Low-Dimensional Representation of Symbolic Sequences

Manuel E. Lladser
Applied Mathematics (Associate)
Computer Science (Affiliate)
University of Colorado - Boulder

AofA 2019 - CIRM Luminy, France

Motivation: Dengue Virus

- Transmitted by mosquitos
- Kills in average 25K human worldwide per year
- Targets octapeptides (**octamers**)
- Target characterization has been elusive, making vaccine development challenging

R	R	R	R	S	V	A	L
R	Q	K	R	S	V	A	L
P	S	M	T	M	R	C	I
M	V	T	A	G	H	G	Q
K	K	Q	R	A	G	V	L
S	K	K	R	S	W	P	L
R	K	R	R	L	T	I	M
K	T	K	R	Y	L	P	A
A	I	K	R	G	L	R	T
A	Q	R	R	G	R	I	G
A	G	R	K	S	L	T	L
N	T	R	R	G	T	G	N

Figure: Some known targets

Motivation: Dengue Virus

- Transmitted by mosquitos
- Kills in average 25K human worldwide per year
- Targets octapeptides (**octamers**)
- Target characterization has been elusive, making vaccine development challenging

R	R	R	R	S	V	A	L
R	Q	K	R	S	V	A	L
P	S	M	T	M	R	C	I
M	V	T	A	G	H	G	Q
K	K	Q	R	A	G	V	L
S	K	K	R	S	W	P	L
R	K	R	R	L	T	I	M
K	T	K	R	Y	L	P	A
A	I	K	R	G	L	R	T
A	Q	R	R	G	R	I	G
A	G	R	K	S	L	T	L
N	T	R	R	G	T	G	N

Figure: Some known targets

Question.

Are there sensible low-dimensional representations of octamers?

Common Numerical Representation Schemes

- One-hot encodings
- k-mer count vectors
- Fisher kernels [1]
- Multidimensional Scaling (PCoA) [2]
- Word2Vec (BioVec [3])
- DeepWalk [4]
- Node2Vec [5]

[1] Jaakkola, Diekhans, and Haussler - 1999
[2] Krzanowski - 2000
[3] Asgari and Mofrad - 2015
[4] Perozzi, Al-Rfou, and Skiena - 2014
[5] Grover and Leskovec - 2016

Common Numerical Representation Schemes (shortcomings)

- One-hot encodings (dimension: ak + sparse)
- k-mer count vectors (non-local + dimension: a^k)
- Fisher kernels [1] (large dataset)
- Multidimensional Scaling (PCoA) [2] (n points map to dimension: $n-1$)
- Word2Vec (BioVec [3]) (large dataset)
- DeepWalk [4] (new sample rerun)
- Node2Vec [5] (new sample rerun)

[1] Jaakkola, Diekhans, and Haussler - 1999
[2] Krzanowski - 2000
[3] Asgari and Mofrad - 2015
[4] Perozzi, Al-Rfou, and Skiena - 2014
[5] Grover and Leskovec - 2016

A Multilateration Approach

Problem.

Develop a sensible low-dimensional representation of points in a large metric space that addresses the shortcomings of the existing methods

A Multilateration Approach

Problem.

Develop a sensible low-dimensional representation of points in a large metric space that addresses the shortcomings of the existing methods

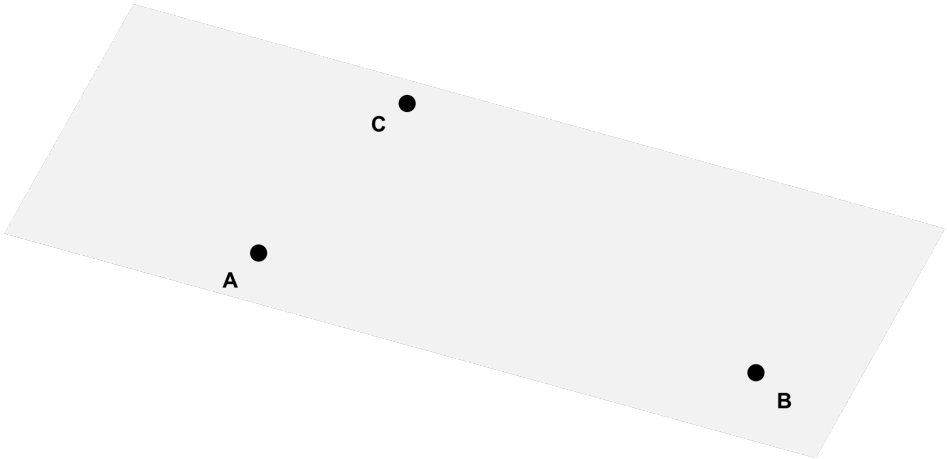


Figure: Three non-colinear points trilaterate the plane

A Multilateration Approach

Problem.

Develop a sensible low-dimensional representation of points in a large metric space that addresses the shortcomings of the existing methods

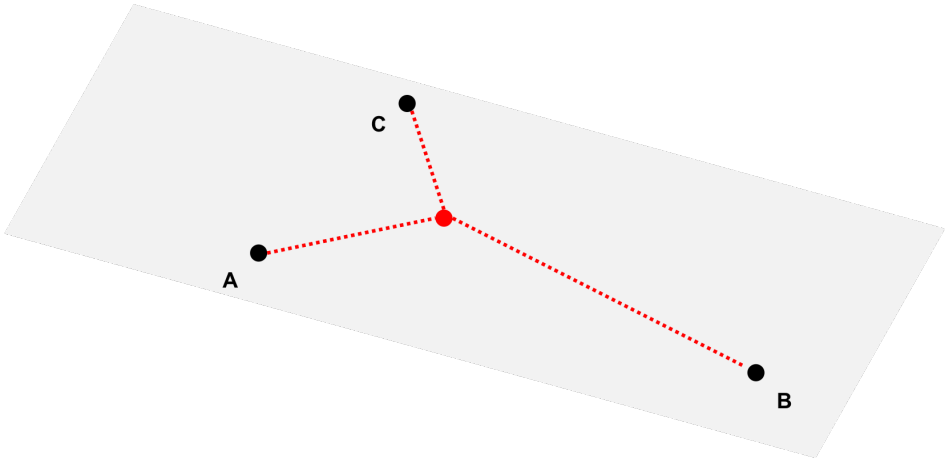


Figure: Three non-colinear points trilaterate the plane

A Multilateration Approach

Problem.

