Published in final edited form as:
J Comput Biol. 2023 April ; 30(4): 420–431. doi:10.1089/cmb.2022.0395.

A Novel Information-theory-based Genetic Distance that Approximates Phenotypic Differences

D. S. Campo[1,*], **A. Mosa**[2], **Y. Khudyakov**[1]

[1]Molecular Epidemiology & Bioinformatics Laboratory, Division of Viral Hepatitis, Centers for Disease Control and Prevention. Atlanta, GA, USA

[2]University Health Network, Toronto, Canada

Abstract

Application of genetic distances to measure phenotypic relatedness is a challenging task, reflecting the complex relationship between genotype and phenotype. Accurate assessment of proximity among sequences with different phenotypic traits depends on how strongly the chosen distance is associated with structural and functional properties. Here, we present a new distance measure (MIH, Mutual Information and entropy H) for categorical data such as nucleotide or amino acid sequences. MIH applies an information matrix (IM), which is calculated from the data and captures heterogeneity of individual positions as measured by Shannon entropy and coordinated substitutions among positions as measured by mutual information. In general, MIH assigns low weights to differences occurring at high entropy positions or at dependent positions. MIH distance was compared to other common distances on two experimental and two simulated datasets. MIH showed the best ability to distinguish cross-immunoreactive sequence pairs from non-cross-immunoreactive pairs of variants of the hepatitis C virus hypervariable region 1 (26,883 pairwise comparisons), and MHC (Major Histocompatibility Complex) binding peptides (n=181) from non-binding peptides (n=129). Analysis of 74 simulated RNA secondary structures also showed that the ratio between MIH distance of sequences from the same RNA structure and MIH of sequences from different structures is 3 orders of magnitude greater than for Hamming distances. These findings indicate that lower MIH between two sequences is associated with greater probability of the sequences to belong to the same phenotype. Examination of rule-based phenotypes generated *in silico* showed that: (i) MIH is strongly associated with phenotypic differences, (ii) IM of sequences under selection is very different from IM generated under random scenarios, and (iii) IM is robust to sampling. In conclusion, MIH strongly approximates structural/functional distances and should have important applications to a wide range of biological problems, including evolution, artificial selection of biological functions and structures, and measuring phenotypic similarity.

[*]Corresponding author.

AUTHOR CONTRIBUTIONS

DSC, AM and YK designed the study. DSC created the MIH distance and performed all calculations. DSC, AM and YK wrote the manuscript. All authors reviewed and approved the final manuscript.

AUTHOR'S DISCLOSURE

The authors declare they have no competing financial interests.

Keywords

Shannon entropy; Mutual information; Natural and artificial selection; Machine learning; Categorical variables; Protein; Genetic distance; Phenotype

1. INTRODUCTION

When analyzing categorical data, e.g., protein, DNA or presence/absence data, with a constant number of dimensions across observations, e.g., in a multiple sequence alignment, a common task in bioinformatics and data science is to compare two sequences or strings. The underlying intuition for this comparison is that genetically close sequences are more likely to encode similar structural and functional properties (or phenotype) than more distant sequences. Thus, various types of distances among sequences are frequently used to create similarity trees or networks, examination of which is expected to identify natural clusters that reflect similarity among encoded phenotypic traits and eventually allow for research on the underlying causes of such similarity.

Several distance measures have been developed for analyzing sequence similarity including such simple similarity-type Hamming and Jaccard indices applicable to any type of categorical data. Although these indices measure sequence similarity or difference, they do not fully capture phenotypic differences, owing to certain limitations:

(i) Simple distances assign equal weight to all substitutions, discounting differential effect of substitutions on phenotype. Application of amino acid physicochemical properties seemingly solves this problem for protein comparison. However, uncertainty in relevance of particular physicochemical properties to a specific phenotypic trait does not allow for the efficient measurement of functional differences among protein sequence variants. Capturing differential effects of substitutions on phenotypes is fundamental to study of evolution. One of the most frequently used measures used for protein comparison is BLOSSUM-62, which is based on substitution matrices empirically identified for proteins over a pre-defined evolutionary range. Considering that specific substitution may affect a particular domain or phenotype differently from the expected general trends observed over evolutionary scales, custom-made substitution models have been developed to replace general-purpose models (Goonesekere and Lee, 2008). Nevertheless, these approaches are not applicable to data science or ecology problems that involve other types of categorical data such as presence/absence rather than protein sequences.

(ii) Equal weight of substitutions at different positions (site or variable) is another limitation of simple distances. Although different positions have different conservation levels, a reflection of their differential contribution to phenotype, this information is not usually considered to improve the resolution of sequence comparisons using these distances.

(iii) Finally, the most insidious problem is that positions are rarely independent, a fact that is frequently underestimated, e.g., in phylogenetic reconstructions. A phenotype is not only affected by state of each site, but by interaction among states of different sites. Regarding proteins, their activities and properties are the result of interactions among constitutive amino acids (aa). Effect of substitutions that tend to destabilize a particular structure and/or function is compensated by substitutions at other sites resulting in stability restoration (Pollock and Taylor, 1997). For example, a substitution involving a reduction of volume in protein core might cause a destabilizing pocket, which only one or a few spatially adjacent residues would be capable of filling. Structurally or functionally linked sites tend to evolve in a correlated fashion due to the compensation process (Pollock and Taylor, 1997). Abundant experimental evidence indicates that proteins contain pairs of covariant sites. Such sites have been identified both by analysis of families of natural proteins with known structures (Altschuh et al., 1988, Bordo and Argos, 1990, Chothia and Lesk, 1982, Mateu and Fersht, 1999, Oosawa and Simon, 1986) and by site-directed mutagenesis (Armstrong et al., 2006, Baldwin et al., 1993, Lim and Sauer, 1989).

