Are Multidimensional Boolean Patterns Dominating Microbiome and Microbial Genome Data?

George Golovko
University of Texas Medical Branch (Galveston)
Kamil Khanipov
University of Texas Medical Branch (Galveston)
Victor Reyes
University of Texas Medical Branch (Galveston)
Irina Pinchuk
Penn State Health Milton Hershey Medical Center
Yuriy Fofanov (yufofano@utmb.edu)
University of Texas Medical Branch (Galveston)

Method Article

Keywords: Microbiome, Multidimensional patterns, Boolean patterns, Multiomics, Co-exclusion, Co-presence, Regulatory Network

Posted Date: April 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1506825/v1

Are Multidimensional Boolean Patterns Dominating Microbiome and Microbial Genome Data?

George Golovko, Kamil Khanipov, Victor Reyes, Irina Pinchuk, Yuriy Fofanov*

Keywords: Microbiome, Multidimensional patterns, Boolean patterns, Multiomics, Co-exclusion, Co-presence, Regulatory Network.

Abstract

Background

Virtually every biological system is governed by the complex relations among its components. Identifying such relations requires a rigorous or heuristics-based search for *patterns* among *variables/features* of a system. A number of algorithms have been developed to identify two-dimensional (involving two *variables*) *patterns* employing correlation, covariation, mutual information, etc. It seems obvious, however, that comprehensive descriptions of complex biological systems may also include more complicated *multidimensional* relations, which can only be described using patterns that simultaneously embrace 3, 4, and more variables. The main challenges in the search for such *multidimensional patterns* include: (a) computational complexity of the search; (b) distinction of statistically significant patterns from false patterns which can be observed in large data sets simply by chance; and (3) integration of heterogeneous data types (numerical, Boolean, categorical, etc.) in a single pattern.

Results

This manuscript presents an attempt to address some of these challenges by defining *multidimensional Boolean patterns* in a way permitting to: (a) accommodate heterogeneous multi-omics data, (b) formulate criteria for separating trivial from non-trivial patterns, and (c) identify conditions, required for a given pattern to predict the values of selected feature(s). Additionally, the proposed definition of the pattern's strength (pattern's score) and minimal population threshold permits estimation of the statistical significance of detected patterns using scores distributions of artificial datasets created by randomizing original data.

Conclusion

To test the proposed approach we performed a search for all possible 2-, 3-, and 4-dimensional patterns in historical data from the Human Microbiome Project (15 body sites) and collection of *H. pylori* genomes associated with gastric ulcers, gastritis, and duodenal ulcers. In all datasets under consideration, we were able to identify hundreds of statistically significant multidimensional patterns. These results suggest that such patterns may dominate the landscape of microbial genomics/microbiomics systems.

Background

Boolean patterns are routinely used to describe non-numerical relations between components (*features*) of complex systems, especially when relations cannot be described (or identified) by correlation, covariance,

and/or other numerical characteristics of statistical relationships [1]. For example, in environmental microbial communities (MC), the functions vital to all microorganisms (e.g., nitrogen fixation) are usually performed by a single, not highly abundant species, when the rest of the species are depending on it. As a result, the pattern describing relations between nitrogen fixing species and any other member of MC cannot be observed as a correlation/covariation but is rather expected to be manifested as Boolean *one-way relation* patterns[1,2], where the presence of each "dependent" species (features) requires the presence of another "provider" feature (nitrogen fixing species), but not vice versa (Figure 1c). Similarly, other Boolean pairwise relations, such as *co-presence* and *co-exclusion,* may be represented as Boolean patterns (Figure 1a, b). In fact, out of 16 (2^4) possible combinations of the presence/absence profiles between two variables, only four may be interpreted as possible relations: *co-presence*, *co-exclusion*, and two *one-way relations* (variable 1 needs variable 2 to be present and the opposite: variable 2 needs variable 1 to be present).

In the previous work[1] we were able to introduce several examples of 2- and 3-dimensional patterns and demonstrate their presence in microbiome data. The focus of this manuscript is to present a more comprehensive examination of the properties of multidimensional patterns; present an approach which allows estimation of the statistical significance of each individual pattern in conjunction with appropriate selection of the presence/absence and minimum population thresholds; as well as demonstrate the presence of a large number of statistically significant 2-,3-, and 4-dimensional patterns in microbial genome and microbiome datasets.