Develop a sensible low-dimensional representation of points in a large metric space that addresses the shortcomings of the existing methods

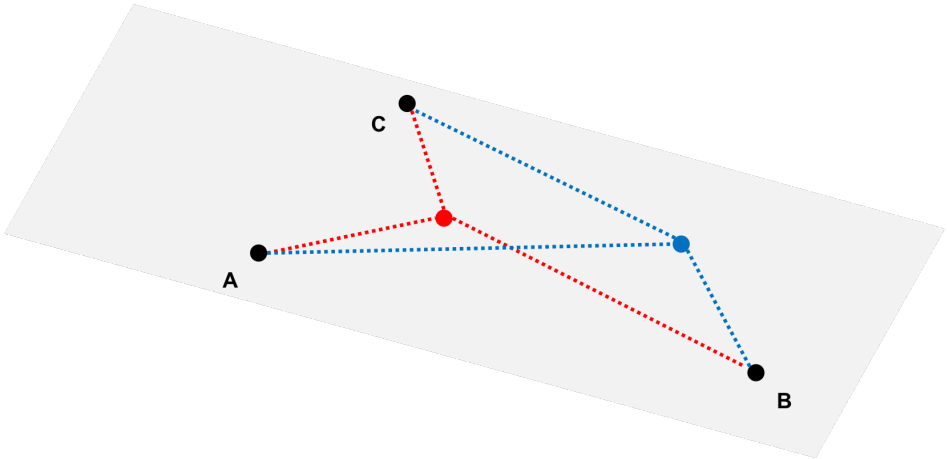
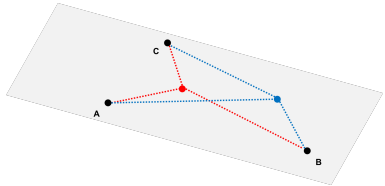


Figure: Three non-colinear points trilaterate the plane

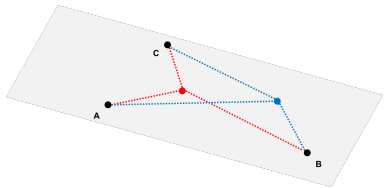
A Multilateration Approach



Question.

Is it possible to multilaterate any finite metric space?

A Multilateration Approach



Question.

Is it possible to multilaterate any finite metric space?

Definitions.

- Call $R \subset V$ **resolving** when $\forall x \neq y \in V \; \exists r \in R$ s.t. $d(x,r) \neq d(y,r)$
- $\Phi(x) := \left(d(x,r)\right)_{r \in R}$ from V to $\mathbb{R}^{|R|}$ is one-to-one and "continuous"
- (**Metric Dimension**.) $\beta(V,d) := \min_{R \subset V \text{resolving}} |R|$

(Terminology borrowed from graph theory!)

Metric Dimension vs Distance Matrix

Observation.

The metric dimension of a metric space (V, d) is the smallest number of columns in its distance matrix D needed to differentiate all the rows

$$D = \begin{bmatrix} 0 & 1 & 2 & 1 & 2 & 3 \\ 1 & 0 & 1 & 1 & 1 & 2 \\ 2 & 1 & 0 & 2 & 2 & 1 \\ 1 & 1 & 2 & 0 & 2 & 3 \\ 2 & 1 & 2 & 2 & 0 & 3 \\ 3 & 2 & 1 & 3 & 3 & 0 \end{bmatrix}$$

Metric Dimension vs Distance Matrix

Observation.

The metric dimension of a metric space (V, d) is the smallest number of columns in its distance matrix D needed to differentiate all the rows

$$D = \begin{bmatrix} 0 & 1 & 2 & 1 & 2 & 3 \\ 1 & 0 & 1 & 1 & 1 & 2 \\ 2 & 1 & 0 & 2 & 2 & 1 \\ 1 & 1 & 2 & 0 & 2 & 3 \\ 2 & 1 & 2 & 2 & 0 & 3 \\ 3 & 2 & 1 & 3 & 3 & 0 \end{bmatrix}$$

Metric Dimension vs Distance Matrix

Theorem: (Tillquist-LI'19)

The general multilateration problem is NP-complete.

Why? Reduction to the *Set Cover Problem*, which is a known [1] NP-complete problem □

[1] Karp - 1972

Metric Dimension of a Graph

- Our terminology is borrowed from a graph theory [1,2]
- A set R of nodes in a graph $G = (V, E)$ is called **resolving** when any vertex in the graph is uniquely determined by its vector of distances to those nodes
- The **metric dimension** of a graph, $\beta(G)$, is the size of a smallest resolving set
- If d is the **geodesic distance** between pairs of nodes in G then

