An experimental study of employing visual appearance as a phenotype

Lior Wolf Yoni Donner
Tel-Aviv University, Israel
www.cs.tau.ac.il/~wolf

Abstract

Visual and non-visual data are often related through complex, indirect links, thus making the prediction of one from the other difficult. Examples include the partially-understood connections between firing of V1 neurons and visual stimuli, the coupling between recorded speech and video of the corresponding lip movements, and the attempts to infer criminal intentions from surveillance videos.

In this study, we explore the exploitation of the visual/non-visual relation between genetic sequences and visual appearance. This exploitation is currently considered infeasible due to the many hidden variables and unknown factors involved, the considerable variability and noise that exist in images and the high-dimensionality of the data.

Despite the difficulties, we show convincing evidence that the application of correlations between genotype and visual phenotype for identification is feasible with current technologies. To this end, we employ sensitive forced-matching tests, that can accurately detect correlations between data sets. These tests are used to compare the performance of several existing algorithms, as well as novel ones that we have designed for the task.

1. Introduction

It is undebatable that our inheritance determines much of our appearance. However, the association of genetic data and visual appearance, on a level that exceeds single traits, would appear to most biologists a remote possibility. Their reasoning may include the following factors:

Complex unknown mechanisms. The genetic influences on visual appearance are complex and involve a multitude of genes and pathways, and are therefore difficult to analyze. The exact causal explanation of the links between genes and appearance is beyond present-day knowledge.

Unknown genes. Not only is a causal model missing, even the identities of the genes that are most relevant to visual appearance are vastly unknown.

Hidden variables. Even given a full causal model, it would still be impossible to fully predict appearance from genetic data alone, due to the influence of environmental factors.

High dimensional data. Images, and their common representations, are much more varied, noisy and high-dimensional than traits that are currently used in genotype/phenotype[1] association studies.

Limited availability of data sets. A study comparing genes to visual appearance would be most beneficial within species and in between nearby species. Currently only few such data sets exist for higher organisms.

In spite of these difficulties, we show that visual appearance, acquired directly from images, and analyzed automatically, is usable as a phenotype. Instead of tackling the task of predicting the appearance given the genetic data, we focus on a more limited identification task: given a previously unseen gene sequence, identify the matching image, out of a set of several previously unseen images. We refer to such tests as forced matching tests.

To achieve our results, we exploit the following:

Sufficiency of correlations. A causal model is not required for identification, even in face of many missing variables.

Avoiding synthesis. The richness of the visual data is a disadvantage for the task of predicting the appearance given the genotype, while it is an advantage for identification.

The Barcode Of Life data sets. Our study is based on having the sequence of the same gene across similar species. This is the kind of data which is collected in the Barcode of Life Database (BOLD) [14]. It should be noted, however, that this database currently contains the sequence of only one gene, which is mitochondrial (see below) and hardly related to appearance.

Informative data representations. The modern image analysis tools provide us with rich feature vectors that contain much more information than the single traits commonly used in association studies.

2. Biological background

Previous work on linking genotype to visually-identifiable phenotypes has focused on univariate or low-

[1] The term *genotype* refers to an organism's exact genetic makeup, while its observable characteristics are referred to as *phenotype*.

dimensional traits such as eye color [19], principal variations in skeletal structure [3] and height [20], as well as on the discovery of specific genes that contribute to these traits. In contrast, our work goes beyond single traits to the direct genotype-phenotype analysis of photographs and illustrations of animal species.

2.1. Mitochondrial DNA

Mitochondrial DNA (mtDNA) is normally inherited unchanged from the mother, with very limited recombination compared to nuclear DNA, yet its mutation rate is higher than that of nuclear DNA [2], resulting in low variance within species and high variance between species, making it a promising candidate for species identification[2].

The mitochondrial gene cytochrome c oxidase I (COI) has been repeatedly employed for DNA barcoding and its discriminative effectiveness has been demonstrated in several species [6]. We use COI sequence data for all experiments, mainly due to the high availability of COI sequences for many species of the same genus[3], publicly available from BOLD [14].

Mitochondrial genes have mostly migrated to the nucleus. Those left in the mtDNA are believed to be involved in the process of cell metabolism. How can mitochondrial genes be used to identify images? The source of the correlations may lie in their common genetic history. Consider the example illustrated in Figure 1, involving an ancestral population A in which some versions of the mitochondrial and the appearance-related genes exist, and an ancestral population B, in which different versions exist. The species which originate from population A may have these mtDNA and appearance-genes versions (in the same species), yet the species that originate from population B are unlikely to have these combinations. Hence, correlative links between mtDNA and appearance-encoding-genes are formed.