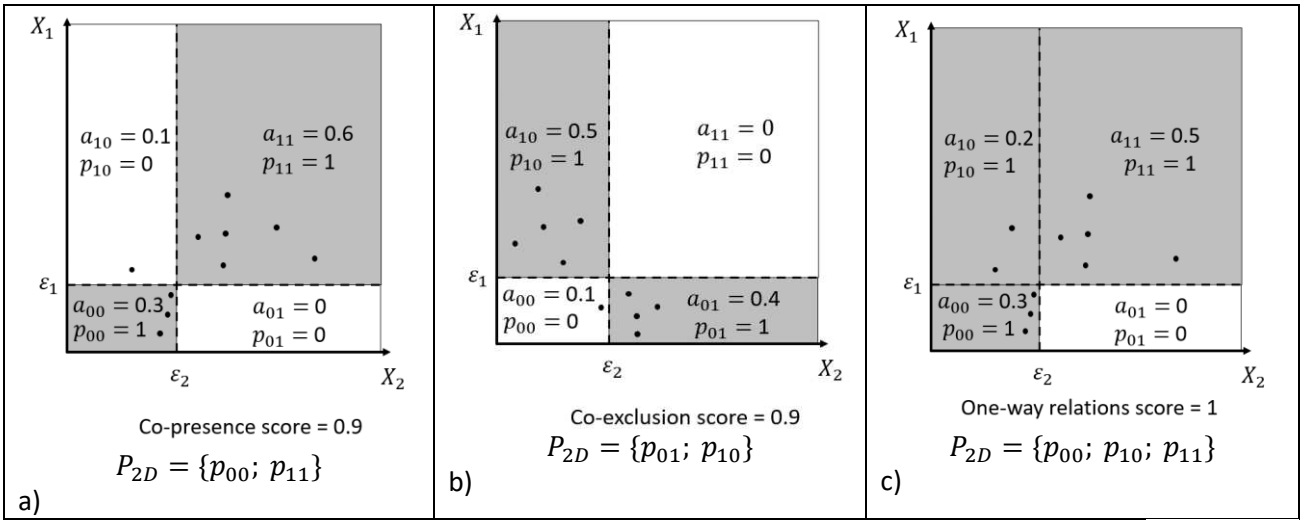

Figure 1. Boolean two-dimensional patterns: (a) *co-presence*; (b) *co-exclusion*; and (c) *one-way relations*. Here ε_1 and ε_2 are presence/absence thresholds; a_{00}, a_{01}, a_{10}, and a_{11} - the proportion of points (observation) located in each partition (p_{00}, p_{01}, p_{10}, and p_{11}). Grey color and pattern description (P_{2D}) denote partitions require to be present in the pattern.

METHODS

Multidimensional Boolean patterns

In general terms, the Boolean pattern (P) can be defined as a presence/absence (0/1) profile of Boolean variables $\{p_{ij}, \ldots\}$ associated with each combination of indices ($i, j, k, \ldots$), thus for two variables, the 2D pattern can be defined as:

$$P_{2D} = \left\{p_{ij} = \{0,1\}; i = \{0,1\}, j = \{0,1\}\right\};$$

For example, the *co-presence* and *co-exclusion* patterns (Figure 1a, b) can be defined by the Boolean presence (1) or absence (0) values in all four partitions of the 2D space:

$$P_{2D\ co-presence} = \{p_{00} = 1; p_{01} = 0; p_{10} = 0; p_{11} = 1;\};$$

$$P_{2D\ co-exclusion} = \{p_{00} = 0; p_{01} = 1; p_{10} = 1; p_{11} = 0;\}$$

Similarly, for higher dimensions:

$$P_{3D} = \left\{p_{ijk} = \{0,1\}; i = \{0,1\}, j = \{0,1\}, k = \{0,1\}\right\};$$

$$P_{4D} = \left\{p_{ijkl} = \{0,1\}; i = \{0,1\}, j = \{0,1\}, k = \{0,1\}, l = \{0,1\}\right\};$$

In the same way, 3D *co-presence* (Figure 2a) can be defined as:

$$P_{3D\ co-presence} = \{p_{000} = 1; p_{001} = 0; p_{010} = 0; p_{011} = 0; p_{100} = 0; p_{101} = 0; p_{110} = 0; p_{111} = 1;\};$$

Interestingly, 3D space allows to define two different types (three in 4D space) of *co-exclusion* patterns (Figure 2b, c):

$$P_{3D\ Type\ 1\ co-exclusion} = \{p_{000} = 0; p_{001} = 1; p_{010} = 1; p_{011} = 0; p_{100} = 1; p_{101} = 0; p_{110} = 0; p_{111} = 0;\};$$

$$P_{3D\ Type\ 2\ co-exclusion} = \{p_{000} = 0; p_{001} = 0; p_{010} = 0; p_{011} = 1; p_{100} = 0; p_{101} = 1; p_{110} = 1; p_{111} = 0;\};$$

Pattern strength (Score)

The pattern strength (score) can be defined as the fraction of observations belonging to the pattern. For example, for 2D patterns, the pattern strength (score) will depend on the fraction of the experimental observations in four partitions of the two-dimensional space: $a_{00},\ a_{01},\ a_{01},\ a_{11}$ (Figure 1a-c).

$$S = \sum_{i=0}^{1} \sum_{j=0}^{1} p_{ij} a_{ij};$$

Several approaches can be used to transform abundance values to the presence/absence profile [3]. We, however, believe that since the same feature may be involved in multiple processes, each of which may require it to be present in different abundance, the identification of the presence/absence threshold must be specific for each pattern and feature combination and require threshold (ε_i) optimization:

$$S_{2D} = \max_{\varepsilon_1, \varepsilon_2} \left(\sum_{i=0}^{1} \sum_{j=0}^{1} p_{ij} a_{ij}\right);$$

$$S_{3D} = \max_{\varepsilon_1, \varepsilon_2, \varepsilon_3} \left(\sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{k=0}^{1} p_{ijk} a_{ijk}\right);$$

$$S_{4D} = \max_{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4} \left(\sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{k=0}^{1} \sum_{l=0}^{1} p_{ijkl} a_{ijkl}\right);$$

While threshold optimization provides the possibility that different variables may have different and even pattern-specific thresholds, it can also produce some misleading results. This situation arises, for example, if partitions required to be present in the pattern are empty or contain a very small number of observations.

This effect can be eliminated by establishing a predefined *minimal presence threshold* (d)[1] for a proportion of the observations ($a_{ij\ldots}$) in each quadrant present in the pattern ($p_{ij\ldots}=1$).