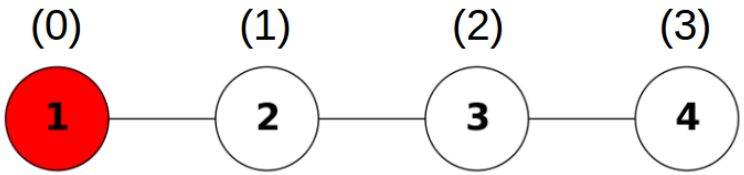
$$\beta(G) = \beta(V, d)$$

[1] Slater - 1975

[2] Harary and Melter - 1976

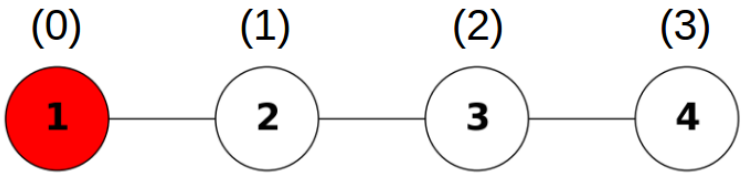
Examples

- $\beta(G) = 1$ if and only if G is a path

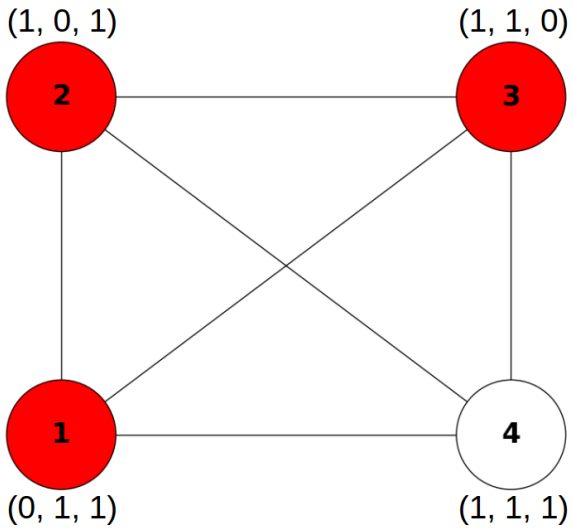


Examples

- $\beta(G) = 1$ if and only if G is a path



- $\beta(G) = (n-1)$ if and only if $G = K_n$



Known Facts About Metric Dimension

- $\beta(G)$ is known for specific graph families such as:
  - trees [1,2]
  - regular bipartite graphs [3]
  - complete n-partite graphs [4]
- Finding $\beta(G)$ is **NP-complete** for a general graph $G = (V, E)$ [5]
- The **Information Content Heuristic** (ICH) finds a resolving set greedily [6]
  - $O(|V|^3)$ time-complexity
  - $O(|V|^2)$ memory-complexity
  - Approximation ratio of $1 + (1 + o(1)) \cdot \ln |V|$

[1] Slater - 1975
[2] Harary and Melter - 1976
[3] Bača, Baskoro, Salman, et. al. - 2011
[4] Saputro, Baskoro, Salman, et. al. - 2009
[5] Hauptmann, Schmied, and Viehmann - 2012; Gary and Johnson - 1979
[6] Hauptmann, Schmied, and Viehmann - 2012

Hamming Graphs

Definition.

$H_{k,a}$ denotes the graph with vertex set $\{0, \ldots, a-1\}^k$, i.e. **k-mers** over an alphabet of size a. Two k-mers are connected by an edge if and only if they are identical except for one character at the same position

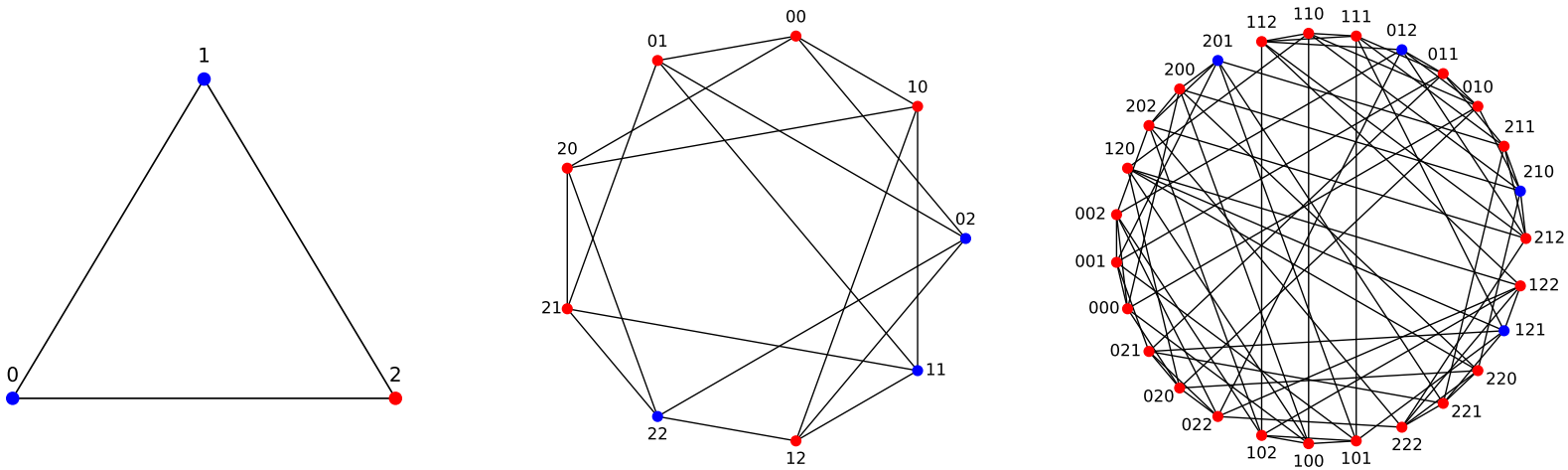


Figure: Visuals of $H_{1,3}$, $H_{2,3}$, and $H_{3,3}$ (resolving sets in blue)

Hamming Graphs

Definition.

$H_{k,a}$ denotes the graph with vertex set $\{0, \ldots, a-1\}^k$, i.e. **k-mers** over an alphabet of size a. Two k-mers are connected by an edge if and only if they are identical except for one character at the same position