3. Algorithms

We formulate the problem we solve as follows. Given a training set $\{m_i, p_i\}_{i=1}^n$ of n matching genetic markers and images, a marker of a new organism m_{new} and a set of images of unseen species $\{p_{new_1}, \ldots, p_{new_k}\}$, choose the image which best matches the new marker. In most of our experiments, $k = 2$.

The most direct approach to solving forced matching tests is to learn the joint probability $Pr(m, p)$ and choose the image p_{new_j} that maximizes $Pr(m_{new}, p_{new_j})$. Learning this joint probability distribution, however, is difficult and may not be necessary. Below we present several algorithms for forced matching tests that do not estimate the above density.

Let x_i and y_i be the centered genotype and phenotype vectors obtained by subtracting the empirical means:

$$x_i = m_i - \frac{1}{n}\sum_{j=1}^{n} m_j, \quad y_i = p_i - \tfrac{1}{n}\sum_{j=1}^{n} p_j$$

x_{new} and y_{new_k} are similarly defined, using the means estimated on the training set only.

3.1. Canonical Correlation Analysis

In Canonical Correlation Analysis (CCA) [7], two transformations are found that cast genes and images to a common target vector-space such that the matching genes and images are transformed to similar vectors in the sense of maximal correlation coefficient. Additional constraints are that the components of the resulting vectors would be pairwise uncorrelated and of unit variance.

The CCA formulation is, therefore:

$$\max_{W_X, W_Y} \sum_{i=1}^{n} x_i^T W_X W_Y^T y_i \ , \ \text{subject to}$$
$$\sum_{i=1}^{n} W_X^T x_i x_i^T W_X = \sum_{i=1}^{n} W_Y^T y_i y_i^T W_Y = I$$

The dimension of the target vector space is the minimum of the two dimensions d_M and d_P and shall be denoted by l. Thus W_X is a matrix of dimensions $d_M \times l$ and W_Y is of dimensions $d_P \times l$.

The matrices W_X and W_Y are then used for forced matching by choosing the image that minimizes $D(W_X^T x_{new}, W_Y^T y_{new_j})$, where $D : \mathbf{R}^l \times \mathbf{R}^l \to \mathbf{R}$ is a distance metric.

Though a standard distance metric such as Euclidean distance can be used, it may be beneficial to weigh the elements of the vectors by the corresponding correlation coefficients, as they vary for different locations in the target vectors. Similar to the transformation to Mahalanobis space, we define a transformation to correlation space, in which each element of the transformed vector is multiplied by the corresponding correlation coefficient. Denote the i'th correlation coefficient by ρ_i, then for $u \in \mathbf{R}^l$ the corresponding vector in correlation space $\hat{u}$ is defined by $\hat{u}_i = \rho_i u_i$.

Inspired by the distance metrics employed in the CSU Face Identification Evaluation System [1], we define the following distance metrics:

[2]Genomic species identification is the process in which given a DNA sample, the species of the donating animal is recovered. DNA barcoding is the process of creating a signature based on a DNA sample, for example, for the purpose of species identification.

[3]Genus (pl. genera) means a taxonomic group. We prefer to work within genus and not between genera since it simplifies the unification of image data gathering, and since we cannot imagine applications that involve describing previously unseen genera.

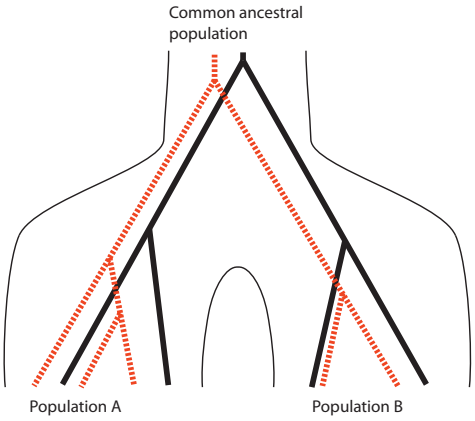

Figure 1. It may seem surprising that mitochondrial genes can be used to identify images, since they do not affect appearance directly. These genes may, however, be correlated with appearance due to the common evolutionary lineage shared by all genes. The figure above demonstrates correlations that are created by an event of population split. The two colors correspond to genes of different functions, the two sides of the graph denote a population split, and the splitting of the lines denote a mutation creating a new gene variant. After the population split, new variants of genes created in one population do not migrate to the other and appear concurrently with new variants of other genes, thus yielding correlations between functionally independent genes.