Patterns properties

Patterns - SubPatterns Consistency

Each n-dimensional pattern can be decomposed into a combination of $2n$ patterns of lower $(n-1)$ dimensions. Such subpatterns can be pictured as "slices" of an n dimensional pattern for each specific value of one of the variables. For example, each 3D pattern

$$P_{3D}=\left\{p_{ijk}=\{0,1\};i=\{0,1\},j=\{0,1\},k=\{0,1\}\right\};$$

contains six 2D subpatterns:

$$P_{2D\ i=0}=\left\{p_{ijk}=\{0,1\};i=0,j=\{0,1\},k=\{0,1\}\right\};$$
$$P_{2D\ i=1}=\left\{p_{ijk}=\{0,1\};i=1,j=\{0,1\},k=\{0,1\}\right\};$$
$$P_{2D\ j=0}=\left\{p_{ijk}=\{0,1\};i=\{0,1\},j=0,k=\{0,1\}\right\};$$
$$P_{2D\ j=1}=\left\{p_{ijk}=\{0,1\};i=\{0,1\},j=1,k=\{0,1\}\right\};$$
$$P_{2D\ k=0}=\left\{p_{ijk}=\{0,1\};i=\{0,1\},j=\{0,1\},k=0\right\};$$
$$P_{2D\ k=1}=\left\{p_{ijk}=\{0,1\};i=\{0,1\},j=\{0,1\},k=1\right\};$$

It is important to emphasize, however, that not every combination of $n-1$ dimensional patterns expressed between n variables are possible. Each combination of $n-1$ dimensional patterns leads to three possible outcomes:

1. The combination of $n-1$ dimensional patterns leads to a unique n-dimensional pattern;
2. The combination of $n-1$ dimensional patterns leads to several n-dimensional patterns;
3. The combination of $n-1$ dimensional patterns leads to no n-dimensional pattern, thereby being inconsistent.
4. For example, the following set of 2D patterns is inconsistent (cannot exist) because the corresponding 3D pattern is impossible.

5. $P_{2D\ i=0}=\{p_{000}=1;p_{001}=0;p_{010}=0;p_{011}=0;\};$
6. $P_{2D\ i=1}=\{p_{100}=0;p_{101}=0;p_{110}=0;p_{111}=0;\};$
7. $P_{2D\ j=0}=\{p_{000}=0;p_{001}=0;p_{100}=0;p_{101}=0;\};$
8. $P_{2D\ j=1}=\{p_{010}=0;p_{011}=0;p_{110}=0;p_{111}=0;\};$
9. $P_{2D\ k=0}=\{p_{000}=0;p_{010}=0;p_{100}=0;p_{110}=0;\};$
10. $P_{2D\ k=1}=\{p_{001}=0;p_{011}=0;p_{101}=0;p_{111}=1;\};$

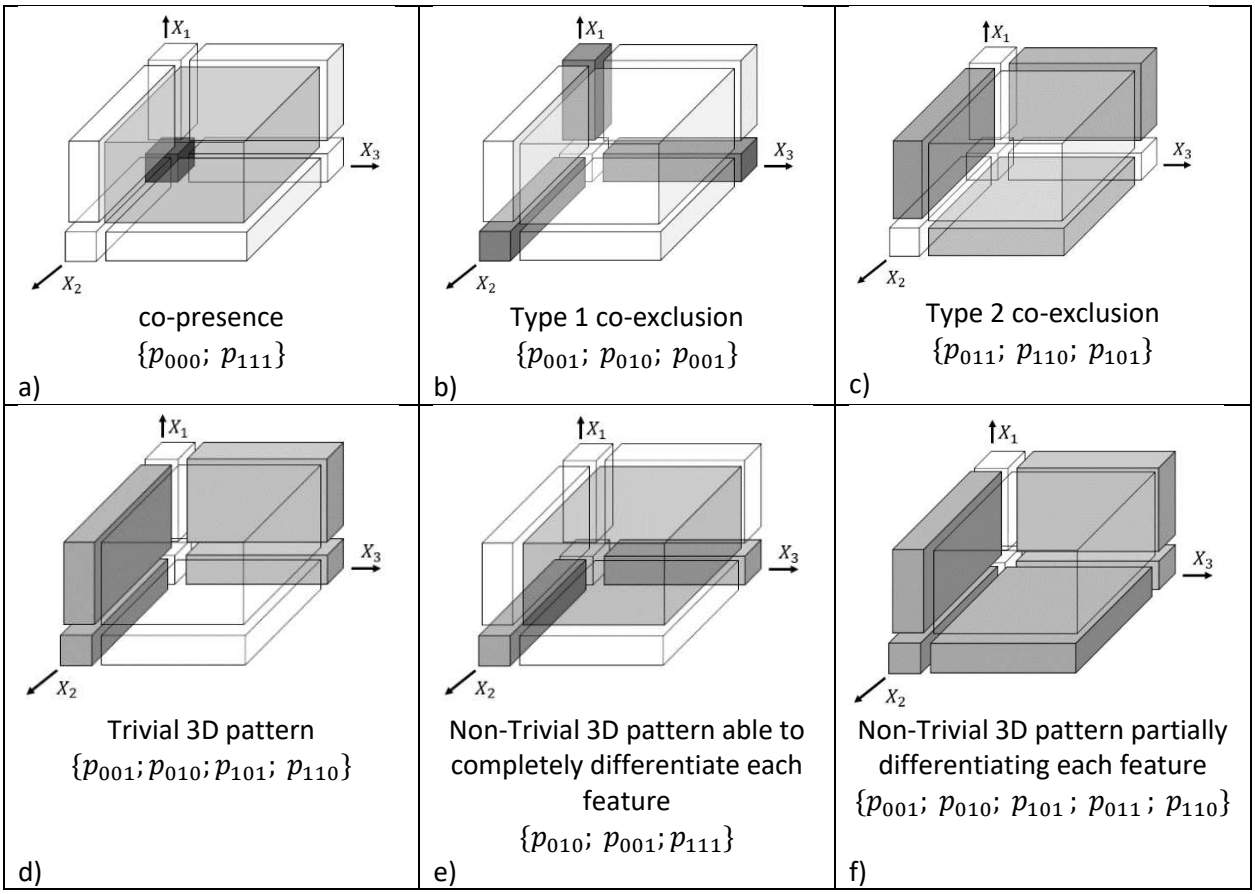

Figure 2. Examples of Boolean three-dimensional patterns. (a) *co-presence*, (b) *Type 1 co-exclusion*, (c) *Type 2 co-exclusion*, (d) Trivial 3D pattern, (e) Non-Trivial 3D pattern able to completely differentiate each feature, (f) Non-Trivial 3D pattern partially differentiating each feature.