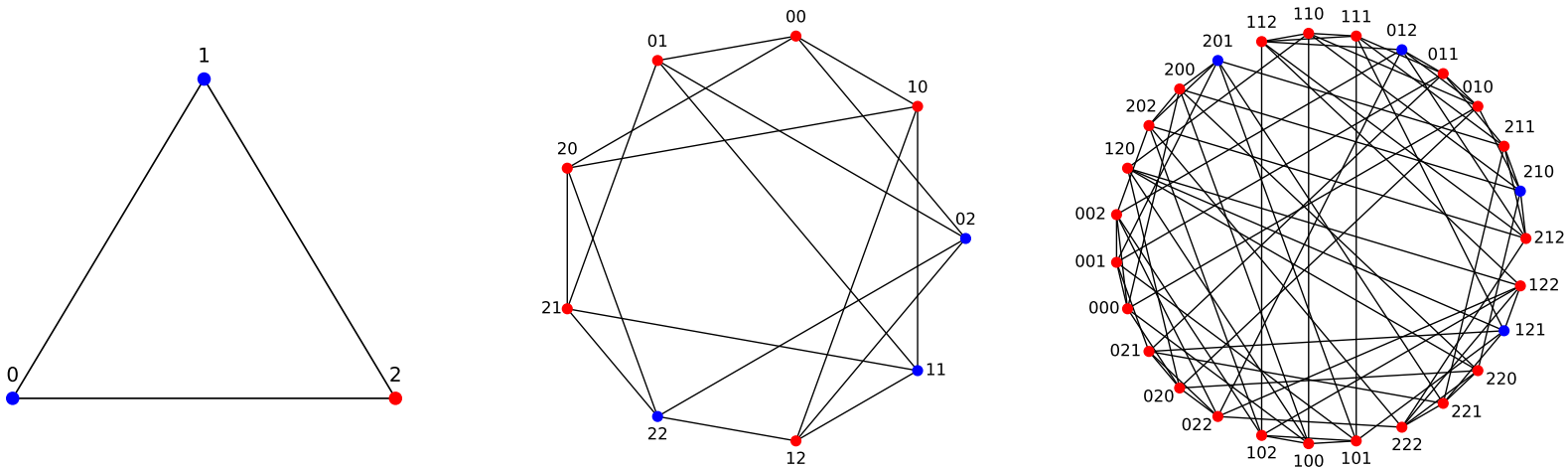


Figure: Visuals of $H_{1,3}$, $H_{2,3}$, and $H_{3,3}$ (resolving sets in blue)

Question.

Can we find small resolving sets in $H_{k,a}$ efficiently?

Theorem: (Tillquist-LI'19)

$\beta(H_{k-1,a}) \leq \beta(H_{k,a}) \leq \beta(H_{k-1,a}) + \lfloor a/2 \rfloor$. *

* Case with $a = 2$ due to Chartrand et al. - 2000

Why? The proof is constructive!

- d:= distance matrix of $H_{k-1,a}$
- D:= distance matrix of $H_{k,a}$

$$D = \begin{array}{c} \\ 0\ldots \\ 1\ldots \\ \vdots \\ (a\text{-}1)\ldots \end{array} \begin{array}{c} \begin{array}{cccc} 0\ldots & 1\ldots & \cdots & (a\text{-}1)\ldots \end{array} \\ \begin{pmatrix} d & d+\mathbf{1} & \cdots & d+\mathbf{1} \\ d+\mathbf{1} & d & \cdots & d+\mathbf{1} \\ \vdots & \vdots & \ddots & \vdots \\ d+\mathbf{1} & d+\mathbf{1} & \cdots & d \end{pmatrix} \end{array}$$

- Any resolving set of $H_{k-1,a}$ distinguishes rows in each block
- Pick additional nodes to resolve rows across blocks □

Algorithm: Hamming Graph Resolving Set Construction

Input: Resolving set r of $H_{k-1,a}$
Output: Resolving set R of $H_{k,a}$
function CONSTRUCTRESOLVINGSET(r,a)
    $R_0 \leftarrow \{\}$
    $R_1 \leftarrow \{\}$
    $i \leftarrow 0$
    for $w \in r$ **do**
        if $i < a$ **then**
            $R_0 \leftarrow R_0 \cup \{iw\}$
            if $i < (a-1)$ **then**
                $R_0 \leftarrow R_0 \cup \{(i+1)w\}$
            end if
        end if
        if $i \geq a$ **then**
            $R_1 \leftarrow R_1 \cup \{0w\}$
        end if
        $i \leftarrow (i+2)$
    end for
    $R \leftarrow (R_0 \cup R_1)$
    return R
end function

Algorithm: Hamming Graph Resolving Set Construction

Input: Resolving set r of $H_{k-1,a}$
Output: Resolving set R of $H_{k,a}$
function CONSTRUCTRESOLVINGSET(r,a)
    $R_0 \leftarrow \{\}$
    $R_1 \leftarrow \{\}$
    $i \leftarrow 0$
    for $w \in r$ **do**
        if $i < a$ **then**
            $R_0 \leftarrow R_0 \cup \{iw\}$
            if $i < (a-1)$ **then**
                $R_0 \leftarrow R_0 \cup \{(i+1)w\}$
            end if
        end if
        if $i \geq a$ **then**
            $R_1 \leftarrow R_1 \cup \{0w\}$
        end if
        $i \leftarrow (i+2)$
    end for
    $R \leftarrow (R_0 \cup R_1)$
    return R
end function

- Time complexity starting with $H_{1,a}$: $O(ak^2)$
- (ICH time complexity: $O(a^{3k})$)
- Finds resolving set for $H_{k,a}$ of size $O(k)$
- $(|H_{k,a}| = a^k)$

Back to Octamers

- ICH gives the following resolving set for $H_{3,20}$:

$$\left\{\begin{array}{llllllllll} \text{AAA}, & \text{RRR}, & \text{NNN}, & \text{DDD}, & \text{CCC}, & \text{QQQ}, & \text{EEE}, & \text{GGG}, & \text{HHH}, & \text{CNS}, \\ \text{III}, & \text{LLL}, & \text{KKK}, & \text{MMM}, & \text{FFF}, & \text{PPP}, & \text{SIS}, & \text{NST}, & \text{TTC}, & \text{QPK} \\ \text{ARW}, & \text{WWD}, & \text{MKY}, & \text{QYE}, & \text{YGL}, & \text{HPV}, & \text{VFR}, & \text{EAG}, & \text{KLQ}, & \text{SVT}, \\ \text{DHF}, & \text{WMP} & & & & & & & & \end{array}\right\}$$

Back to Octamers

- ICH gives the following resolving set for $H_{3,20}$:

$$\left\{\begin{array}{cccccccccc} \text{AAA}, & \text{RRR}, & \text{NNN}, & \text{DDD}, & \text{CCC}, & \text{QQQ}, & \text{EEE}, & \text{GGG}, & \text{HHH}, & \text{CNS}, \\ \text{III}, & \text{LLL}, & \text{KKK}, & \text{MMM}, & \text{FFF}, & \text{PPP}, & \text{SIS}, & \text{NST}, & \text{TTC}, & \text{QPK} \\ \text{ARW}, & \text{WWD}, & \text{MKY}, & \text{QYE}, & \text{YGL}, & \text{HPV}, & \text{VFR}, & \text{EAG}, & \text{KLQ}, & \text{SVT}, \\ \text{DHF}, & \text{WMP} & & & & & & & & \end{array}\right\}$$