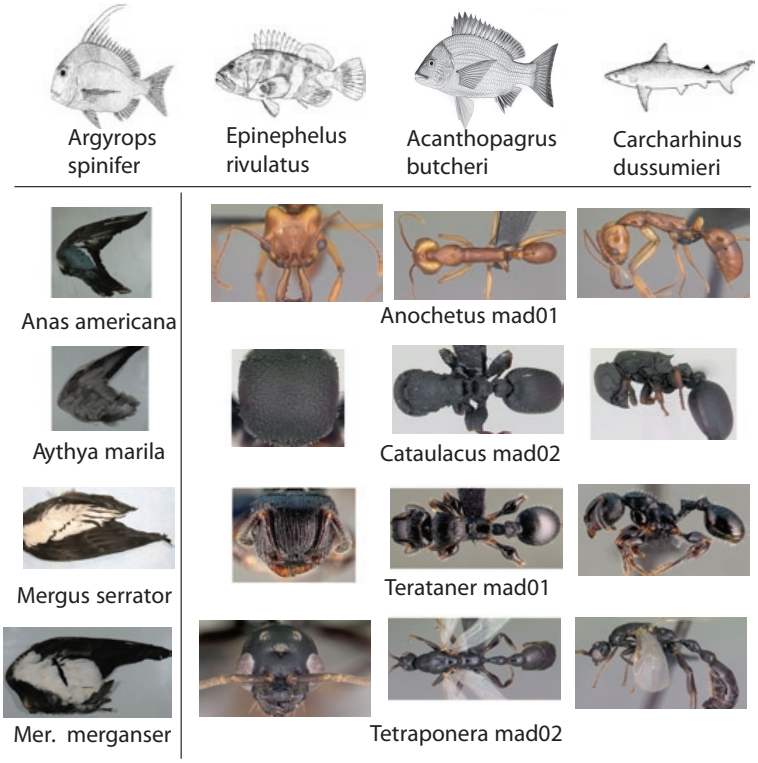

Figure 2. Example of images used in our experiments. (top) illustrations of fish from FishBase; (bottom right) Ant images from head, dorsal and profile views, retrieved from AntWeb; (bottom left) Wing images available at BOLD.

Regular space	Correlation space
$D_{L_1}(u,v) = \sum_{i=1}^{l} \lvert u_i - v_i \rvert$	$D_{CoL_1}(u,v) = \sum_{i=1}^{l} \lvert \hat{u}_i - \hat{v}_i \rvert$
$D_{L_2}(u,v) = \sum_{i=1}^{l} (u_i - v_i)^2$	$D_{CoL_2}(u,v) = \sum_{i=1}^{l} (\hat{u}_i - \hat{v}_i)^2$
$D_C(u,v) = 1 - \frac{u \cdot v}{\lVert u \rVert \lVert v \rVert}$	$D_{CC}(u,v) = 1 - \frac{\hat{u} \cdot \hat{v}}{\lVert \hat{u} \rVert \lVert \hat{v} \rVert}$

We have compared the matching accuracy using different distance metrics, and our experiments (see section 4) show that cosine distance in correlation space (D_{CC}) dominates the other distance measures on all data sets. For this reason we use it as the standard distance metric.

Since the feature vectors for both genes and images are of dimensions significantly higher than the number of training samples, statistical regularization must be used to avoid overfitting. We use the regularized version of CCA suggested by [22]. Generally, two regularization parameters need to be determined: η_m and η_p. We use a single regularization parameter instead, η as follows. Let $X = [x_1\ x_2\ \ldots\ x_n]$ and $Y = [y_1\ y_2\ \ldots\ y_n]$, and denote by λ_M, λ_P the largest eigenvalues of $XX^\top$ and $YY^\top$. We set $\eta_M = \eta\lambda_M$ and similarly for η_P. This way of choosing the regularization parameters is invariant to scale and can be used uniformly across all data sets.

