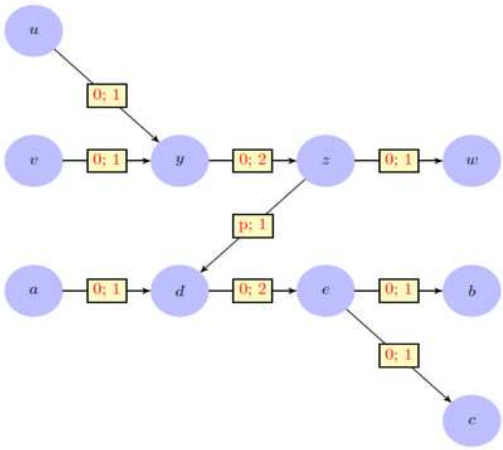

Fig. 1. Edges corresponding to relations (10),(11)

the vertex set V contains vertices $y \in V_{p_1}$ and $z \in V_{a_1}$ with $m(y) = m(z) = 0$, $l(y) = o$, $l(z) = l(u)\sigma$. The set A contains arcs uy, vy, yz, zw with $cost(uy) = cost(vy) = cost(yz) = cost(zw) = 0$, $cap(uy) = cap(vy) = cap(zw) = 1$, $cap(yz) = 2$. See Figure 1. The subgraph $N[u, v, w, y, z]$ induced by vertices u, v, w, y, z will be referred as left fork.

6) Analogously, for every $a \in V_1$ and $b, c \in V_2$ such that
   a) $l(a) = l(b) = l(c)$,
   b) there exists $\sigma \in \Sigma$ such that $|m(a) - m(b) - m(c) - m(\sigma)| < \delta$,

   the set V contains vertices $d \in V_{a_2}$ and $e \in V_{p_2}$ with $m(d) = m(e) = 0$, $l(e) = o$, $l(d) = \sigma l(b)$. The set A contains arcs ad, de, eb, ec with $cost(ad) = cost(de) = cost(eb) = cost(ec) = 0$, $cap(ad) = cap(eb) = cap(ec) = 1$, $cap(de) = 2$. See Figure 1. Further the subgraph $N[a, b, c, d, e]$ will be referred as right fork.
7) For vertices $u \in V_{a_1}$, $v \in V_{a_2}$ the set A contains an arc uv with $cost(uv) = p$ and $cap(uv) = 1$, if $l(u) = l(v)$. See Figure 1.
8) For every $u \in V_1$ and $v \in V_2$ such that
   a) $l(u) = l(v)$,
   b) there exists $\sigma_1, \sigma_2 \in \Sigma$ such that $|m(u) - m(\sigma_1) - m(v) + m(\sigma_2)| < \delta$,

   the set V contains vertices $y \in V_{d_1}$ and $z \in V_{d_2}$ with $m(y) = m(z) = 0$, $l(y) = l(z) = \sigma_1\sigma_2$. The set A contains arcs uy, yz, zv with $cost(uy) = cost(yz) = cost(zv) = 0$, $cap(uy) = cap(zv) = 1$, $cap(yz) = 0$. See Figure 2.
9) for all distinct vertices $y, a \in V_{d_1}$, $z, b \in V_{d_2}$ such that $yz, ab \in A$, $cap(yz) = cap(ab) = 0$ and $l(y) = l(b)$, the set A contains arcs yb,az with $cost(yb) = cost(az) = \frac{q}{2}$, $cap(yb) = cap(az) = 1$. See Figure 2.
10) For every $v \in V_1$ there exists an arc vs with $cost(vs) = g$ and $cap(vs) = 1$.

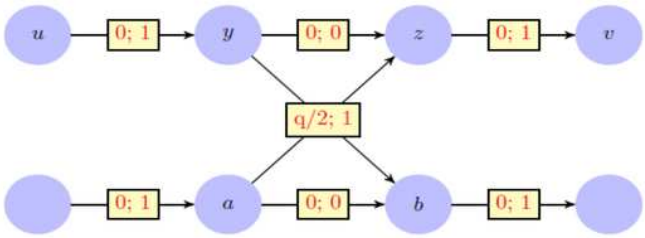

Fig. 2. Edges corresponding to relations (12),(13)