- Then five iterations of our algorithm give this resolving set for $H_{8,20}$:

$$\left\{\begin{array}{ccccc} \text{AAAAAAAA}, & \text{AAAAAAAR}, & \text{AAAAAARA}, & \text{AAAAARAA}, & \text{AAAARAAA}, \\ \text{AAARAAAA}, & \text{ARWAAAAA}, & \text{CCCHHHHH}, & \text{CCCHHHHI}, & \text{CCCHHHIA}, \\ \text{CCCHHIAA}, & \text{CCCHIAAA}, & \text{CCCIAAAA}, & \text{CNSAAAAA}, & \text{DDDEEEEE}, \\ \text{DDDEEEEG}, & \text{DDDEEEGA}, & \text{DDDEEGAA}, & \text{DDDEGAAA}, & \text{DDDGAAAA}, \\ \text{DHFAAAAA}, & \text{EAGAAAAA}, & \text{EEEFAAAA}, & \text{EEEMFAAA}, & \text{EEEMMFAA}, \\ \text{EEEMMMFA}, & \text{EEEMMMMF}, & \text{EEEMMMMM}, & \text{FFFAAAAA}, & \text{GGGPPPPP}, \\ \text{GGGPPPPS}, & \text{GGGPPPSA}, & \text{GGGPPSAA}, & \text{GGGPSAAA}, & \text{GGGSAAAA}, \\ \text{HHHTTTTT}, & \text{HHHTTTTW}, & \text{HHHTTTWA}, & \text{HHHTTWAA}, & \text{HHHTWAAA}, \\ \text{HHHWAAAA}, & \text{HPVAAAAA}, & \text{IIIVAAAA}, & \text{IIIYVAAA}, & \text{IIIYYVAA}, \\ \text{IIIYYYVA}, & \text{IIIYYYYV}, & \text{IIIYYYYY}, & \text{KKKAAAAA}, & \text{KLQAAAAA}, \\ \text{LLLAAAAA}, & \text{MKYAAAAA}, & \text{MMMAAAAA}, & \text{NNNCCCCC}, & \text{NNNCCCCQ}, \\ \text{NNNCCCQA}, & \text{NNNCCQAA}, & \text{NNNCQAAA}, & \text{NNNQAAAA}, & \text{NSTAAAAA}, \\ \text{PPPAAAAA}, & \text{QPKAAAAA}, & \text{QQQKAAAA}, & \text{QQQLKAAA}, & \text{QQQLLKAA}, \\ \text{QQQLLLKA}, & \text{QQQLLLLK}, & \text{QQQLLLLL}, & \text{QYEAAAAA}, & \text{RRRDAAAA}, \\ \text{RRRNDAAA}, & \text{RRRNNDAA}, & \text{RRRNNNDA}, & \text{RRRNNNND}, & \text{RRRNNNNN}, \\ \text{SISAAAAA}, & \text{SVTAAAAA}, & \text{TTCAAAAA}, & \text{VFRAAAAA}, & \text{WMPAAAAA}, \\ \text{WWDAAAAA}, & \text{YGLAAAAA} & & & \end{array}\right\}$$

Back to Octamers

- ICH gives the following resolving set for $H_{3,20}$:

$$\left\{\begin{array}{cccccccccc} \text{AAA}, & \text{RRR}, & \text{NNN}, & \text{DDD}, & \text{CCC}, & \text{QQQ}, & \text{EEE}, & \text{GGG}, & \text{HHH}, & \text{CNS}, \\ \text{III}, & \text{LLL}, & \text{KKK}, & \text{MMM}, & \text{FFF}, & \text{PPP}, & \text{SIS}, & \text{NST}, & \text{TTC}, & \text{QPK} \\ \text{ARW}, & \text{WWD}, & \text{MKY}, & \text{QYE}, & \text{YGL}, & \text{HPV}, & \text{VFR}, & \text{EAG}, & \text{KLQ}, & \text{SVT}, \\ \text{DHF}, & \text{WMP} & & & & & & & & \end{array}\right\}$$

- Then five iterations of our algorithm give this resolving set for $H_{8,20}$:

$$\left\{\begin{array}{ccccc} \text{AAAAAAAA}, & \text{AAAAAAAR}, & \text{AAAAAARA}, & \text{AAAAARAA}, & \text{AAAARAAA}, \\ \text{AAARAAAA}, & \text{ARWAAAAA}, & \text{CCCHHHHH}, & \text{CCCHHHHI}, & \text{CCCHHHIA}, \\ \text{CCCHHIAA}, & \text{CCCHIAAA}, & \text{CCCIAAAA}, & \text{CNSAAAAA}, & \text{DDDEEEEE}, \\ \text{DDDEEEEG}, & \text{DDDEEEGA}, & \text{DDDEEGAA}, & \text{DDDEGAAA}, & \text{DDDGAAAA}, \\ \text{DHFAAAAA}, & \text{EAGAAAAA}, & \text{EEEFAAAA}, & \text{EEEMFAAA}, & \text{EEEMMFAA}, \\ \text{EEEMMMFA}, & \text{EEEMMMMF}, & \text{EEEMMMMM}, & \text{FFFAAAAA}, & \text{GGGPPPPP}, \\ \text{GGGPPPPS}, & \text{GGGPPPSA}, & \text{GGGPPSAA}, & \text{GGGPSAAA}, & \text{GGGSAAAA}, \\ \text{HHHTTTTT}, & \text{HHHTTTTW}, & \text{HHHTTTWA}, & \text{HHHTTWAA}, & \text{HHHTWAAA}, \\ \text{HHHWAAAA}, & \text{HPVAAAAA}, & \text{IIIVAAAA}, & \text{IIIYVAAA}, & \text{IIIYYVAA}, \\ \text{IIIYYYVA}, & \text{IIIYYYYV}, & \text{IIIYYYYY}, & \text{KKKAAAAA}, & \text{KLQAAAAA}, \\ \text{LLLAAAAA}, & \text{MKYAAAAA}, & \text{MMMAAAAA}, & \text{NNNCCCCC}, & \text{NNNCCCCQ}, \\ \text{NNNCCCQA}, & \text{NNNCCQAA}, & \text{NNNCQAAA}, & \text{NNNQAAAA}, & \text{NSTAAAAA}, \\ \text{PPPAAAAA}, & \text{QPKAAAAA}, & \text{QQQKAAAA}, & \text{QQQLKAAA}, & \text{QQQLLKAA}, \\ \text{QQQLLLKA}, & \text{QQQLLLLK}, & \text{QQQLLLLL}, & \text{QYEAAAAA}, & \text{RRRDAAAA}, \\ \text{RRRNDAAA}, & \text{RRRNNDAA}, & \text{RRRNNNDA}, & \text{RRRNNNND}, & \text{RRRNNNNN}, \\ \text{SISAAAAA}, & \text{SVTAAAAA}, & \text{TTCAAAAA}, & \text{VFRAAAAA}, & \text{WMPAAAAA}, \\ \text{WWDAAAAA}, & \text{YGLAAAAA} & & & \end{array}\right\}$$