Functional constraints integrate all protein sites in a very specific structure, which restricts independence of mutations. In such integrated structure, individual mutations are frequently disadvantageous. The detrimental effect of these mutations can be compensated by mutations at other functionally linked sites (Govindarajan et al., 2003), making all these sites covariable. Genetically close sequences of proteins sharing a common evolutionary path should contain the vestiges of these effects in the form of covariant pairs of sites (Clarke, 1995). These interactions should manifest themselves in covariation between pairs of sites in a multiple sequence alignment. Analysis of covariation has been used in viral evolutionary studies (Campo et al., 2008), protein engineering (Voigt et al., 2001), detection of sequence-function correlations (Atchley et al., 2000, Fukami-Kobayashi et al., 2002), protein structure prediction (Afonnikov et al., 2001, Altschuh et al., 1988, Benner et al., 1997, Chen and Wang, 2005, Clarke, 1995, Göbel et al., 1994, Larson et al., 2000, Nagl et al., 1999, Neher, 1994, Nemoto et al., 2004, Shackelford and Karplus, 2007, Shindyalov et al., 1994, Taylor and Hatrick, 1994) and identification of important motifs in viral proteins (Altschuh et al., 1987, Amon et al., 2005, Kolli et al., 2006, Korber et al., 1993). Several studies confirmed that highly coordinated sites are often functionally and/or spatially coupled, with coevolving positions being frequently located in regions critical for protein function, such as active sites and surfaces involved in molecular interactions with other proteins (Atchley et al., 2000, Benner et al., 1997, Gloor et al., 2005, Poon and Chao, 2005, Yeang and Haussler, 2007). Recently, information on covariation observed among protein sites in sequence alignment was successfully used in the AlphaFold algorithm for identification of contacting aa residues (Jumper et al., 2021).

Thus, coordinated mutations at more than one protein site contribute to phenotypic variations and must be taken into consideration for measuring functional and structural similarity among proteins. Here, we propose application of coordinated substitutions as measured by Mutual Information (MI) in addition to heterogeneity of individual positions as measured by Shannon Entropy (H) to approximate structural/functional distances among naturally or artificially selected sequence variants. The novel MIH (Mutual Information and Entropy H) distance was evaluated using two experimental and two simulated datasets and showed better performance than other commonly used distances.

2. METHODS

2.1 Datasets

2.1.1. MHC-binding dataset: A dataset of 310 peptides, for which the MHC (Major Histocompatibility Complex) class I (Kb) binding activity was measured in a binary (yes/no) fashion (Hofmeyr, 2001, Milik, 1998), was used in this study. The dataset of 181 binders and 129 non-binders was originally obtained by random sampling from a large (>10^7) library of peptides. Multiple sequence alignment of the peptide sequences was generated using ClustalW with the BLOSUM protein weight matrix (as implemented in MEGAX) (Kumar et al., 2018). Two distance sets were defined by measuring distances within the binding phenotype (binders to binders) and between binding and non-binding phenotypes (binders to non-binders).

2.1.2. HVR1 cross-immunoreactivity dataset: HVR1 (Hyper-Variable region 1) is a 27aa hypervariable region located at the N-terminus of the E2 protein of hepatitis C virus. The dataset contains information on cross-immunoreactivity of 26,883 pairs among 262 HVR1 variants (Campo et al., 2012). Two distance sets were established: (1) distances between cross-immunoreactive HVR1 variants (antibody against one variant bind peptide containing another HVR1 variant) and (2) distances between non-cross-immunoreactive variants (no antibody binding). Considering that each 27 aa HVR1 peptide likely contains more than one epitope, these HVR1 sequence was divided into 20 possible overlapping 8-mers (step = 1). Distances were calculated for each 8-mer subregion. Only best results obtained for each distance in any subregion are reported here.

The information matrix (see section 2.2) for this region was calculated from 12,245 HVR1 nucleotide sequences obtained from the Virus Pathogen Database and Analysis Resource (ViPR)(Pickett et al., 2012). HVR1 nucleotide sequences were translated into protein sequences. Only one sequence per patient was used and sequences with insertions, deletions or stop codons were removed from analysis.

2.1.3. RNA secondary structure dataset: The RNA secondary structure is an essential phenotypic trait, as documented by its conservation in evolution and by convergent *in vitro* evolution toward a similar secondary structure when selecting for a specific function (Ekland and Bartel, 1996, van Nimwegen et al., 1999). RNA secondary structure prediction based on free-energy minimization is a standard tool in experimental biology and has been shown to be reliable (Huynen et al., 1997). We generated the full sequence space of A^D sequences, where A=2 (C and G only, to reduce the computational burden) and D = 13. Prediction of the minimum free energy structure for each sequence was done using seqFold (v0.7.14, https://pypi.org/project/seqfold/). Seqfold is an implementation of a commonly used dynamic programming algorithm (Zuker and Stiegler, 1981). The number of sequences in each structure was calculated and only structures with ≥4 sequences were used. Finally, two groups of distances among all sequences for each phenotypes were defined: (1) Inside: for distances between sequences forming the structure and (2) Outside: for distances between sequences forming the structure and those sequences that do not belong to the structure.

2.1.4. Rule-based random phenotypes:

We wanted to generate thousands of *in silico* datasets, each with sequences satisfying different randomly generated constraints:

1. Space generation: The full sequence space of A^D sequences was generated. For D=12 and A=2, a total of 4,096 sequences were generated.
2. A set of rules for this phenotype was defined. First, a random set of "conserved" positions (from 1 to D) was chosen. For each conserved position, a random set of states (from 1 to A-1 states) allowed in this position was defined. In addition, a random set of pairs of "linked" positions (from 0 to the number of combinations of 2 positions) was identified. For each linked pair, we define randomly its polarity as "positive" or "negative" (see next step).
3. Every sequence in the full space is evaluated by measuring how many of the rules it satisfies. For the "conserved" rules, we simply check if the sequence presents an allowed state in the specified position. For the "linked" pairs, we check the states at the two specified positions: if the pair is "positive", the states should be identical; if the pair is "negative", the states should be different.
4. Finally, each sequence has a measure of fitness that is the ratio between the number of satisfied rules and the total number of rules.

Only sequences with fitness equal 1 (all rules are matched) were considered to be members of the phenotype. This process was repeated to create 1,000 different random phenotypes, and for each one we defined two groups of distances among all sequences: (1) Inside: for distances between sequences that belong to this phenotype and (2) Outside: for distances between sequences that belong to this phenotype and those sequences that do not.

2.2. The MIH distance

The intuition behind MIH distance is the same as for the Mahalanobis distance for continuous data. The Mahalanobis distance was used to account for the fact that the variance of each variable is different and that there may be covariance between variables. This distance is commonly used to detect outliers (Etherington, 2021). The Mahalanobis distance is reduced to the Euclidean distance for uncorrelated variables with unit variance.