3.2. Kernel Canonical Correlation Analysis

CCA only examines linear transformations. To overcome this limitation, kernelized versions have been proposed (e.g. [24, 5]). Given two non-linear transformations $\phi : M \rightarrow \hat{M}$, $\psi : P \rightarrow \hat{P}$ to high-dimensional spaces, Kernel CCA solves a problem similar to CCA, where x_i is replaced by $\overline{\phi(m_i)} = \phi(m_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(m_j)$, and y_i is replaced by $\overline{\psi(p_i)} = \psi(p_i) - \frac{1}{n}\sum_{j=1}^{n}\psi(p_j)$.

If ψ and ϕ satisfy certain conditions, the solution is efficiently obtained by employing the "kernel trick" which allows the solution to be found without explicitly evaluating ϕ and ψ. Rather, kernel functions $K_M : M \times M \rightarrow R$ and $K_P : P \times P \rightarrow R$ are used. We use Gaussian kernels: $K_M(m_1, m_2) = exp(\frac{-\lVert m_1 - m_2 \rVert^2}{2\sigma_M^2})$, $K_P(p_1, p_2) = exp(\frac{-\lVert p_1 - p_2 \rVert^2}{2\sigma_P^2})$. σ_M and σ_P are Gaussian widths, and similarly to the choice of regularization parameters η_M, η_P above, we use only one parameter instead, τ, and set $\sigma_M = \tau\sqrt{\frac{2}{n(n-1)}\sum_{1 \le i < j \le n}\lVert m_i - m_j \rVert^2}$ and similarly for σ_P. Regularization, the matching process and the distance

metrics are applied similarly to the linear case above.

3.3. Common Discriminant Feature Extraction

CCA minimizes the distances between matching genes and images, while disregarding the distances between non-matching genes and images. An alternative optimization goal combines the minimization of distances between matching pairs with the maximization of distances between non-matching pairs, thus attempting to achieve maximal discriminative ability. This approach has been developed in [10] for the general case where the samples come from several classes.

We have applied the algorithm to the match/no-match problem, where each class contains a single gene/image pair. Our experiments show considerably better performance using ridge-like regularization instead of the local consistency of the original method, so we have modified the algorithm to use ridge regularization.

3.4. Maximal Margin Robot

The MMR method [15] transforms the CCA formalization to a maximal-margin problem, and the corresponding solution can be seen as a generalization of Support Vector Machines for the case where both inputs and outputs are vectors. Rather than maximizing the sum of correlations, MMR maximizes the minimal correlation. Robustness to outliers is maintained through the inclusion of slack variables. This formalization produces the following quadratic programming problem:

$$\min_{W,\xi} \quad \frac{1}{2}\|W\|_F^2 + C\mathbf{1}^T\xi \ , \quad \text{subject to}$$
$$\forall 1 \le i \le n \quad y_i^T W x_i \ge 1 - \xi_i \quad \xi_i \ge 0$$

As a matching score between a genotype x_{new} and a phenotype y_{new_k} we employ $y_{new_k}^T W x_{new}$.

3.5. Preference Margin Optimization

The MMR method considers the minimal correlation as the margin. We propose an alternative formulation which we term "Preference Margin Optimization" (PMO), where the margin is defined as the difference between the correlation of matching pairs and the correlation of non-matching pairs. This is similar to Support Vector Machines (SVMs), in which margins measure separation between classes.

The problem is formulated as the following QP problem:

$$\min_{W,\xi} \quad \frac{1}{2}\|W\|_F^2 + \frac{C}{n}\mathbf{1}^T\xi \ , \quad \text{subject to:}$$
$$\forall i \ne j \quad \xi_{ij} \ge 0, \quad y_i^T W x_i - y_j^T W x_i \ge 1 - \xi_{ij}$$

Since the $n(n-1)$ constraints are computationally demanding, for the fish data set we use only $r = 20$ constraints per training pair, selected as the closest matches using CCA. Our experiments show little improvement for $r > 20$.

3.6. Binary Classification

A simple approach involves reduction to a binary classification problem, with one class representing matching pairs and the other non-matching pairs. To avoid scaling issues, the two data sets are preprocessed by subtracting the mean value and dividing by the mean norm, after which each genotype/phenotype pair is concatenated to one vector. We test this approach by using the SVM classifier, trained on a data set containing positive (+1) samples of matching pairs and negative (-1) samples of non-matching pairs. Then, for matching, we choose the image that is more likely to belong to the positive class according to the trained SVM.