- i.e. octamers may be represented as 82-dimensional vectors!

WORK IN PROGRESS

Question.

Is the metric dimension problem still NP-complete on Hamming graphs?

Question.

Is the metric dimension problem still NP-complete on Hamming graphs?

ANSWER: We do not know yet!

Question.

Is the metric dimension problem still NP-complete on Hamming graphs?

ANSWER: We do not know yet!

Problem.

Find characterizations of resolvability in $H_{k,a}$

Question.

Is the metric dimension problem still NP-complete on Hamming graphs?

ANSWER: We do not know yet!

Problem.

Find characterizations of resolvability in $H_{k,a}$

Observation.

Brute force is not practical to determine whether a set R of k-mers resolves or not $H_{k,a}$ because, for each of the $\Theta(a^{2k})$ pairs (x, y) of k-mers, one would need to check if there is $r \in R$ such that $d(x, r) \neq d(y, r)$

Lower-complexity Characterization of Resolvability

- $Q_k := H_{k,a=2}$ (**k-dimensional hypercube**)

k	1	2	3	4	5	6	7	8	9	10	*11*	*12*	*13*	*14*	*15*	*16*	*17*
$\beta(Q_k)$	1	2	3	4	4	5	6	6	7	7	*8*	*8*	*8*	*9*	*9*	*10*	*10*

Table: Exact [1] up to $k = 10$, conjectured [2] up to $k = 17$

[1] Harary and Melter - 1976

[2] Mladenović, Kratica, Kovačević-Vujčić et al. - 2012

[3] Beardon - 2013

Lower-complexity Characterization of Resolvability

- $Q_k := H_{k,a=2}$ (**k-dimensional hypercube**)

k	1	2	3	4	5	6	7	8	9	10	*11*	*12*	*13*	*14*	*15*	*16*	*17*
$\beta(Q_k)$	1	2	3	4	4	5	6	6	7	7	*8*	*8*	*8*	*9*	*9*	*10*	*10*

Table: Exact [1] up to $k = 10$, conjectured [2] up to $k = 17$

- There is already a resolvability characterization of hypercubes [3]

[1] Harary and Melter - 1976

[2] Mladenović, Kratica, Kovačević-Vujčić et al. - 2012

[3] Beardon - 2013

Lower-complexity Characterization of Resolvability

- $Q_k := H_{k,a=2}$ (**k-dimensional hypercube**)

k	1	2	3	4	5	6	7	8	9	10	*11*	*12*	*13*	*14*	*15*	*16*	*17*
$\beta(Q_k)$	1	2	3	4	4	5	6	6	7	7	*8*	*8*	*8*	*9*	*9*	*10*	*10*

Table: Exact [1] up to $k = 10$, conjectured [2] up to $k = 17$

- There is already a resolvability characterization of hypercubes [3]
- ... but Hamming graphs are **transitive**

Theorem: (with Laird, Tillquist & Becker)

Suppose $1^k \in R$ and define $A := (\text{—}r\text{—})_{r \in R}$. Then, R is resolves Q_k iff there does not exist a non-zero $z \in \{0, \pm 1\}^k$ such that $Az = 0$.

[1] Harary and Melter - 1976

[2] Mladenović, Kratica, Kovačević-Vujčić et al. - 2012

[3] Beardon - 2013

Theorem: (with Laird, Tillquist & Becker)

Let $v_1, \ldots, v_n$ be k-mers with **one-hot encodings** $V_1, \ldots, V_n$, respectively. If

$$A := \begin{pmatrix} — \ \mathrm{vec}(V_1) \ — \\ \vdots \\ — \ \mathrm{vec}(V_n) \ — \end{pmatrix}$$

then $R = \{v_1, \ldots, v_n\}$ resolves $H_{k,a}$ iff $z = 0$ is the only solution to the linear system $Az = 0$ which satisfies this additional constraint: if z is decomposed into k non-overlapping blocks of dimension a as follows

$$z = \Big((z_1, \ldots, z_a), (z_{a+1}, \ldots, z_{2a}), \ldots, (z_{(k-1)a+1}, \ldots, z_{ka}) \Big)^T,$$

then each block is the difference of two canonical vectors in $\mathbb{R}^a$.

Can we throw points away keeping a set resolving?

Using this new characterization:

- Integer linear programming (ILP) implies that

$$\left\{\begin{array}{lllll}
\sout{\text{AAAAAAAA}}, & \sout{\text{AAAAAAAR}}, & \sout{\text{AAAAAARA}}, & \sout{\text{AAAAARAA}}, & \sout{\text{AAAARAAA}}, \\
\text{AAARAAAA}, & \text{ARWAAAAA}, & \text{CCCHHHHH}, & \text{CCCHHHHI}, & \text{CCCHHHIA}, \\
\text{CCCHHIAA}, & \text{CCCHIAAA}, & \text{CCCIAAAA}, & \text{CNSAAAAA}, & \text{DDDEEEEE}, \\
\text{DDDEEEEG}, & \text{DDDEEEGA}, & \text{DDDEEGAA}, & \text{DDDEGAAA}, & \text{DDDGAAAA}, \\
\text{DHFAAAAA}, & \text{EAGAAAAA}, & \text{EEEFAAAA}, & \text{EEEMFAAA}, & \text{EEEMMFAA}, \\
\text{EEEMMMFA}, & \text{EEEMMMMF}, & \text{EEEMMMMM}, & \text{FFFAAAAA}, & \text{GGGPPPPP}, \\
\text{GGGPPPPS}, & \text{GGGPPPSA}, & \text{GGGPPSAA}, & \text{GGGPSAAA}, & \text{GGGSAAAA}, \\
\text{HHHTTTTT}, & \text{HHHTTTTW}, & \text{HHHTTTWA}, & \text{HHHTTWAA}, & \text{HHHTWAAA}, \\
\text{HHHWAAAA}, & \text{HPVAAAAA}, & \text{IIIVAAAA}, & \text{IIIYVAAA}, & \text{IIIYYVAA}, \\
\text{IIIYYYVA}, & \text{IIIYYYYV}, & \text{IIIYYYYY}, & \text{KKKAAAAA}, & \text{KLQAAAAA}, \\
\text{LLLAAAAA}, & \text{MKYAAAAA}, & \text{MMMAAAAA}, & \text{NNNCCCCC}, & \text{NNNCCCCQ}, \\
\text{NNNCCCQA}, & \text{NNNCCQAA}, & \text{NNNCQAAA}, & \text{NNNQAAAA}, & \text{NSTAAAAA}, \\
\text{PPPAAAAA}, & \text{QPKAAAAA}, & \text{QQQKAAAA}, & \text{QQQLKAAA}, & \text{QQQLLKAA}, \\
\text{QQQLLLKA}, & \text{QQQLLLLK}, & \text{QQQLLLLL}, & \text{QYEAAAAA}, & \text{RRRDAAAA}, \\
\text{RRRNDAAA}, & \text{RRRNNDAA}, & \text{RRRNNNDA}, & \text{RRRNNNND}, & \text{RRRNNNNN}, \\
\text{SISAAAAA}, & \text{SVTAAAAA}, & \text{TTCAAAAA}, & \text{VFRAAAAA}, & \text{WMPAAAAA}, \\
\text{WWDAAAAA}, & \text{YGLAAAAA} & & &
\end{array}\right\}.$$

resolves $H_{8,20}$ with high probability.

Can we throw points away keeping a set resolving?

Using this new characterization:

- Integer linear programming (ILP) implies that

$$\left\{\begin{array}{lllll}
\sout{\text{AAAAAAAA}}, & \sout{\text{AAAAAAAR}}, & \sout{\text{AAAAAARA}}, & \sout{\text{AAAAARAA}}, & \sout{\text{AAAARAAA}}, \\
\text{AAARAAAA}, & \text{ARWAAAAA}, & \text{CCCHHHHH}, & \text{CCCHHHHI}, & \text{CCCHHHIA}, \\
\text{CCCHHIAA}, & \text{CCCHIAAA}, & \text{CCCIAAAA}, & \text{CNSAAAAA}, & \text{DDDEEEEE}, \\
\text{DDDEEEEG}, & \text{DDDEEEGA}, & \text{DDDEEGAA}, & \text{DDDEGAAA}, & \text{DDDGAAAA}, \\
\text{DHFAAAAA}, & \text{EAGAAAAA}, & \text{EEEFAAAA}, & \text{EEEMFAAA}, & \text{EEEMMFAA}, \\
\text{EEEMMMFA}, & \text{EEEMMMMF}, & \text{EEEMMMMM}, & \text{FFFAAAAA}, & \text{GGGPPPPP}, \\
\text{GGGPPPPS}, & \text{GGGPPPSA}, & \text{GGGPPSAA}, & \text{GGGPSAAA}, & \text{GGGSAAAA}, \\
\text{HHHTTTTT}, & \text{HHHTTTTW}, & \text{HHHTTTWA}, & \text{HHHTTWAA}, & \text{HHHTWAAA}, \\
\text{HHHWAAAA}, & \text{HPVAAAAA}, & \text{IIIVAAAA}, & \text{IIIYVAAA}, & \text{IIIYYVAA}, \\
\text{IIIYYYVA}, & \text{IIIYYYYV}, & \text{IIIYYYYY}, & \text{KKKAAAAA}, & \text{KLQAAAAA}, \\
\text{LLLAAAAA}, & \text{MKYAAAAA}, & \text{MMMAAAAA}, & \text{NNNCCCCC}, & \text{NNNCCCCQ}, \\
\text{NNNCCCQA}, & \text{NNNCCQAA}, & \text{NNNCQAAA}, & \text{NNNQAAAA}, & \text{NSTAAAAA}, \\
\text{PPPAAAAA}, & \text{QPKAAAAA}, & \text{QQQKAAAA}, & \text{QQQLKAAA}, & \text{QQQLLKAA}, \\
\text{QQQLLLKA}, & \text{QQQLLLLK}, & \text{QQQLLLLL}, & \text{QYEAAAAA}, & \text{RRRDAAAA}, \\
\text{RRRNDAAA}, & \text{RRRNNDAA}, & \text{RRRNNNDA}, & \text{RRRNNNND}, & \text{RRRNNNNN}, \\
\text{SISAAAAA}, & \text{SVTAAAAA}, & \text{TTCAAAAA}, & \text{VFRAAAAA}, & \text{WMPAAAAA}, \\
\text{WWDAAAAA}, & \text{YGLAAAAA} & & &
\end{array}\right\}.$$

  resolves $H_{8,20}$ with high probability.

- Gröbner bases imply this set resolves $H_{8,20}$ with certainty

Can we throw points away keeping a set resolving?