MIH distance can be applied to any type of categorical data, considering variability of each position as measured by entropy and existence of coordinated substitutions as measured by mutual information. The MIH distance between two sequences x and y is given by the following formula:

$$MIH(x, y) = xy^T . I^{-1} . xy$$

Where xy is the mismatch vector (with 1 where the symbols are different and 0 where they are the same) and xy^T is its transposed form. I is the symmetric information matrix (IM), with Shannon entropy (H) in the diagonals and mutual information (MI) between position pairs in all other entries. Formulation of Shannon entropy and mutual information is described elsewhere (Cover and Thomas, 2006). Logarithm was calculated using base equal to the number of states to ensure it ranges from 0 to 1.

For each dataset considered in this paper, IM was calculated using all available sequences. Each sequence had frequency equal to 1.

2.3 Comparison with other distances

2.3.1. Other distances: The MIH distance was compared with (i) Hamming distances: the number of mismatches divided by the sequence length; (ii) BLOSSUM62 distance: 1 minus the normalized similarity scores (from 0 to 1), (iii) Mahalanobis distances between vectors of dummy binary variables, and (iv) Minkowski distances (p=1, the Manhattan distance) between physiochemical vectors (see section 2.3.2).

2.3.2. Physiochemical vectors: There are many reported aa properties. Selection of suitable properties for a particular problem is a difficult task. A solution to this problem was proposed by Atchley et al. (Atchley et al., 2005) who used multivariate statistical analyses on 494 aa properties (Kawashima and Kanehisa, 2000) to produce a small set of highly interpretable numeric patterns of aa variability that can be used in a wide variety of analyses directed toward understanding the evolutionary, structural, and functional aspects of protein variability. This transformation summarizes the high level of redundancy in the original physicochemical attributes and produces much smaller, statistically independent, and well-conditioned variables for subsequent statistical analysis (Atchley and Zhao, 2007). The resultant five factors are linear functions of the original data, fewer in number than the original, and reflect clusters of covarying traits that describe the underlying structure of the variables (Atchley et al., 2005). Atchley et al. (Atchley and Zhao, 2007) showed how the transformation into one of the five multidimensional factors of physicochemical properties was useful in the analysis of Basic Helix-Loop-Helix proteins that bind DNA.

2.3.3. Statistical comparison between groups: To establish whether there were differences between the mean distances of the two groups in each examined dataset, we applied a Multi-Response Permutation procedure (MRPP). MRPP is a non-parametric permutation test for testing the hypothesis of no difference between two or more groups of entities (McCune and Grace, 2002). Permutation tests represent the ideal situations where one can derive the exact probabilities associated with a test statistic, rather than approximate values obtained from common probability distributions, such as t, F and X^2 (Cai, 2004). In most studies, the population distribution is unknown and assuming a normal distribution is inappropriate for many biological datasets, which often are skewed, discontinuous, and multi-modal. The distance-functions that form the basis of the MRPP are used to detect differences in distributions, sensitive to both dispersion and shifts in central tendency (Cade and Richards, 2001). In this study, we applied the MRPP test to the different groups of distances, with 10,000 permutations.

All calculations in this paper were performed with Python 3 and all programs are available upon request.

3. RESULTS

3.1 MIH shows superior performance on all datasets

The MIH distance weighs changes according to a symmetric information matrix (IM) calculated from all sequences under consideration, with Shannon entropy (H) being in diagonals and mutual information (MI) between position pairs in the corresponding entries. The intuition is that differences between two sequences at positions with high entropy are expected to be less essential for the phenotypic trait and accordingly receives low weight. Similarly, differences at a pair of highly associated positions also receive low weight since these positions are not independent and thus should not add their contribution ot the final distance. In the absence of selection constraints, the MIH distance is reduced to the Hamming distance when positions have maximum entropy, and every pair of positions has mutual information equal to zero (the null IM).

On the experimental datasets, MIH was compared to the Hamming, BLOSSUM62, Mahalanobis (dummy variables) and physiochemical distances. On the simulated datasets, MIH was compared only to Hamming distance. These distances were measured among all sequence variants to generate two sets, one for distances among sequence variants belonging to a phenotypic trait and another for distances among variants from two different phenotypic traits. There were significant differences between the two sets for all distances (MRPP test, $p < 0.0001$), indicating that sequence relatedness do captures differences in the phenotypic traits separating them. However, there were substantial variations in the magnitude of these differences.

Comparison among the distances was conducted using ratio (R) between the distance sets; R≈1 indicates lack of differentiation between presence or absence of the measured phenotypic trait. For the MHC dataset, R was greater for MIH than for any other distance, with the mean R=3.52 and the median R=7.08 (Figure 1A), indicating that low MIH between two sequences is a strong indication that both bind MHC molecules. For the HVR1 cross-immunoreactivity dataset, the mean R=1.89 and median R=1.40 were again greater than for other distances (Figure 1B), indicating that the lower MIH between two sequences is associated with the higher probability of them being cross-immunoreactive.

The simulated RNA dataset contained 79 different secondary structures (or phenotypes), 74 of which were represented with ≥4 sequences and analyzed further. The MIH showed a ratio 3 orders of magnitude greater than for Hamming distances (Figure 1C), indicating that lower MIH between two sequences is associated with greater probability of the sequences to belong to the same RNA structure.

For the rule-based simulated phenotypes (n=1,000), each phenotype was defined by sequences that do not break any of the randomly defined rules. Distance sets were generated for each of 1,000 phenotypes, with one set for within phenotype and another between the phenotype and all other phenotypes. As shown in Figure 1D, mean MIH ratio is consistently higher than the mean ratio of Hamming distances, with the value being proportional to the size of selected datasets, reaching R=30.91. However, the mean ratio of Hamming distance followed the opposite trend, with its value being inversely proportional to the size of the selected dataset, quickly approaching R=1.18. Thus, the lower the MIH distance between two sequences, the higher the probability that they belong to the same rule-based phenotype.