3.7. Nearest Neighbor Transfer

Given a new genetic marker m_{new}, this simple method chooses out of the existing markers the closest one $\arg\min_{i=1}^{n} \|m_i - m_{new}\|$ and selects the image most similar to the corresponding image p_i: $\arg\min_k \|p_{new_k} - p_i\|$.

3.8. Phylogenetic tree based algorithms

Two forms of phylogenetic-tree based identification were used for comparison. In the first method (PT1), a phylogenetic tree is constructed using the UPGMA method [18] for the pairs of genetic markers and images (m_i, p_i) in the training set, where the distances are determined by Euclidean distance between the vector representations of the genes (alignment is not needed since the sequences in BOLD are aligned for each data set). Then, given the new marker m_{new}, the node $v = (m_i, p_i)$ in the tree that minimizes $\|m_i - m_{new}\|$ is found, as well as the nodes $u_k = (m_{i_k}, p_{i_k})$ that minimize $\|p_{i_k} - p_{new_k}\|$ for each of the new images $\{p_{new_k}\}$. The chosen image p_k is that for which the tree distance, measured as number of edges, between v and u_k is minimal. To give this algorithm the benefit of the doubt, ties are not counted in the overall accuracy.

The second method (PT2) is based on a similarity measure between phylogenetic trees, where the similarity between two trees is defined as the linear correlation between the two corresponding distance matrices as in [4]. The chosen image using this method maximizes, by searching over the unknown matching index k, this similarity between the distance matrices of genetic markers $\{m_i\}_{i=1}^N \cup \{m_{new}\}$ and images $\{p_i\}_{i=1}^N \cup \{p_{new_k}\}$.

4. Experiments

We employ the above methods in order to validate the plausibility of genotype-visual-phenotype exploitation, and

to test the suitability of each algorithm for this task. We employ three genetic and five visual data-sets. The genotype is vectorized in two conventional manners: base frequencies (see section 4.2) and the Kimura two-parameter model (K2P) [9]. The images are represented by either the SIFT descriptor [11] or by the C_1 [16] descriptors.

4.1. Data sets

Three gene sequence data sets are employed: Fishes of Australia [23], Ant Diversity in Northern Madagascar [17], and Birds of North America - Phase II project [8]. By locating matching images we constructed data sets containing multiple genotype-phenotype pairs (M, P), where M stands for the COI gene sequence and P is a single image of an animal of that species. No fish or ant species is associated with more than one image, while the bird data set contains several genetic markers and images for each species.

The images were extracted from several sources. A total of 157 fish species illustrations with varying style and quality were retrieved from FishBase (`http://filaman.ifm-geomar.de/`). Images of 26 relevant ant species were available from AntWeb (`http://www.antweb.org/`), with each ant photographed from a profile view, a head view and a dorsal view. The fish and ant images were matched to the BOLD record by the species name. The BOLD repository itself contains 125 wing images of sequenced birds covering 25 species. The images were cropped as needed to remove foreign objects such as text signs, lines and digits. Some bird images were flipped horizontally so that all birds would face the same direction. Some examples are shown in Figure 2.

4.2. Data representation

Representation of the genetic sequences The length of the COI gene used in the experiments is about is about 650 base pairs, depending on the dataset, and it is represented as a string over the alphabet $\{A, C, G, T\}$ of the same length. In this work, we represent genetic sequences of length N as vectors of length $4N$, where each element in the vector corresponds to one nucleotide (i.e., character) in one location in the sequence. In the case where only one sequence is available for a species, the values are all 0 or 1. When r sequences $seq^1, \ldots, seq^r$ are available from different individuals in the same species, the values represent the portion of samples with the corresponding nucleotide in the corresponding location, and thus the resulting vector m is defined by: $m_{4i+j} = \frac{1}{r}|\{1 \leq k \leq r : seq_i^k = j\}|$ for $1 \leq i \leq N$, $1 \leq j \leq 4$, where we enumerate the alphabet as $1..4$. We refer to this representation as "base frequencies". An alternative representation, based on Kimura's two-parameter model (K2P) [9], was also tested for comparison.