Let $x : A \to \mathbb{N}, a \mapsto x_a$ is a flow in the network N. Problem 2 could be formulated as the following variant of the network flow problem:

$$\text{minimize} \sum_{a \in A} cost(a) x_a \tag{15}$$

subject to

$$\sum_{uv \in A} x_{uv} - \sum_{vw \in A} x_{vw} = 0, \quad v \in V \setminus \{s, t\}; \tag{16}$$

$$\sum_{sv \in A, v \neq v_\epsilon} x_{sv} = |V_1| - 1; \tag{17}$$

$$x_{uy} - x_{vy} = 0, \quad y \in V_{p_1}; \tag{18}$$

$$x_{eb} - x_{ec} = 0, \quad e \in V_{p_2}; \tag{19}$$

$$x_{uy} - x_{zv} = 0; \;\; yz \in A, cap(yz) = cost(uy) = cost(zv) = 0 \tag{20}$$

$$0 \leq x_a \leq cap(a), \;\; a \in A. \tag{21}$$

This formulation differs from the classical network flow problem formulation by additional constraints which require flow to be equal on some prescribed pairs of arcs.

Arcs from 4) provide the possibility of match between peaks with close masses with 0 penalty. Vertices and arcs from 5)-7) and constraints (18)-(19) allow to match peaks with pairs of peaks according to relations (10),(11). The capacities of arcs defined in 5)-7) are chosen in such a way that if flow goes through the left fork, then it should also go through the right fork indicating the same mutation, thus forcing a fulfillment of requirement (iii) of Problem 2. Moreover, if flow goes through some pair of forks, exactly one arc of cost p between those forks is involved, thus forcing penalty for mutation. Vertices and arcs from 8)-9) and constraints (20) play the same role for relations (12),(13). Constraint (17) for total size of the flow ensures that every peak is either matched or penalized for mismatch, which is encoded by arcs from 10). Moreover, arcs from 10) ensure that the problem (15)-(21) always has a feasible solution. (16) and (21) are standard flow conservation and capacity constraints.

If P^1 and P^2 are samples of single genomes with isolated SNPs, then the number of SNPs could be estimated as $|\{a \in A : x_a > 0, cost(a) = p\}|$.

4 Testing results, conclusions and future work

The algorithm was tested on simulated data. 80 pairs of sequences of lengths 40-60bp with 2-4 isolated SNPs were randomly generated. NEW TEXT For each position one of possible symbols was chosen with equal probability to generate first sequence, and then random mutations were introduced on the prescribed positions to generate the second sequence. END OF NEW TEXT MS profiles of generated sequences were simulated using masses $m(A) = 329.21$ DA, $m(T) = 306.17$ DA, $m(G) = 345.21$ DA, $m(C) = 305.18$ DA. The ILP formulation (15)-(21) was solved using GNU Linear Programming Kit (GLPK) (http://www.gnu.org/software/glpk/) on a computer with two 2.67GHz processors and 12 GB RAM. Since ILP solving is usually time-consuming, the time limit 30 seconds per problem was established. For 90% (72 of 80) of test instances correct number of SNPs were estimated within the time limit.

Thus the proposed approach enables an accurate comparison of MS profiles and provides a direct evaluation of genetic distances between DNA molecules without invoking sequences.

The proposed spectral alignment method is expected to be highly effective in evaluating genetic relatedness between viral samples and identifying transmission clusters from viral outbreaks. The reasons behind this presumption is based on the fact, that simple Hamming distance between samples (which corresponds to the score function (3)) was shown to effectively separate transmission clusters [4][8], and Hamming distance could be calculated as a special case of our model with $p = q = \infty$.

ILP-based approach to solving the problem (15)-(21) is time-consuming. Therefore more computationally effective approaches may be required to handle larger samples. It is expected that direct applications of network flow-based methods, Lagrangian relaxations or Bender decompositions should dramatically increase performance of the algorithm. The generalizations of relations (10)-(13)

in order to obtain a model allowing estimation of actual number of mutations in the presence of SNP clusters is also of interest for the future research.