Such a relationship between lower and higher dimensional patterns offers an interesting opportunity to evaluate the overall consistency of the set of patterns describing each given biological system and allows detection of contradictions between patterns. Additionally, the presence of higher dimensional patterns could be used to identify the presence of missed lower-dimensional subpatterns.

It is also important to mention the special role of the p_{00}, p_{000}, p_{0000} etc. partitions in the pattern definition. In some cases, for example when it represents co-absence of particular microorganisms in a microbial community, the observations in this partition can be interpreted as irrelevant (all features/values under consideration are absent) and do not contribute to the pattern strength. Conversely, for patterns involving features like SNPs, or gene/protein expression profiles, consideration of this partition will be necessary.

Trivial vs. non-trivial patterns

While the total number of n-dimensional patterns is 2^{2^n}, it is important to keep in mind that not all patterns can be interpreted as a possible relation between variables. For example, the 2D pattern in which $p_{ij} = 1$ for all combinations of i, j (all partitions are present) cannot be interpreted as any meaningful relation and can be considered as *trivial*. Two simple criteria can be used to identify (and exclude from future consideration) such trivial (not allowing any relation-like interpretation) patterns:

1. For a pattern to be *non-trivial,* each variable in it must be observed in both presence (1) and absence (0) states.

For example, in the case of a 2D pattern:

$$\sum_{j=0}^{1} p_{0j} > 0;\ \ \sum_{j=0}^{1} p_{1j} > 0;\ \ \sum_{i=0}^{1} p_{i0} > 0;\ \ \sum_{i=0}^{1} p_{i1} > 0;$$

Specifically, if a variable is always present or always absent in all observations, it cannot be part of any non-trivial pattern.

2. For a pattern to be *non-trivial, for* each variable its $(n-1)$ dimensional subpattern when this variable is present (1) must be different from the $(n-1)$ dimensional subpattern when this variable is absent (0). For example, 3D pattern:

$$P_{3D} = \left\{p_{ijk} = \{0,1\}; i = \{0,1\}, j = \{0,1\}, k = \{0,1\}\right\};$$

containing six 2D subpatterns:

$$P_{2D\ i=0} = \left\{p_{ijk} = \{0,1\}; i = 0, j = \{0,1\}, k = \{0,1\}\right\};$$
$$P_{2D\ i=1} = \left\{p_{ijk} = \{0,1\}; i = 1, j = \{0,1\}, k = \{0,1\}\right\};$$
$$P_{2D\ j=0} = \left\{p_{ijk} = \{0,1\}; i = \{0,1\}, j = 0, k = \{0,1\}\right\};$$
$$P_{2D\ j=1} = \left\{p_{ijk} = \{0,1\}; i = \{0,1\}, j = 1, k = \{0,1\}\right\};$$
$$P_{2D\ k=0} = \left\{p_{ijk} = \{0,1\}; i = \{0,1\}, j = \{0,1\}, k = 0\right\};$$
$$P_{2D\ k=1} = \left\{p_{ijk} = \{0,1\}; i = \{0,1\}, j = \{0,1\}, k = 1\right\},$$

will be considered non-trivial only if: $P_{2D\ i=0} \neq P_{2D\ i=1}\ and\ \ P_{2D\ j=0} \neq P_{2D\ j=1}\ and\ P_{2D\ k=0} \neq P_{2D\ k=1}$ simultaneously. Examples of trivial and non-trivial 3D patterns can be found in Figure 2: the pattern in Figure 2e is non-trivial, while the pattern in Figure 2d is trivial because 2D subpatterns are the same for both values of feature X_1.

One special type of multidimensional pattern-subpattern relationship occurs when the n-dimensional pattern is not trivial, but is a direct consequence of its subpatterns. For example, a 3D type 1 *co-exclusion* pattern (Figure 2b) may be considered a result of three 2D *co-exclusions*. Such patterns can be easily detected and excluded from consideration by requiring the n-dimensional pattern to have a unique (only one) way $(n-1)$-dimensional pattern representation. However, in the presented study, we decided to include such patterns so that all statistically significant (see definition below) n-dimensional patterns are included regardless of whether one or more of the $(n-1)$-dimensional subpatterns passed the statistical significance threshold.

Patterns which can distinguish selected features

Biomedical studies are often focused on the identification of features related to specific features of interest such as: combination of SNPs which can be predictive genetic markers of cancer; patterns in gut microbial composition associated with Crohn's disease; or combinations of environmental and genetic factors associated with the development of autism.

To identify whether a certain n-dimensional pattern can discriminate features of interest, one can perform a comparison between two $(n-1)$- dimensional subpatterns appearing for each value (0/1) of that feature. For example, the 2-dimensional pattern in Figure 1a can use the value of one feature (X_1) to "predict" the value of another (X_2). In contrast, some patterns can be used to discriminate selected features only for a specific condition (specific combination of features). Figure 1b shows an example where the presence of one feature ($X_1 = 1$) cannot be used to discriminate X_2: it can be either 0 or 1; but when $X_1 = 0$ the only possible value for X_2 is 0.