3.3 Properties of the rule-based phenotypes

IM of each phenotype was compared to the null IM, which is the identity matrix (1s in diagonals and 0s everywhere else). The null IM can be obtained in absence of selection when all possible sequences in the space are used to calculate IM. The use of the null IM reduces MIH to Hamming distances because all positions are maximally heterogenous and independent from each other.

To study whether the differences between the observed IM in each of the 1,000 phenotypes of constant size (n=32) and the null IM can be randomly obtained, two random scenarios were considered: (i) a random sample of the entire space and (ii) a connected sample generated from one seed by randomly choosing a neighbour sequence, which is in turn used to select a random neighbour, and iterating the procedure until the desired size is reached. Figure 2A shows that the rule-based constraints generate phenotypes with IM being more different from the null IM (mean RMSE, Root Mean Squared Error = 0.3140) than IM from the random (mean RMSE = 0.0389) or connected scenarios (mean RMSE = 0.1612).

These results show that IM, as a foundation for MIH, is associated with phenotypic differences. However, simulations considered here allow for using all possible sequences that belong to a certain phenotype, which is hardly attainable experimentally. To simulate the effect of sampling on the IM calculation, 1,000 replica subsamples (for each level ranging from 5% to 95%) of the members of 1,000 phenotypes of the same size (n=256) were generated. First, we measured the RMSE between the subsampled IM and the null IM (Figure 2B), showing that even at very low levels of sampling (5%) RMSE=0.2408 is only 6.2% greater than RMSE=0.2266 obtained without sampling. Second, we measured RMSE of each subsampled IM vs the full IM obtained without sampling. Although it takes a sample of ~20% to get close to the full IM, even at very low sample levels, the sampled IM is 2.5-times closer to the full IM (RMSE=0.0970) than to the null IM at that sampling level. These results suggest that the IM calculation is robust to sampling variations.

3.4. Fitness and MIH distances in rule-based phenotypes

The rule-based constraints provide opportunity to measure fitness values as a fraction of satisfied rules for every sequence in the space, and thus allow to explore strength of association between MIH and phenotype differences.

Between-phenotype analysis: Comparison of 499,500 phenotype pairs was conducted using RMSE between the fitness values of all sequences. The fitness RMSE were compared to the mean distances. A small association was found with Hamming distances (Pearson correlation, r = 0.2611, p value = 0) (Figure 3A). It is important to consider that MIH between two sequences belonging to different phenotypes can be calculated in three ways: (i) MIH-1 using IM of phenotype 1; (ii) MIH-2 using IM of phenotype 2; and (ii) MIH-3 using a joint IM calculated from the union of both sequence sets. A strong association was found with the average of MIH-1 and MIH-2 (r = 0.8861, p = 0) (Figure 3B), maximum of

MIH-1 and MIH-2 (r = 0.8415, p = 0), and minimum of MIH-1 and MIH-2 (r = 0.6374, p = 0) but no association was found using the joint MIH (r = −0.0009, p = 0.5535).

Within-phenotype analysis: Within a single rule-based phenotype (n=1,000), the standard deviation of all fitness values is inversely proportional to the RMSE between MIH and Hamming distances (r = 0.9305, p = 0). Thus, higher discrepancy between MIH and Hamming distances is strongly associated with greater heterogeneity of fitness values (Figure 3C). Furthermore, correlation between distance and difference in fitness for each pair of sequences is always positive and proportional to the phenotype size for all distances (Figure 3D). However, the correlation value is much higher for MIH than for Hamming distance at all phenotype sizes (average ratio = 2.00). It is important to note that this correlation between MIH and fitness difference is even greater locally among one-step neighbors than globally among all possible pairs as described above (average ratio = 5.86), especially at a small phenotype size. This local correlation cannot be calculated for Hamming distances because all distances are identical (d=1).

4. DISCUSSION

The MIH distance is a novel measure of differences among genetic variants that have been naturally or artificially selected. The process of selection creates profound constraints on the sequence variability, while keeping constant the selected structure or function. In both our experimental datasets, MIH provided a better separation between sets than other distances. However, it was surprising that simple Hamming distance performed equally well to more complex measures specific to proteins (BLOSSUM62 and physiochemical properties of amino acids). As the Mahalanobis distance shares many of the properties of MIH, we expected a better performance, but at least in these two datasets the change from categorical (20aa) to dummy variables seems to make a substantial difference.

Although it is important to have validation with experimental and/or biologically inspired datasets (such as the RNA secondary structures), these are only three examples out of the immense space of possible phenotypes and experimental datasets suitable for such validation are not readily available. Therefore, here we devised a new *in silico* framework to generating datasets of sequences under strong constraints to demonstrate application of MIH to evaluation of differences among artificially selected phenotypes. This framework extends the intuition behind the NK model for generating random fitness landscapes (Kauffman and Weinberger, 1989). The NK model could be adapted to create datasets of sequences by using a threshold fitness and tunable epistasis level (parameter K, the number of positions associated with each position). However, the pairs of associated positions considered by this model are fixed and fully linked, whereas the Rule-based model offers a greater flexibility in the strength of association between pairs of positions and in the structure of epistatic networks. In addition, the model provides a complete control over the allowed heterogeneity levels at each position, which is lacking in the NK model. Thus, the Rule-based model affords a strong control over heterogeneity of individual positions, strength of epistatic interactions and structure of epistatic networks, while providing a ruleset allowing easy interpretability of results and simple computational implementation.

There are three practical issues associated with calculation of MIH. (i) In rare cases, the determinant of the IM is equal to zero, preventing calculation of the inverse matrix required for measuring MIH. In these cases, addition of one over the number of sequences to a random number of entries in IM (making sure of keeping the matrix symmetric) resolves this numerical problem but simultaneously adds slight variation to distances calculated at every run, which could be potentially problematic when the phenotypic differences are small. (ii) MIH distances can have a very broad range, reaching very high values, as shown here for the RNA structure dataset. A rank transformation mitigates this issue. However, it seems important to examine other ways to normalize MIH distances. (iii) MIH distance performs best when comparing sequences within a single simple phenotype (for instance, to detect true phenotypic clusters or merge apparent but equivalent clusters) or when comparing related sequences encoding a certain phenotypic trait against sequences that do not encode this trait. However, when comparing sets of sequences that encode different phenotypes, the task is to choose one of possible distances depending on the used IM as we have experienced here for the between-phenotype comparisons using the rule-based simulations. Although our results showed that MIH is more correlated with phenotypic differences than Hamming distances, the choice of IM for one of the compared phenotypes or another, average IM for both phenotypes, and joint-phenotype IM may potentially cause interpretation problems in some settings where there is not an evidence-based way of choosing one. We found in this study that using the average was the best option, followed by the maximum.