References

1. S. Böcker. SNP and mutation discovery using base-specific cleavage and MALDI-TOF mass spectrometry, Bioinformatics Vol. 19 Suppl. 1 2003, pages i44–i53
2. S. Böcker. Sequencing from Compomers: Using Mass Spectrometry for DNA de novo Sequencing of 200+ nt, Journal Of Computational Biology Volume 11, Number 6, 2004, P. 1110-1134
3. S. Böcker, H.-M. Kaltenbach. Mass spectra alignments and their significance, Journal of Discrete Algorithms Vol. 5, Issue 4, December 2007, P. 714-728
4. Z. Dimitrova, D.S. Campo, S. Ramachandran, G. Vaughan, L. Ganova-Raeva, Y. Lin, J.C. Forbi, G. Xia, P. Skums, B. Pearlman and Y. Khudyakov. Evaluation of viral heterogeneity using next-generation sequencing, end-point limiting-dilution and mass spectrometry, In Silico Biology 11 (2011/2012) 183–192
5. M. Ehrich, S. Böcker and D. van den Boom. Multiplexed discovery of sequence polymorphisms using base-specific cleavage and MALDI-TOF MS, Nucleic Acids Res. 2005; 33(4): e38.
6. Ehrich M, Nelson MR, Stanssens P, Zabeau M, Liloglou T, Xinarianos G, Cantor CR, Field JK and van den Boom D. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America 2005; 102:15785-15790.
7. L. Ganova-Raeva, S. Ramachandran, C. Honisch, J. C. Forbi, X. Zhai and Y. Khudyakov. Robust Hepatitis B Virus Genotyping by Mass Spectrometry, J. Clin. Microbiol. 2010, 48(11):4161.
8. L. Ganova-Raeva, Z. Dimitrova, D.S. Campo, L. Yulin, S. Ramachandran, G.-L. Xia, C. Honisch, C. Cantor, Y. Khudyakov. Detection of hepatitis C virus transmission using DNA mass spectrometry, J Infect Dis. 2013 Jan 31
9. Kirpekar F, Nordhoff E, Larsen LK, Kristiansen K, Roepstorff P and Hillenkamp F. DNA sequence analysis by MALDI mass spectrometry. Nucleic acids research 1998; 26:2554-2559.
10. M. Lefmann, C. Honisch, S. Böcker, N. Storm, F. von Wintzingerode, C. Schlötelburg, Annette Moter, Dirk van den Boom and Ulf B. Göbel. Novel Mass Spectrometry-Based Tool for Genotypic Identification of Mycobacteria, Journal Of Clinical Microbiology, Jan. 2004, p. 339–346
11. V. Mäkinen. Peak alignment using restricted edit distances, Biomolecular Engineering 10/2007; 24(3):337-42
12. P.A. Pevzner, V. Dancik, and C.L. Tang. Mutation-tolerant protein identification by mass-spectrometry. Journal of Computational Biology, 7:777–787, 2000.
13. Pusch W, Kraeuter KO, Froehlich T, Stalgies Y and Kostrzewa M. Genotools SNP manager: a new software for automated high-throughput MALDI-TOF mass spectrometry SNP genotyping. Biotechniques 2001; 30:210-215.
14. Rees JC, Voorhees KJ. Simultaneous detection of two bacterial pathogens using bacteriophage amplification coupled with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 2005; 19:2757-2761.

15. Sampath R, Hall TA, Massire C, Li F, Blyn LB, Eshoo MW, Hofstadler SA and Ecker DJ. Rapid identification of emerging infectious agents using PCR and electrospray ionization mass spectrometry. Ann. NY Acad. Sci. 2007; 1102:109-120.
16. Sjoholm MI, Dillner J, Carlson J. Multiplex detection of human herpesviruses from archival specimens by using matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol 2008; 46:540-545.
17. P. Stanssens, M. Zabeau, G. Meersseman, G. Remes, Y. Gansemans, N. Storm, R. Hartmer, C. Honisch, C. P.Rodi, S. Böcker and D. van den Boom. High-Throughput MALDI-TOF Discovery of Genomic Sequence Polymorphisms, Genome Res. 2004 Jan;14(1):126-33.
18. Tost J, Schatz P, Schuster M, Berlin K and Gut IG. Analysis and accurate quantification of CpG methylation by MALDI mass spectrometry. Nucleic Acids Res 2003; 31:e50.
19. von Wintzingerode F, Bocker S, Schlotelburg C, Chiu NH, Storm N, Jurinke C, Cantor CR, Gobel UB and van den Boom D. Base-specific fragmentation of amplified 16S rRNA genes analyzed by mass spectrometry: a tool for rapid bacterial identification. Proc. Natl. Acad. Sci. USA 2002; 99:7039-7044.
20. Yang H, Yang K, Khafagi A, Tang Y, Carey TE, Opipari AW, Lieberman R, Oeth PA, Lancaster W, Klinger HP, Kaseb AO, Metwally A, Khaled H and Kurnit DM. Sensitive detection of human papillomavirus in cervical, head/neck, and schistosomiasis-associated bladder malignancies. Proc. Natl. Acad. Sci. USA 2005; 102:7683-7688.