The following criteria/definition can be formulated to determine if a given pattern can completely differentiate the feature.

Feature (X_1) can be completely differentiated by pattern $P_{ijk\dots}$ if and only if $p_{0jk\dots} + p_{1jk\dots} \leq 1$ for every combination of j, k, etc. In other words, to completely discriminate the feature, the corresponding presence/absence values of the subpatterns appearing for each value (0/1) of this feature must never be co-present.

Partial feature discrimination can be defined by relaxing complete differentiation criteria:

a. *To be able to conditionally (partially) discriminate the presence of the feature 1, corresponding subpatterns ($P_{0jk\dots}$ and $P_{1jk\dots}$) must have at least one partition (the combination of j,k,…) where corresponding values $p_{0jk\dots} = 0$ and $p_{1jk\dots} = 1$ simultaneously.*
b. *To be able to conditionally (partially) discriminate the absence of the feature 1, corresponding subpatterns ($P_{0jk\dots}$ and $P_{1jk\dots}$) must have at least one partition (the combination of j,k,…) where corresponding values $p_{0jk\dots} = 1$ and $p_{1jk\dots} = 0$ simultaneously.*

Pattern Type	2D	3D	4D
Total number of patterns	16	256	65,536
Non-trivial patterns $p_{0..}$ included	6	174	62,224
Non-trivial patterns $p_{0..}$ excluded	4	96	31,479
Non-trivial patterns able to differentiate single feature ($p_{0..}$ included)	2	38	5,401
Non-trivial patterns able to differentiate single feature ($p_{0..}$ excluded)	1	15	1,909
Non-trivial patterns able to conditionally differentiate the feature's presence ($p_{0..}$ included)	4	136	56,823
Non-trivial patterns able to conditionally differentiate the feature's presence ($p_{0..}$ excluded)	2	66	27,770
Non-trivial patterns able to conditionally differentiate feature's absence ($p_{0..}$ included)	4	136	56,570
Non-trivial patterns able to conditionally differentiate feature's absence ($p_{0..}$ excluded)	3	81	29,522

Table 1. The number of different types of patterns in 2, 3, and 4 dimensions.

Figure 2e shows an example of the pattern that completely differentiates (discriminates) each feature. In contrast, the pattern in Figure 2f presents an example of complex partial discrimination: it can be used to predict the absence of X_1 if X_2 and X_3 are present; the presence of X_3 if X_1 and X_2 are absent; the presence of X_3 if X_1 is present and X_2 is absent; absence of X_3 if X_1 and X_2 are present; the presence of X_2 if X_1 and X_3 are absent; the presence of X_2 if X_1 is present and X_3 is absent; and absence of X_2 if X_1 and X_3 are present. Table 1 presents the summary of the discriminative properties of 2, 3, and 4-dimensional patterns. The complete list of all patterns and their characteristics is available in the Supplementary Table 1.

Patterns Search Implementation

The critical challenge of the search for 3, 4, and higher-dimensional patterns is the high computational complexity of the task. Depending on the implementation, the search for each pattern could reach the complexity of $O\left((fm)^n 2^{2^n}\right)$ where f is the number of features, m the number of samples, n is the pattern dimension, where $O\left(2^{2^n}\right)$ is the complexity associated to the number of patterns and $O((fm)^n)$ is the complexity of the search for the maximum score values. Even with basic parallelization by calculating each model for each combination of features and thresholds, the naïve straight forward implementation of the patterns search will require an enormous amount of computational power. However, by using several optimization steps, such as precomputing the optimization grid, early detection (and exclusion) of outliers and futures unable to participate in non-trivial patterns, as well as an implementation search for all possible patterns in a single step (since for each combination of features, *minimal presence threshold* (d) leads to a unique pattern), we were able to reduce the overall computational complexity of the process to $O((fm)^n)$.

All of the examples presented in this manuscript have been calculated using a single processor workstation, where the analysis of all 2D patterns takes 1-2 hours, 3D patterns 1-3 days and 4D patterns 1-2 weeks. For studies involving an extremely large number of samples (such as single cell mRNA sequencing) further reduction of computational complexity could be achieved by clustering both samples and features which will make computational complexity dependent only on the number of clusters, but not on the number of samples or features.

The search for the multidimensional patterns belongs to the category of multiple hypotheses testing algorithms, so there is significant concern that many patterns will be detected simply by chance. The fraction of such random patterns as well as the overall number of patterns detected depends on the data itself, as well as the choice of the minimal presence threshold (m) and required minimal pattern strength/score (S_{min}). To avoid making any assumptions regarding the type of features value distributions or making an arbitrary choice of the *minimal presence threshold* (d) and *minimal pattern score required* (S_{min}), one can perform several cycles of the patterns search in randomized (shuffled) datasets and identify the Pareto set (Pareto frontier) of (S_{min},d_{max}) pairs beyond which zero (on predefined minimum threshold number) patterns are detected in randomized data and use it as a criterion to select patterns in the real dataset under consideration.

It is necessary to mention that the shuffling method must reflect the underlying assumption about what is considered as the random alternative to the observed dataset (Zero Model) [4, 5]. To accommodate the scenario when some features may be represented by values always lower or always higher than average, the shuffling must be performed across individual feature profiles.