Application of MIH offers several advantages: (i) it is better associated with phenotypic differences than other evaluated distances, (ii) IM of sequences under selection is very different from IMs generated under random scenarios, thus potentially allowing novel applications for evaluation of selection pressures of genomic regions, and (iii) IM is robust to sampling. If a given phenotype is the target of natural or artificial selection, IM for the viable sequences under this constraint is fundamentally different from IM of a sample of random sequences as we have observed here using rule-based datasets. This result suggests that deviation of the observed IM from the null IM (built using all sequences without constraints) can be used as a measure for the strength of selection. With further validation, this property can allow comparison between sequence sets (e.g., viral populations from different patients) to estimate differences in the selective constraints acting on both populations, even at the equal or comparable levels of heterogeneity for both sets. This is particularly important for analyzing noncoding regions where dN/dS ratio is not applicable.

One of the most important goals of data visualization is to assess proximity of data points. As was shown here, MIH is greatly associated with the underlying structural or functional differences. This important property of MIH is invaluable for visualization of genetic proximity and phenotype similarity. Simulation experiments presented in this study show that MIH is highly correlated with fitness differences. As fitness values and phenotypic differences are almost always unknown, MIH provides a simple approximation that can be used for cluster detection or other exploratory analyses. It is noteworthy that the MIH distance is calculated with the IM coming only from the sequences with maximum fitness that did not break any rule, but it still provides information about the fitness of all other sequences. With further evaluation and development, this important property offers applications of MIH to unsupervised analysis of the data, e.g., for building networks or similarity trees to identify phenotypic clustering of sequences. Recently, this MIH property was used to detect previously undetectable clusters of Hepatitis C Virus HVR1 variants to guide development of broadly cross-neutralizing vaccine candidates (Mosa et al., 2021).

The results shown here suggest that MIH is a useful addition to the field of supervised machine learning, specifically by improving the k-nearest-neighbor algorithm. This is a simple but powerful approach for classification that stores all the available cases and classifies the new data based on a similarity measure. However, this similarity is obtained using all variables, even though some are obviously more important than others. Although the problem is alleviated by performing feature selection prior to this step, the problem persists as the remaining variables are likely to differentially contribute to traits and are not independent from each other. Application of MIH, which is associated with differences in structural, functional, or phenotypic traits, could improve performance of the nearest-neighbor approach, which warrants further investigation. Given that MIH distance is correlated with fitness differentials among close neighbors, the selection of neighbors itself could be more meaningful as well. In addition, it is well known that as dimensionality increases, the distance from a given point to the nearest data point approaches the distance to the farthest data point (Beyer et al., 1999). Thus, distances lose discriminatory power very rapidly in a phenomenon known as the dimensionality curse. Consideration of coordinated substitutions implemented in IM reduces dimensionality of data *vs.* application of null IM with as many dimensions as many positions in sequences. As MIH effectively reduces the dimensionality, it may help mitigate this common problem.

In conclusion, a new distance based on information theory presented here for categorical data such as nucleotide and amino acid sequences approximates underlying structural/ functional distances and could have important applications to a wide range of biological problems, including evolution, artificial selection of biological functions and structures, and measuring phenotypic similarity.

ACKNOWLEDGMENTS

An early version of this paper was published as part of the 2021 11th International Conference on Computational Advances in Bio and medical Sciences (ICCABS). The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention (CDC).

REFERENCES

Afonnikov D, Oshchepkov D, Kolchanov N. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. Bioinformatics 2001;17(11):1035–1046. [PubMed: 11724732]

Altschuh D, Lesk A, Bloomer A, et al. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol 1987;193(4):693–707. [PubMed: 3612789]

Altschuh D, Vernet T, Berti P, et al. Coordinated amino acid changes in homologous protein families. Protein Eng 1988;2(3):193–199. [PubMed: 3237684]

Amon JJ, Devasia R, Xia G, et al. Molecular epidemiology of foodborne hepatitis a outbreaks in the United States, 2003. J Infect Dis 2005;192(8):1323–1330; doi: 10.1086/462425. [PubMed: 16170748]

Armstrong GL, Wasley A, Simard EP, et al. The prevalence of hepatitis C virus infection in the United States, 1999 through 2002. Annals of internal medicine 2006;144(10):705–714. [PubMed: 16702586]

Atchley W, Wollenberg K, Fitch W, et al. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 2000;17(1):164–178. [PubMed: 10666716]

Atchley W, Zhao J. Molecular architecture of the DNA-binding region and its relationship to classification of basic helix-loop-helix proteins. Mol Biol Evol 2007;24(1):192–202. [PubMed: 17041153]

Atchley W, Zhao J, Fernandes A, et al. Solving the protein sequence metric problem. Proc Natl Acad Sci USA 2005;102(18):6395–6400. [PubMed: 15851683]

Baldwin E, Hajiseyedjavadi O, Baase W, et al. The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. Science 1993;262(5140):1715–1718. [PubMed: 8259514]

Benner S, Cannarozzi G, Gerloff D, et al. Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. Chem Rev 1997;97:2725–2844. [PubMed: 11851479]

Beyer K, Goldstein J, Ramakrishnan R, et al. When Is “Nearest Neighbor” Meaningful? In: Database Theory — ICDT’99. (Beeri C, Buneman P ed.) Springer, Berlin, Heidelberg. : 1999; pp. 217–235.

Bordo D, Argos P. Evolution of protein cores. Constraints in point mutations as observed in globin tertiary structures. . J Mol Biol 1990;211(4):975–988. [PubMed: 2313703]

Cade B, Richards J. User manual for BLOSSOM statistical software. Midcontinent Ecological Science Center US Geological Survey 2001.

Cai L Multi-response permutation procedure as an alternative to the analysis of variance: An SPSS implementation. Depart Psychol University N Carolina 2004.