Representation of the images The visual descriptors of the images are computed by the bag-of-sift implementation of Andrea Vendaldi available at `http://vision.ucla.edu/~vedaldi/`. This implementation uses hierarchical K-means [12] for partitioning the descriptor space. Keypoints are selected at random locations [13]. Note that the dictionary for this representation was recomputed at each run in order to avoid the use of testing data in the training stage. Using the default parameters, this representation results in vectors of length $11,111$. We also tested an alternative representation, where all images were rescaled to 100×100 pixels, and their C_1 features [16] were computed, using the implementation at `http://cbcl.mit.edu/software-datasets/`. For this representation, each data vector is of dimension $3,728$.

4.3. Experimental procedure

Experiments are carried out using holdout-style cross-validation. In each iteration, 90% of the data is randomly chosen as training set $\{m_i^{tr}, p_i^{tr}\}_{i=1}^{n_{tr}}$ and the remaining 10% is used as test set $\{m_i^{ts}, p_i^{ts}\}_{i=1}^{n_{ts}}$. The matching classifier is trained on the training set, yielding a matching function $f : M \times P \times P \rightarrow \{1, 2\}$. Then for $1 \leq i \leq n_{ts}$, for the genetic marker m_i^{ts}, and for all $j \neq i : 1 \leq j \leq n_{ts}$, the trained classifier matches either the correct image p_i^{ts} or p_j^{ts} to m_i^{ts}. Thus a total of $n_{ts}(n_{ts}-1)$ forced matching tests are carried out in each iteration. The accuracy of an iteration is then the portion of correct matchings out of all matchings: $acc = \frac{1}{n_{ts}(n_{ts}-1)}|\{1 \leq i \neq j \leq n_{ts} : f(m_i^{ts}, p_i^{ts}, p_j^{ts}) = 1\}|$. (In practice the order between p_i^{ts} and p_j^{ts} was randomly switched with probability 0.5 so that the correct matching result would be randomly distributed in $\{1, 2\}$). The results of each experiment include the mean and standard deviation of the accuracy measured over 100 iterations.

In the bird-wing data-set there are several samples for each species, and to avoid trivial identification where we match an image of a previously seen species, no species appears in both the training and test sets, and the choice is always between images of animals from two different previously unseen species.

4.4. Results

We first examine the influence of various parameters on the performance of the algorithms.

Figure 3 (left) shows forced matching results using CCA for varying values of η, the parameter controlling ridge regularization. We set $\eta = 0.05$ for all further experiments. τ is the parameter controlling Gaussian kernel width, and figure 3 (middle) shows how KCCA matching accuracy varies when τ is changed. We set $\tau = 0.25$ for comparison purposes. The DCFE α parameter controls the trade-off between matching pairs distances and non-matching pairs dis-

Metric	%	SD	Metric	%	SD
L_1	75.2	5.77	CoL_1	85.1	4.76
L_2	71.1	7.26	CoL_2	89.3	4.17
C	73.1	7.94	CC	90.5	4.06

Table 1. Matching accuracy (% and standard deviation) for fish using CCA with different distance metrics.

tances. Figure 3 (right) shows DCFE matching results using various values of α, with $\eta = 0.05$. For comparison between the algorithms, we use $\alpha = 2$.

It can be seen from figure 3 that matching accuracy remains stable across a wide range of parameter choices. Our results would change little if other parameters are used. This is also true with regard to the influence of C, the trade-off parameter between margin and slack, in the maximal-margin methods (MMR, PMO, and SVM). We found their performance to be nearly identical over a very wide range of C values, and we use $C = 1$ for all experiments below.

4.4.1 Distance metrics

A distance metric is required to choose the more similar pair of vectors after transformation. We compare several distance metrics for matching with CCA ($\eta = 0.05$). Table 1 shows a comparison that was performed on the fish dataset, between the distance metrics that are described in section 3.1.

4.4.2 Comparison between algorithms

Table 4 shows a comparison between the matching algorithms on the 5 data sets. It can be seen that regularized linear CCA achieves best performance on nearly all data sets. The correlation-based variants are close behind, and the non-correlation-based methods (SVM, NNT, PT1, PT2) are far behind.

The relatively low scores of NNT, PT1 and PT2 suggest that the high matching accuracy is a result of learning correlations between genotype and phenotype, and is harder to achieve by analyzing each separately and then relating the learned structures.

The results also suggest that the reduction to a binary classification problem may not be suitable. A possible reason may be the impoverished nature of this representation, which projects all the data to just one dimension. Another reason may be that non-matches that are "almost-matches" do not affect training of correlation methods much (or at all, depending on the method), while they can harm the training of a binary classifier considerably.