Example 1: Two-, Three-, and Four- dimensional patterns in Human Microbiome Project data

To evaluate the presented approach's ability to identify the presence of 2, 3, and 4-dimensional patterns, we performed analysis of 18 datasets associated with 16 body sites from the NIH Human Microbiome Project[6]. Since the purpose of the study was just a demonstration of the approach, we decided to use a relatively old collection of microbial profiles (December 2016). We downloaded 2,910 samples from the project website in text format (HMQCP–QIIME Community Profiling v13 OTU table). Samples representing a significantly low (less than 2,000) and significantly high (over 50,000) number of sequencing reads were excluded from the analysis. The microbial profiles of the remaining 2380 samples, varying from 67 for Posterior Fornix to 200 for Antecubital Fossa, were normalized against the total number of reads in each sample and transformed into relative abundance profiles merged to *genus* taxonomy level for each body site.

Analysis was performed for each body site individually. The sample profiles for each body site were normalized to make the sum of all abundances equal 100%. The OTUs outliers have been excluded using multiple random sampling of 50% of values from each OTU. The exclusion criteria were set to 25 standard deviations from average. After outliers were excluded, all the data were normalized again to make the sum of all abundances equal 100% for each sample. The outliers were treated as missing values. In pattern score calculations, all observations containing at least one of the missing values were ignored.

In the pattern search the population threshold (m) was the subject of optimization (maximization of the pattern score value) under the condition that each partition required to be present in the pattern is populated above the predefined which was set to equal 10%. Each original data set was randomized using the Mercer twister algorithm and a search for all 2, 3, and 4 dimensional patterns was performed 4 times (shuffling data again for each iteration). This step allowed us to establish a pareto set of pairs possible in random data, so that when the same analysis was performed on the real data, only patterns for which both m and S were higher than any combination observed in the randomized data was considered statistically significant. The original data as well as all the patterns detected can be found in the supplementary data.

Table 2 shows that in each analyzed dataset the total number of features (OTUs) involved in detected 3D and 4D patterns is larger than in 2D patterns. The number of detected higher dimensional patterns also are significantly larger, suggesting that multidimensional patterns could be dominating the overall landscape of the inter-organisms' relationship in these microbiomes. The decline in the number of detected 4D patterns can be caused by the relatively small number of samples preventing existing patterns from passing the statistical significance threshold. In some cases, such as $Posterior\ fornix$, only 2 OTUs appeared to be present in single 2D patterns, however including higher dimensions increased this number to 26.

Another example (Figure 3) shows patterns associated with the $Blautia$ genus in stool samples. Over recent decades, species of this genus have attracted a lot of attention in association to human health, obesity, etc. While there are no strong 2D patterns, this genus seems to be present in several strong ($S > 0.99$) 3D and 4D patterns.

Samples origin	Number of OTUs (genus level)	Number of patterns detected			Number of OUTs participating in patterns			Number of pattern types			Number of OTUs across all patterns
		2D	3D	4D	2D	3D	4D	2D	3D	4D	
Attached Keratinized gingiva	74	46	187	132	33	50	47	3	32	36	56
Anterior nares	129	11	43	11	11	28	25	2	15	4	40
Buccal mucosa	105	33	267	12	31	52	48	3	28	6	58
Hard palate	117	41	159	12	37	41	30	4	23	9	54
Retroauricular crease	108	39	101	66	28	36	28	4	29	29	49
Mid vagina	85	20	149	64	10	32	27	5	32	26	35
Posterior fornix	67	1	25	14	2	19	20	1	6	3	26
Palaine Tonsils	99	44	378	74	40	64	58	3	48	29	74
Saliva	113	35	77	10	41	57	19	3	24	4	69
Subgingival plaque	94	113	632	134	49	71	56	4	37	51	76
Supragingival plaque	90	80	928	562	45	69	72	4	43	151	77
Stool	99	53	511	222	36	56	70	4	34	78	75
Tongue dorsum	81	76	2219	176	47	62	56	4	50	97	64
Throat	115	51	468	31	40	64	46	3	36	12	69
Vaginal introitus	96	16	51	28	17	22	20	3	15	17	32

Table 2. The summary of patterns in 2, 3, and 4 dimensional patterns identified in the Human Microbiome Project data.