Campo D, Dimitrova Z, Mitchell R, et al. Coordinated evolution of the hepatitis C virus. PNAS 2008;105(28):9685–9690. [PubMed: 18621679]

Campo DS, Dimitrova Z, Yokosawa J, et al. Hepatitis C virus antigenic convergence. Sci Rep 2012; (2):267–277. [PubMed: 22355779]

Chen S, Wang Y-M. Multigene tracking os quasispecies in viral persistence and clearance of hepatitis C virua. World J Gastroenterol 2005;11(19):2874–2884. [PubMed: 15902722]

Chothia C, Lesk A. Evolution of proteins formed by beta-sheets. I. Plastocyanin and azurin. J Mol Biol 1982;160(2):309–323. [PubMed: 6816943]

Clarke N Covariation of residues in the homeodomain sequence family. Protein Sci 1995;4(11):2269–2278. [PubMed: 8563623]

Cover TM, Thomas JA. Elements of information theory. Wiley-Interscience: Hoboken, N.J.; 2006.

Ekland EH, Bartel DP. RNA-catalysed RNA polymerization using nucleoside triphosphates. Nature 1996;382(6589):373–376; doi: 10.1038/382373a0. [PubMed: 8684470]

Etherington TR. Mahalanobis distances for ecological niche modelling and outlier detection: implications of sample size, error, and bias for selecting and parameterising a multivariate location and scatter method. Peer J 2021;9:e11436; doi: 10.7717/peerj.11436. [PubMed: 34026369]

Fukami-Kobayashi K, Schreiber D, Benner S. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J Mol Biol 2002;319:729–743. [PubMed: 12054866]

Gloor G, Martin L, Wahl L, et al. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 2005;44(19):156–165.

Göbel U, Sander C, Schneider R, et al. Correlated mutations and residue contacts in proteins. Proteins 1994;18(4):309–317. [PubMed: 8208723]

Goonesekere NC, Lee B. Context-specific amino acid substitution matrices and their use in the detection of protein homologs. Proteins 2008;71(2):910–919; doi: 10.1002/prot.21775. [PubMed: 18004781]

Govindarajan S, Ness J, Kim S, et al. Systematic variation of Amino acid substitutions for stringent assesment of pairwise covariation. J Mol Biol 2003;328:1061–1069. [PubMed: 12729741]

Hofmeyr S An Interpretative Introduction to the IS. Design Principles for ISs and Other Distributed Autonomous Systems 2001:3–28.

Huynen M, Gutell R, Konings D. Assessing the reliability of RNA folding using statistical mechanics. J Mol Biol 1997;267(5):1104–1112; doi: 10.1006/jmbi.1997.0889. [PubMed: 9150399]

Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–589; doi: 10.1038/s41586-021-03819-2. [PubMed: 34265844]

Kauffman S, Weinberger E. The NK model of rugged fitness landscapes and its application to maturation of the immune response. J Theor Biol 1989;141(2):211–245. [PubMed: 2632988]

Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res 2000;28:374. [PubMed: 10592278]

Kolli M, Lastere S, Schiffer C. Co-evolution of nelfinavir-resistant HIV-1 protease and the p1-p6 substrate. Virology 2006;347(2):405–409. [PubMed: 16430939]

Korber B, Farber R, Wolpert D, et al. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci USA 1993;90(15):7176–7180. [PubMed: 8346232]

Kumar S, Stecher G, Li M, et al. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 2018;35(6):1547–1549; doi: 10.1093/molbev/msy096. [PubMed: 29722887]

Larson S, Di Nardo A, Davidson A. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol 2000;303(3):433–446. [PubMed: 11031119]

Lim W, Sauer R. Alternative packing arrangements in the hydrophobic core of lambda repressor. Nature 1989;339(6219):31–36. [PubMed: 2524006]

Mateu M, Fersht A. Mutually compensatory mutations during evolution of the tetramerization domain of tumor supressor p53 lead to impaired hetero-oligomerization. Proc Natl Acad Sci USA 1999;96:3595–3599. [PubMed: 10097082]

McCune B, Grace J. Analysis of ecological communities. MjM Software Design: Gleneden Beach; 2002.

Milik M, S D. Brunmark A. Yuan L. Vitiello A. Jackson M. Peterson P. Skolnick J. Glass C. Application of an artificial neural network to predict specific class I MHC binding peptide sequences. Nat Biotechnol 1998;16(8):753–756. [PubMed: 9702774]

Mosa A, Campo D, Urbanowicz R, et al. Pentavalent HCV Vaccine Candidate Elicits Broadly Neutralizing Antibodies to Neutralization-Resistant Variants. In: The liver meeting. AASLD. 2021.

Nagl S, Freeman J, Smith T. Evolutionary constraint networks in ligand-binding domains: an information-theoretic approach. Pac Symp Biocomput 1999:90–101. [PubMed: 10380188]

Neher E How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA 1994;91(1):98–102. [PubMed: 8278414]

Nemoto W, Imai T, Takahashi T, et al. Detection of pairwise residue proximity by covariation analysis for 3D-structure prediction of G-protein-coupled receptors. Protein J 2004;23(6):427–435. [PubMed: 15517989]

Oosawa K, Simon M. Analysis of mutations in the transmembrane region of the aspartate chemoreceptor in Escherichia coli. Proc Natl Acad Sci USA 1986;83(18):6930–6934. [PubMed: 3018752]

Pickett BE, Greer DS, Zhang Y, et al. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses 2012;4(11):3209–3226; doi: 10.3390/v4113209. [PubMed: 23202522]

Pollock D, Taylor W. Effectiveness of correlation analysis in identifying protein residues. Protein Eng 1997;10(6):647–657. [PubMed: 9278277]

Poon A, Chao L. The rate of compensatory mutation in the DNA bacteriophage phiX174. Genetics 2005;170(3):989–999. [PubMed: 15911582]

Shackelford G, Karplus K. Contact prediction using mutual information and neural nets. Proteins 2007;69(Suppl 8):159–164. . [PubMed: 17932918]

Shindyalov I, kolchanov N, Sander C. Can three dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 1994;7:349–358. [PubMed: 8177884]