Using this new characterization:

- Integer linear programming (ILP) implies that

$$\left\{\begin{array}{lllll}
\sout{\text{AAAAAAAA}}, & \sout{\text{AAAAAAAR}}, & \sout{\text{AAAAAARA}}, & \sout{\text{AAAAARAA}}, & \sout{\text{AAAARAAA}}, \\
\text{AAARAAAA}, & \text{ARWAAAAA}, & \text{CCCHHHHH}, & \text{CCCHHHHI}, & \text{CCCHHHIA}, \\
\text{CCCHHIAA}, & \text{CCCHIAAA}, & \text{CCCIAAAA}, & \text{CNSAAAAA}, & \text{DDDEEEEE}, \\
\text{DDDEEEEG}, & \text{DDDEEEGA}, & \text{DDDEEGAA}, & \text{DDDEGAAA}, & \text{DDDGAAAA}, \\
\text{DHFAAAAA}, & \text{EAGAAAAA}, & \text{EEEFAAAA}, & \text{EEEMFAAA}, & \text{EEEMMFAA}, \\
\text{EEEMMMFA}, & \text{EEEMMMMF}, & \text{EEEMMMMM}, & \text{FFFAAAAA}, & \text{GGGPPPPP}, \\
\text{GGGPPPPS}, & \text{GGGPPPSA}, & \text{GGGPPSAA}, & \text{GGGPSAAA}, & \text{GGGSAAAA}, \\
\text{HHHTTTTT}, & \text{HHHTTTTW}, & \text{HHHTTTWA}, & \text{HHHTTWAA}, & \text{HHHTWAAA}, \\
\text{HHHWAAAA}, & \text{HPVAAAAA}, & \text{IIIVAAAA}, & \text{IIIYVAAA}, & \text{IIIYYVAA}, \\
\text{IIIYYYVA}, & \text{IIIYYYYV}, & \text{IIIYYYYY}, & \text{KKKAAAAA}, & \text{KLQAAAAA}, \\
\text{LLLAAAAA}, & \text{MKYAAAAA}, & \text{MMMAAAAA}, & \text{NNNCCCCC}, & \text{NNNCCCCQ}, \\
\text{NNNCCCQA}, & \text{NNNCCQAA}, & \text{NNNCQAAA}, & \text{NNNQAAAA}, & \text{NSTAAAAA}, \\
\text{PPPAAAAA}, & \text{QPKAAAAA}, & \text{QQQKAAAA}, & \text{QQQLKAAA}, & \text{QQQLLKAA}, \\
\text{QQQLLLKA}, & \text{QQQLLLLK}, & \text{QQQLLLLL}, & \text{QYEAAAAA}, & \text{RRRDAAAA}, \\
\text{RRRNDAAA}, & \text{RRRNNDAA}, & \text{RRRNNNDA}, & \text{RRRNNNND}, & \text{RRRNNNNN}, \\
\text{SISAAAAA}, & \text{SVTAAAAA}, & \text{TTCAAAAA}, & \text{VFRAAAAA}, & \text{WMPAAAAA}, \\
\text{WWDAAAAA}, & \text{YGLAAAAA} & & &
\end{array}\right\}.$$

  resolves $H_{8,20}$ with high probability.

- Gröbner bases imply this set resolves $H_{8,20}$ with certainty
- So octamers may be represented as 77-dimensional vectors!

COLLABORATORS:

Lucas Laird
APPM/CSCI

Richard C. Tillquist
IQ Bio/CSCI

Stephen Becker
APPM

FUNDING:

- BioFrontiers Institute Computing Core
- NSF Grant No. 1836914 (PI)

... Thank You!

Proof of Concept

- Fruit fly genome ($\sim 1.75 \times 10^8$ base-pairs)
- Problem: Characterize 20-mers centered at intron-exon boundaries
- Pos. examples: ~87K; Neg. examples: Random 20-mers

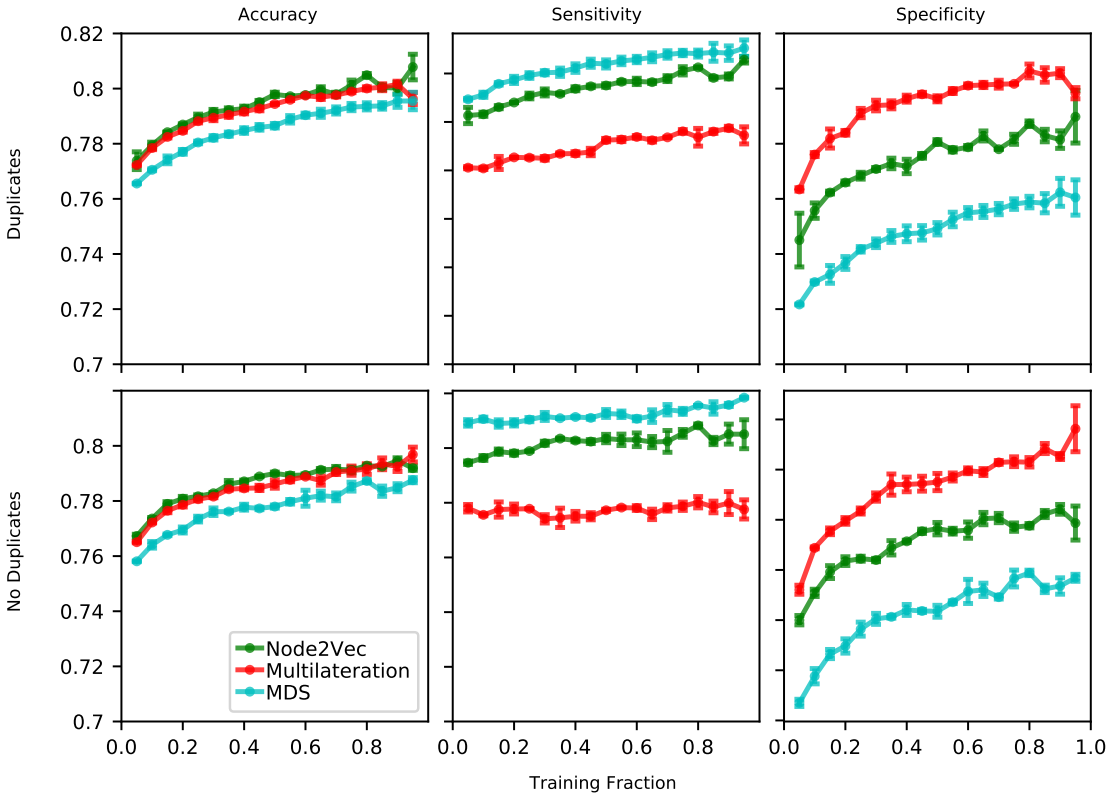


Figure: Cross-validation of KNN with 3 different embeddings