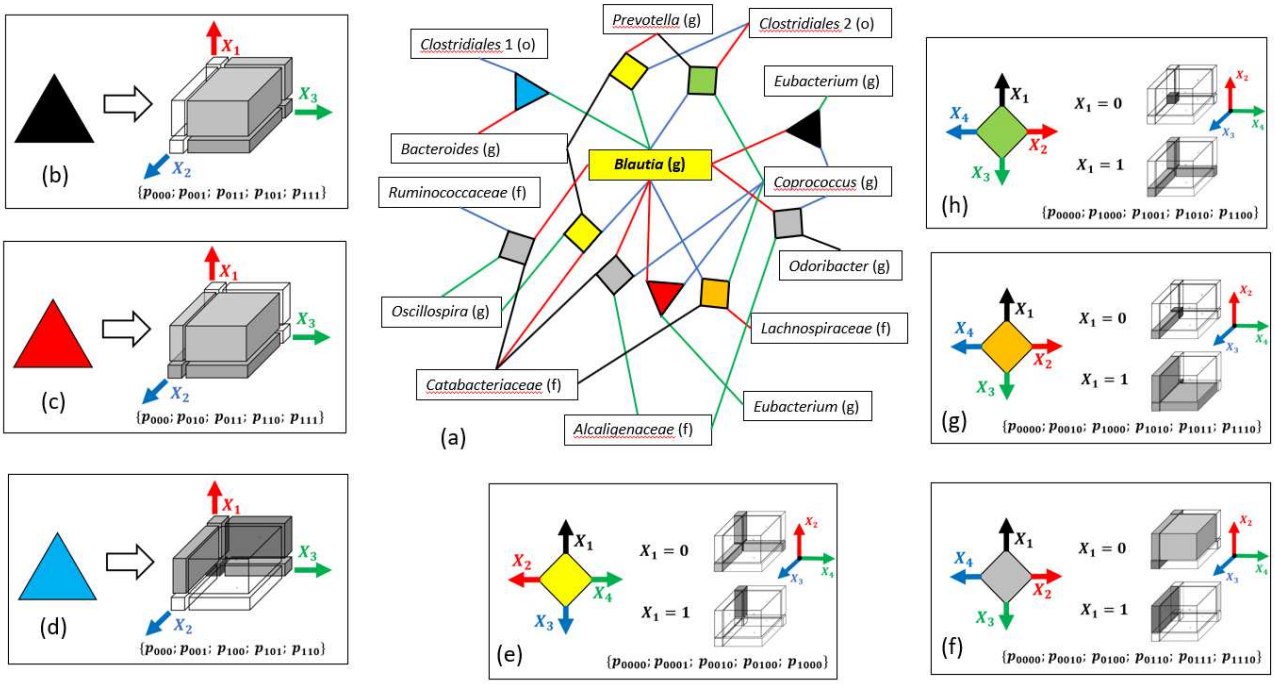

Figure 3. Three- and four-dimensional patterns containing *Blautia* genus in Human Microbiome Project stool samples. Labels (o), (g), and (f) represent order, genus, and family respectively. Each of the 3D and 4D patterns cannot be reduced to the combination of 2D patterns. Pattern (e) for example, shows that all four microorganisms cannot coexist in any combination.

Example 2: 3D Patterns in *H. pylori* Gene Composition Associated with Duodenal ulcers, Gastritis, and Gastric ulcers

Fully annotated *Helicobacter pylori* genomes (93 total) were downloaded from the PATRIC BRC database [7]. The database included 58 *H. pylori* genomes from chronic gastritis, 21 from gastric ulcers, and 13 from duodenal ulcers. The features of every genome were extracted and merged based upon their protein product annotation from the PATRIC BRC database. The unannotated features were excluded from the analysis as they could not be merged across samples based on the annotation identifier.

To demonstrate the possibilities of performing a search for multidimensional patterns we limited the focus of the analysis to identification of metadata (disease-type) associated patterns which included duodenal ulcers, gastritis, or gastric ulcers as one of the features. There were no 2D patterns with a minimal pattern score of 0.99 and a population threshold of 0.1 identified and there were 28 3D patterns (Figure 4) detected in the dataset.

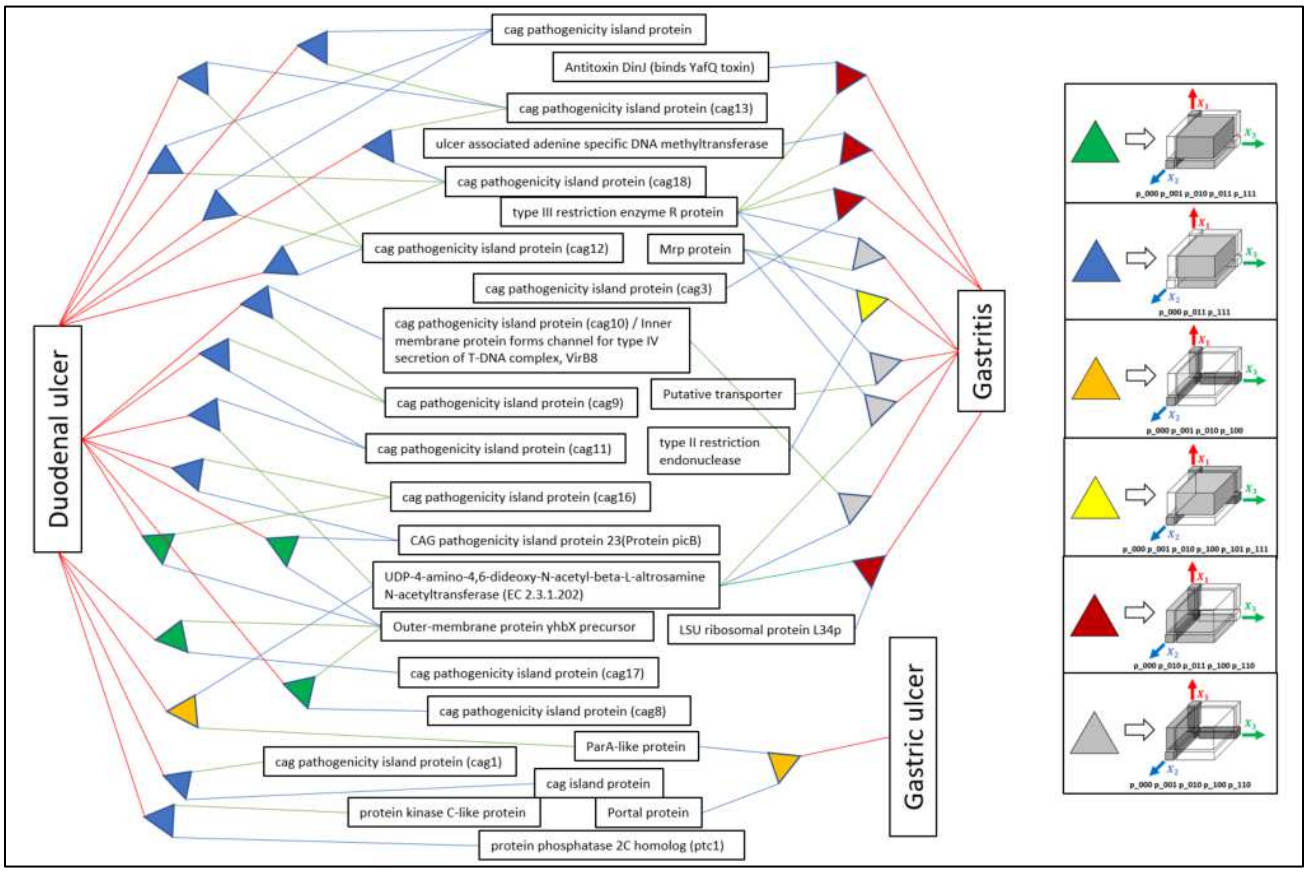

Figure 4. Subset of 3D patterns in *H. pylori* gene composition associated to duodenal ulcers, gastritis, and gastric ulcers.