Taylor W, Hatrick K. Compensating changes in protein multiple sequence alignments. Protein Eng 1994;7(3):341–348. [PubMed: 8177883]

van Nimwegen E, Crutchfield JP, Huynen M. Neutral evolution of mutational robustness. Proc Natl Acad Sci U S A 1999;96(17):9716–9720. [PubMed: 10449760]

Voigt C, Mayo S, Arnold F, et al. Computational method to reduce the search space for directed protein evolution. Proc Natl Acad Sci USA 2001;98:3778–3783. [PubMed: 11274394]

Yeang C, Haussler D. Detecting coevolution in and among protein domains. PLoS Comput Biol 2007 3(11):e211. [PubMed: 17983264]

Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 1981;9(1):133–148; doi: 10.1093/nar/9.1.133. [PubMed: 6163133]

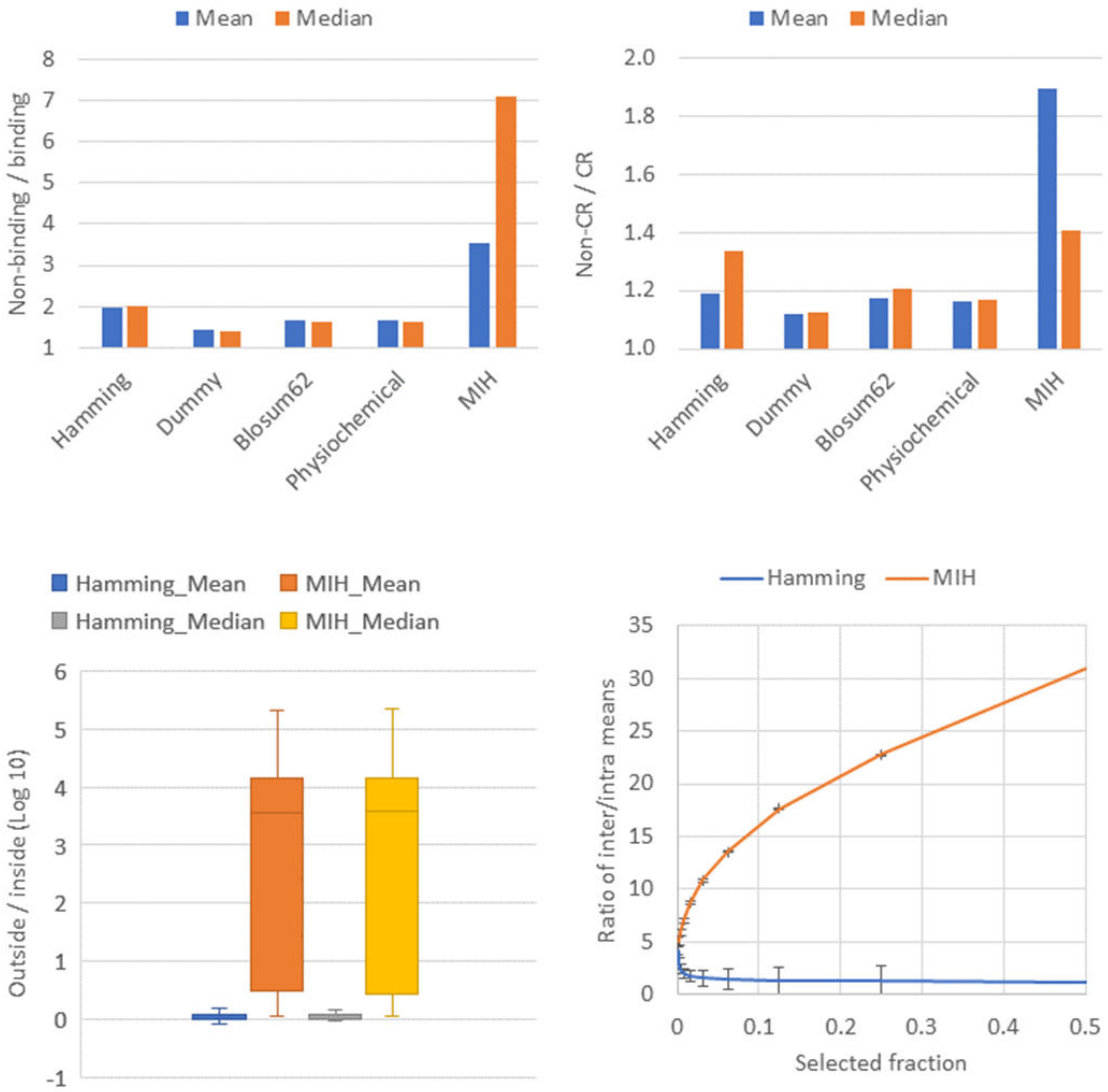

Figure 1.

Distance comparison among four datasets. A) MHC-binding dataset, comparing 5 different distance types. B) Cross-reactivity dataset, comparing 5 different distance types. C) RNA secondary structure dataset, comparing Hamming and MIH and showing a boxplot of all the ratios. D) Rule-based dataset, comparing Hamming and MIH depending on the size of the selected phenotypes (x-axis). The error bars show the standard deviation among 1000 phenotypes.