4.4.3 Alternative gene sequence representations

In the representation used for the experiments described above, the similarity between two gene sequences is the number of agreements between the sequences. An alternative representation in which different probabilities are used for transitions and transversions as in the Kimura two-parameter model [9], was tested as well. Results (omitted) show that both representations performs similarly.

4.4.4 Alternative image representations

Besides the experiments using the SIFT "bag-of-features" representation, we also performed additional experiments using an alternative image representation called C_1 [16]. The results depicted in Figure 5 suggest that the C_1 identification accuracy is almost as high as when using the SIFT descriptors, although performance seems to be more sensitive to variations in the regularization parameter η.

5. Summary

By employing forced-matching tests, we present here a large body of evidence that shows that visual identification based on DNA sequences is not "Science Fiction". Our results are statistically significant, and are consistent across data sets and data representations.

From the algorithmic point of view, we have compared a wide variety of existing and new algorithms. The results suggest that that regularized CCA may be the most suitable algorithm for such applications.

Our results may open the way to new and exciting applications, such as "virtual line ups" based on DNA evidence from the crime scene, which can use footage of suspects captured on nearby surveillance cameras. Another application, which may not be as remote as it sounds may be to predict the appearance of extinct animals.

Acknowledgments

The authors would like to thank Tomaso Poggio for his insightful comments and suggestions. This research is supported by the Israel Science Foundation (grants No. 1440/06, 1214/06), the Colton Foundation, and a Raymond and Beverly Sackler Career Development Chair.

References

[1] J. R. Beveridge et al. The csu face identification evaluation system. *Machine Vision and Applications*, 16, 2004.

[2] W. M. Brown, M. George, and A. C. Wilson. Rapid evolution of animal mitochondrial dna. *PNAS*, 76(4), 1979.

[3] K. Chase et al. Genetic basis for systems of skeletal quantitative traits: Principal component analysis of the canid skeleton. *PNAS*, 99:9930 – 9935, 2002.

[4] C. Goh et al. Co-evolution of proteins with their interaction partners. *J. Molecular Biology*, 299(2), 2000.

[5] D. R. Hardoon and J. Shawe-Taylor. Kcca for different level precision in content-based image retrieval. *Int. Workshop on Content-Based Multimedia Indexing*, 2003.

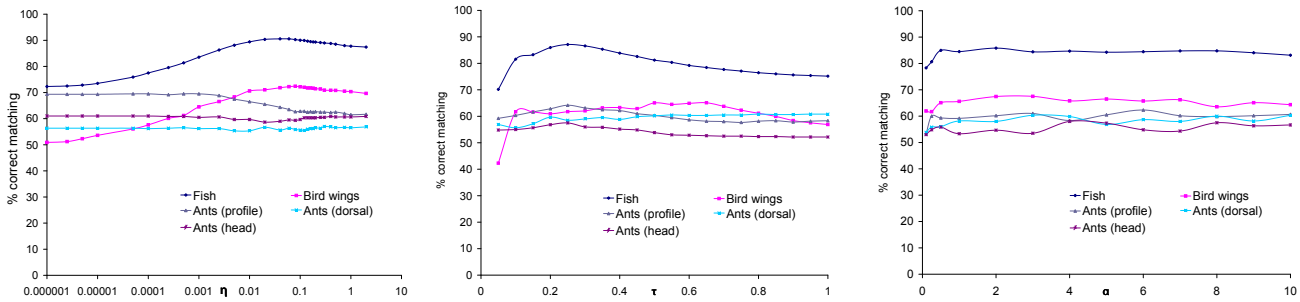

Figure 3. The effects on forced matching accuracy of changing several parameters of the algorithms. (left) CCA performance with varying values of η. (middle) KCCA performance with varying values of τ. (right) DCFE performance with varying values of α. Error bars were omitted to reduce clutter: standard deviation is about 5% for fish, 15% for bird wings and 20% for ants.