Discussion

In contrast with dimensionality reduction (such as principal components analysis and multidimensional covariation) and AI solutions (neural network, deep learning, etc.), patterns identified by the proposed approach may provide better background for meaningful mechanistic interpretation of the underlying biological processes. Additionally a variety of mutual information-based methods are not suitable for estimating the strength of Boolean patterns because of the effects of the number of populated partitions and disbalance of the partitions' population on the pattern's score[1]. Our preliminary analysis suggests that multidimensional patterns may not only be present but could also dominate the landscape of the microbiome as well as other types of data (including multi-omics), which is not surprising because complex interactions between components of biological systems are unlikely to be reduced to simple pairwise interactions.

In contrast with AI, one of the distinguishable properties of the proposed approach is its ability to identify cases when no prediction is possible.

It also seems that different patterns represent different properties of biological systems. While some patterns, such as *co-presence* or *one-way relations* can be interpreted as an interaction between system components (proteins, microorganisms, etc.), other patterns, such as *co-exclusion* may represent the "design rules" of the systems allowing prediction of which microbiome compositions can and cannot exist.

The proposed approach can be extended in several ways. More complex patterns can be defined by using multiple thresholds for each feature., so instead of limiting consideration to absence/presence thresholds, consideration of patterns containing several levels of abundance, such as absence and presence in low, medium, and high abundance are possible. In this case, optimization of 3 thresholds: $\varepsilon_{absence}; \varepsilon_{low}; \varepsilon_{medium};$ would be necessary for each feature. Another interesting direction would be to allow already identified patterns to become metadata-like features and recurrently included in consideration as parts of input data.

It is important to keep in mind that the search for multidimensional patterns (as a search for any other type of pattern) is exploratory. Each finding must be validated. The most important outcome of the proposed analysis approach is that many hypotheses are filtered out allowing the validation effort to focus on a much smaller number of patterns.

Conclusion

Proposed definition of *multidimensional Boolean patterns* allows to accommodate heterogeneous multi-omics data, formulate criteria for separating trivial from non-trivial patterns, and identify conditions, required for a given pattern to predict the values of selected feature(s). Used in concert with the proposed definition of the pattern's strength (pattern's score) and minimal population threshold it permits estimation of the statistical significance of detected patterns by comparing scores distributions of artificial (randomized) datasets to original data.

Performed examination of all 2-, 3-, and 4-dimensional patterns in the Human Microbiome Project (15 body sites) data as well as in collection of $H.\ pylori$ genomes suggest that multidimensional patterns may dominate the landscape of microbial genomics/microbiomics systems.

Availability of data and materials

All data are incorporated into the article and its online supplementary material.

References

1. Golovko, G. et al. Identification of multidimensional Boolean patterns in microbial communities. *Microbiome* **8**, 131 (2020).
2. Saito, M.A. et al. Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph Crocosphaera watsonii. *Proc Natl Acad Sci U S A* **108**, 2184-2189 (2011).
3. Sahoo, D., Dill, D.L., Tibshirani, R. & Plevritis, S.K. Extracting binary signals from microarray time-course data. *Nucleic Acids Research* **35**, 3705-3712 (2007).
4. Faust, K. & Raes, J. CoNet app: inference of biological association networks using Cytoscape. *F1000Res* **5**, 1519 (2016).
5. Friedman, J. & Alm, E.J. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* **8**, e1002687 (2012).
6. Peterson, J. et al. The NIH Human Microbiome Project. *Genome Res* **19**, 2317-2323 (2009).
7. Davis, J.J. et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Research* **48**, D606-D612 (2019).

Acknowledgements

Authors are grateful to Brenda Boyko for her help and patience editing the manuscript.

Funding

Not applicable.

Author information

Author notes

1. George Golovko and Kamil Khanipov contributed equally to this work.

Affiliations

1. **Department of Pharmacology and Toxicology, University of Texas Medical Branch, Galveston, TX, USA**
   George Golovko, Kamil Khanipov & Yuriy Fofanov
2. **Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX, USA**
   George Golovko, Kamil Khanipov & Yuriy Fofanov
3. **Glass Bottom Analytics Inc, League City, TX, USA**
   Yuriy Fofanov
4. **Department of Pediatrics, University of Texas Medical Branch, Galveston, Texas, USA**

Victor Ries
5. **Department of Medicine, Penn State Health Milton Hershey Medical Center, Hershey, PA, USA**
Irina Pinchuk

Contributions

YF, GG and KK designed the approach and the computational framework. YF, GG, and KK analyzed the data. YF, GG, and KK carried out the implementation and performed the calculations. VR collected and analyzed *H.pylori* data. KK and IP collected and analyzed microbiome data. YF conceived the study and were in charge of the overall direction and planning. All authors read and approved the final manuscript.

Corresponding author

Correspondence to Yuriy Fofanov

Ethics declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

YF is co-founder of the Glass Bottom Analytics Inc.

Additional information

Publisher's Note

Supplementary information

Additional file 1 of Are Multidimensional Boolean Patterns Dominating Microbiome and Microbial Genome Data?

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuplementaryData.zip