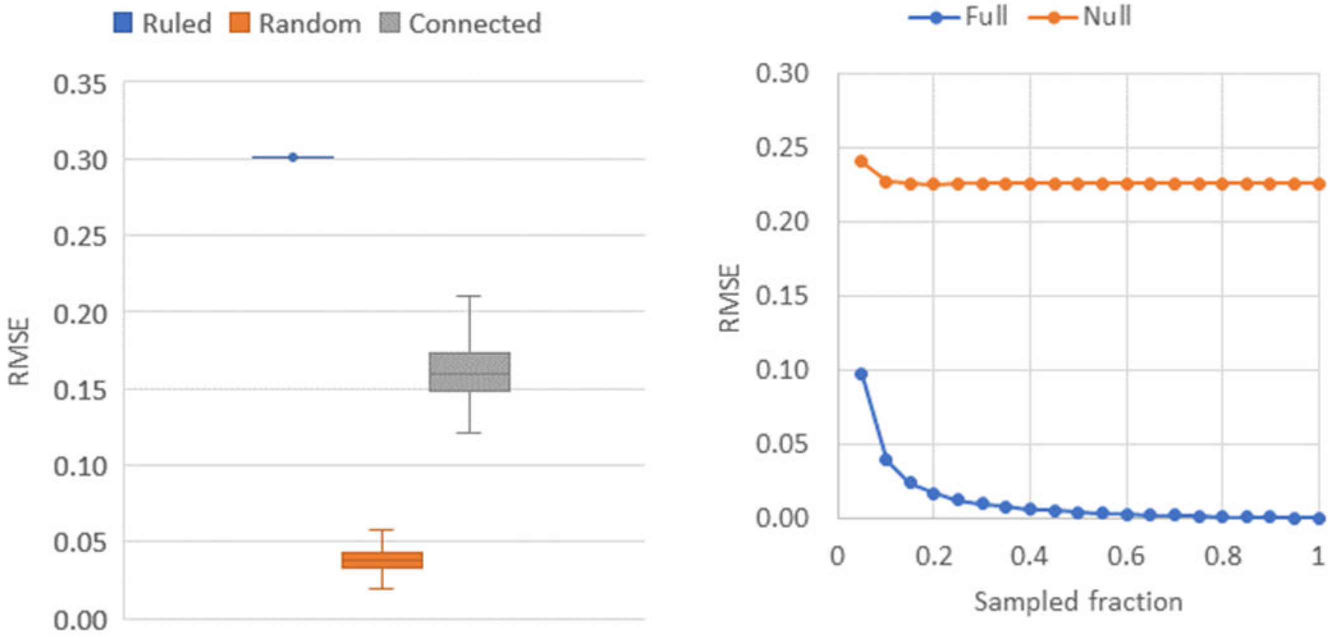

Figure 2.

Information matrix of rule-based phenotypes. A) Boxplot of the RMSE between the IM of observed samples and the null IM with three scenarios: rule-based, random and connected. C) RMSE (Root Mean Squared Error) between the IM of subsamples and the null IM or the full IM. The x-axis shows the different sampling levels.

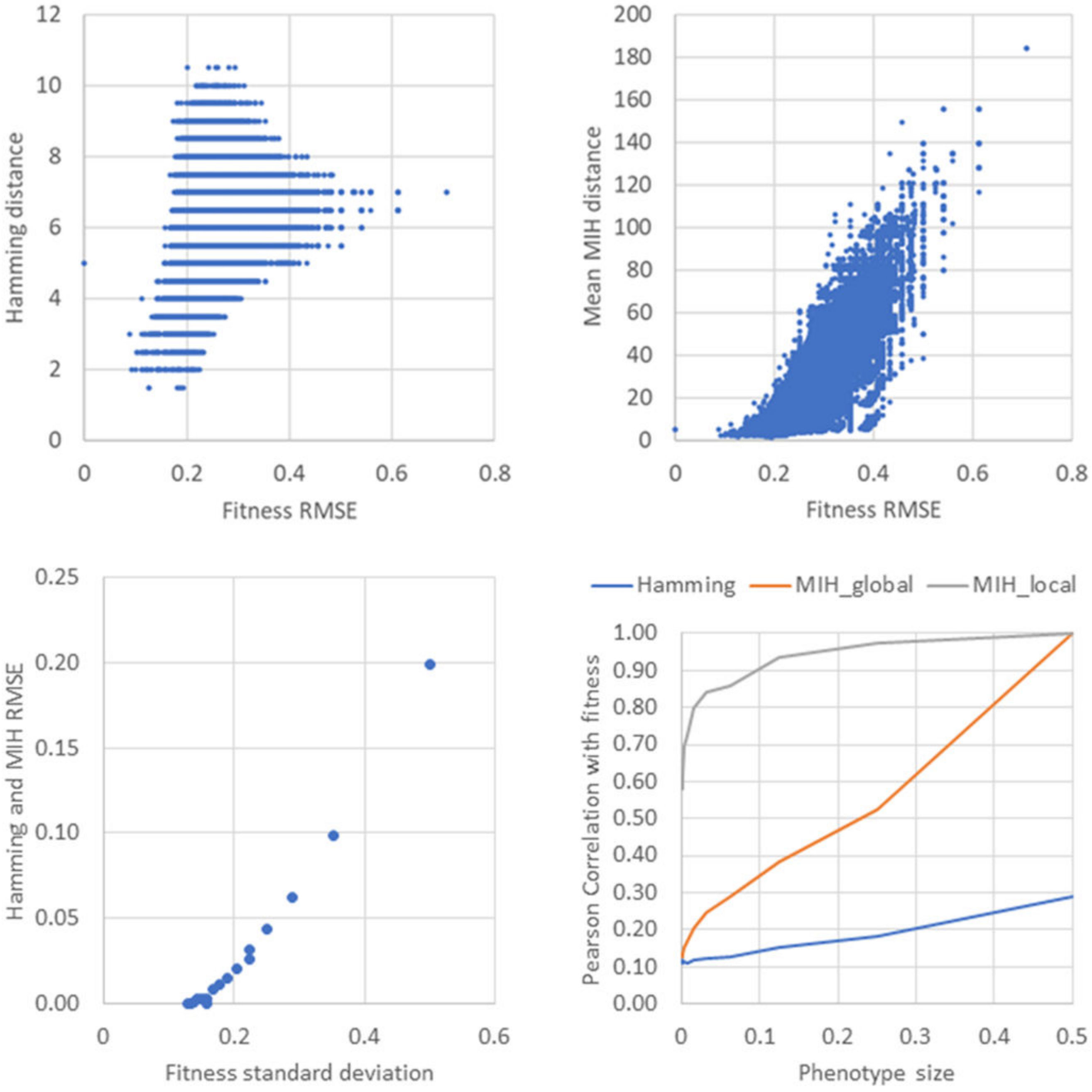

Figure 3.

Fitness and rule-based phenotypes. A) Between-phenotype comparison: Scatterplot between fitness RMSE and Hamming distance. B) Between-phenotype comparison: Scatterplot between fitness RMSE and Hamming distance. C) Within-phenotype scatterplot of RMSE (between hamming and MIH distances) and the fitness standard deviation of the entire space. D) Within-phenotype scatterplot of the Pearson correlation between fitness differentials and three distance types: Hamming, MIH global and local MIH (including only one-step neighbors). The x-axis shows the different sizes of the selected rule-based phenotypes.