Algorithm	Fish	Wing	Ants profile	Ants head	Ants dorsal
CCA	90.5 ± 4.0	72.0 ± 13.7	63.8 ± 21.3	55.8 ± 20.3	59.3 ± 19.9
KCCA	87.1 ± 5.0	61.8 ± 13.7	64.2 ± 20.0	58.5 ± 19.9	57.5 ± 17.8
DCFE	85.8 ± 5.8	67.5 ± 14.6	60.2 ± 23.0	58.0 ± 21.6	54.7 ± 23.5
MMR	86.2 ± 4.9	69.6 ± 13.9	61.7 ± 22.8	56.2 ± 19.9	55.5 ± 23.1
PMO	86.8 ± 5.5	68.5 ± 14.8	60.3 ± 24.4	58.5 ± 18.3	54.2 ± 21.6
SVM	57.6 ± 2.9	53.0 ± 5.2	56.8 ± 17.3	54.7 ± 15.5	54.8 ± 16.8
NNT	68.6 ± 5.6	53.8 ± 9.9	53.8 ± 10.8	54.2 ± 12.3	51.6 ± 8.74
PT1	77.6 ± 6.6	63.2 ± 16.2	57.2 ± 16.0	50.0 ± 20.7	61.8 ± 21.5
PT2	64.1 ± 5.7	57.9 ± 13.5	47.1 ± 22.5	44.2 ± 21.1	50.9 ± 22.9

Figure 4. Matching accuracy (% and standard deviation) on all data sets using all algorithms: Regularized Canonical Correlation Analysis (CCA), Kernel Canonical Correlation Analysis (KCCA), Discriminative Common Feature Extraction (DCFE), Maximal Margin Robot (MMR), Preference Margin Optimization (PMO), Support Vector Machine (SVM), Nearest Neighbor Transfer (NNT), Phylogenetic Tree distance (PT1), Phylogenetic Tree similarity (PT2).

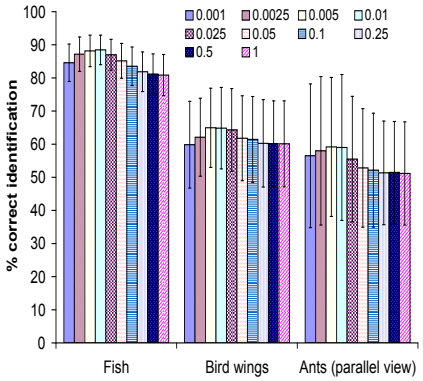

Figure 5. Identification accuracy of fish, bird, and ants (profile-view) using C_1 features, shown for several values of the regularization parameter η. Error bars represent standard deviation.

[6] P. D. N. Hebert et al. Biological identifications through dna barcodes. *Proc. Royal Society Biological Sciences*, 270(1512), 2003.

[7] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, December 1936.

[8] K. C. Kerr et al. Comprehensive dna barcode coverage of north american birds. *Molecular Ecology Notes*,7(4),2007.

[9] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, 1980.

[10] D. Lin and X. Tang. Inter-modality face recognition. In *ECCV*, 2006.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 60(2), 2004.

[12] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[13] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV* 2006.

[14] S. Ratnasingham and P. D. N. Hebert. bold: The barcode of life data system. *Molecular Ecology Notes*,7(10),2007.

[15] T. Sandor Szedmak and D. Hardoon. A metamorphosis of canonical correlation analysis into multivariate maximum margin learning.*Euro. Symp. Artificial Neural Nets.*,2007.

[16] T. Serre et al. Robust object recognition with cortex-like mechanisms. *PAMI*, 29(3), 2007.

[17] M. A. Smith, B. L. Fisher, and P. D. N. Hebert. Dna barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of madagascar. *Phil. Trans. Roy. Soc. Biological Sciences*, 360(1462), 2005.

[18] R. R. Sokal and C. D. Michner. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.

[19] R. A. Sturm and T. N. Frudakis. Eye colour: portals into pigmentation genes and ancestry. *Trends in Genetics*, 20(8):327–332, 2004.

[20] N. B. Sutterm et al. A single igf1 allele is a major determinant of small size in dogs. *Science*, 316(5821), 2007.

[21] G. J. Van Tonder, M. J. Lyons and Y. Ejima. Perception psychology: Visual structure of a Japanese Zen garden. *Nature*, 19, 2002.

[22] H. D. Vinod. Canonical ridge and econometrics of joint production. *J. of Econometrics*, 4(2):147–166, 1976.

[23] R. D. Waard et al. Dna barcoding australia's fish species. *Phil.Trans. Roy. Soc. Biological Sciences*,360(1462),2005.

[24] L. Wolf and A. Shashua. Learning over Sets using Kernel Principal Angles. *J. Machine Learning Res.*, 4(10), 2003.