Quantitative Classification and Natural Clustering of *C. elegans* Behavioral Phenotypes

Wei Geng,[*] Pamela Cosman,[*] Joong-Hwan Baek, † Charles C. Berry, ‡
William R. Schafer[§,1]

*Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407, † School of Electronics, Telecomm. & Computer Engineering, Hankuk Aviation University, Koyang City, South Korea, ‡Biostatistics Unit, UCSD Cancer Center, San Diego, CA 92093, and [§,1]Division of Biology, University of California at San Diego, La Jolla, California 92093-0349.

[1] *Corresponding author*: William R. Schafer, Division of Biology, University of California at San Diego, La Jolla, California 92093-0349, USA. E-mail: wschafer@ucsd.edu Office: (858)-822-0508. Fax: (855)-822-2003.

Running head: Classification of *C. elegans* behavioral phenotypes

Keywords: C. elegans; Classification; Computer Vision; Natural clustering; Machine Learning; Phenotypic analysis

ABSTRACT

Genetic analysis of nervous system function relies on the rigorous description of behavioral phenotypes. However, standard methods for classifying the behavioral patterns of mutant *Caenorhabditis elegans* rely on human observation and are therefore subjective and imprecise. Here we describe the application of machine learning to quantitatively define and classify the behavioral patterns of *C. elegans* nervous system mutants. We have used an automated tracking and image processing system to obtain measurements of a wide range of morphological and behavioral features from recordings of representative mutant types. Using principal component analysis, we represented the behavioral patterns of eight mutant types as data clouds distributed in multidimensional feature space. Cluster analysis using the k-means algorithm made it possible to quantitatively assess the relative similarities between different behavioral phenotypes and to identify natural phenotypic clusters among the data. Since the patterns of phenotypic similarity identified in this study closely paralleled the functional similarities of the mutant gene products, the complex phenotypic signatures obtained from these image data appeared to represent an effective diagnostic of the mutants' underlying molecular defects.

INTRODUCTION

Among the organisms most amenable to the genetic analysis of behavior is the nematode *Caenorhabditis elegans*. *C. elegans* has a simple nervous system consisting of 302 neurons of known position, cell lineage, and synaptic connectivity (Sulston and Horvitz 1977; Sulston *et al.*1983; White *et al*. 1986). Moreover, because of their short generation time, small genome size, and accessibility to germline transformation, these animals are highly amenable to molecular and classical genetics. In principle, the well-defined nervous system of *C. elegans* makes it possible to obtain a reductionist understanding of the neuronal and molecular basis for phenotypes of behavioral mutants. Although precise assays for behavioral abnormalities are critical for neurogenetic studies in *C. elegans*, standard assays for complex behaviors such as locomotion are typically imprecise and subjective. For example, mutants displaying abnormal or uncoordinated ("Unc") movement (Brenner 1974; Hodgkin 1983) are usually classified into descriptive categories such as "kinker", "coiler", "shrinkers", "loopy", "slow", and "sluggish". Although mutants with common molecular defects generally have qualitatively similar behavioral phenotypes, the subjectivity inherent in classifying behavioral patterns by eye makes it difficult if not impossible to assess which mutants have genuinely similar phenotypes based on published descriptions alone.

To address this problem, we have explored the use of machine vision approaches to quantitatively characterize and classify *C. elegans* uncoordinated mutants. In previous work, we built a tracking and imaging system that could follow and record an individual animal's movements over long time periods and save digital image data representing the animal's body posture over the course of the recording (Baek *et al.* 2002). Algorithms were also devised to measure 94 features of a given mutant's body shape or locomotion pattern, making it possible to comprehensively assay multiple aspects of behavior simultaneously. By using these features, it was possible to reliably distinguish examples

of representative mutant types from one another using a binary decision tree algorithm (CART). We therefore reasoned that it might also be possible to use these features to obtain a specific, quantitative definition of a particular mutant phenotype that would be diagnostic of a specific molecular defect and would facilitate quantitative comparisons between different mutant strains.

In this study, we have used image data collected by our automated tracking system to investigate the natural clustering of *C. elegans* behavioral phenotypes. From a complex data set consisting of 253 features measured from behavioral recordings of 797 individuals representing 8 distinct genotypes, we used principal component analysis to represent each mutant type as a cloud of data points in low-dimensional feature space. We have also used k-means clustering and Euclidean distance measurements to explore the natural structure of the behavioral data and to compare the similarities of mutant phenotypic patterns. These results therefore constitute a quantitative definition of several important *C. elegans* behavioral phenotypes, and demonstrate that mutant phenotypes can be clustered using a complex behavioral signature based on quantitative image features.

MATERIALS AND METHODS

Strains and culture methods: Routine culturing of *C. elegans* was performed as described (Brenner 1974). All worms analyzed in these experiments were young adults; fourth-stage larvae were picked the evening before the experiment and tracked the following morning after cultivation at 22°. Experimental animals were allowed to acclimate for 5 minutes before their behavior was analyzed. Plates for tracking experiments were prepared fresh the day of the experiment; a single drop of a saturated LB culture of E. coli strain OP50 was spotted onto a fresh NGM agar plate and allowed to dry for 1 hour before use.

The alleles and predicted products of the genes used in these experiments were as follows: *unc-38 (x20)*, nicotinic acetylcholine receptor alpha-subunit (null allele); *unc-29(x29)*, nicotinic acetylcholine receptor non-alpha-subunit (null allele); *goa-1(n1134)*, G-protein$_o$-alpha-subunit (strong loss-of-function allele); *unc-36 (e251)*; voltage-gated calcium channel alpha-2-subunit (strong loss-of-function allele); *unc-2(mu74)*; N-type voltage-gated calcium channel alpha-1-subunit (null allele); *egl-19(n582)*; L-type voltage-gated calcium channel alpha-1-subunit (partial loss-of-function allele); *nic-1(lj22)*, type 1 glycosyltransferase (partial loss-of-function allele).

Acquisition of image data: *C. elegans* locomotion was tracked with a Zeiss Stemi 2000-C Stereomicroscope mounted with a Cohu High Performance CCD video camera essentially as described (Baek *et al.* 2002). Briefly, a computer-controlled tracker (Parker Automation, SMC-1N) was used to maintain the worms in the center of the optical field of the stereomicroscope during observation. To record the locomotion of an animal, an image frame of the animal was snapped every 0.5 second for at least five minutes. Among those image pixels with values less than or equal to the average value minus three times the standard deviation, the largest connected component was found. The image was then trimmed to the smallest axis-aligned rectangle that contained this component, and saved as eight-bit grayscale data. The dimensions of each image, and the coordinates of

the upper left corner of the rectangle box containing the worm body in the tracker field were also saved simultaneously as the references for the location of an animal in the tracker field at the corresponding time point when the images are snapped. The stereomicroscope was fixed to its largest magnification (50 X) during operation. Depending on the type and the posture of a worm, the number of pixels per trimmed image frame varied. The number of pixels per millimeter was fixed at 312.5 pixel/mm for all worms.

Image Pre-processing: To obtain the clean binary image, the background intensity level of the grayscale image was found first by taking the maximum of the values of the four corner points of the trimmed image (at least one of the corner points is always not part of the worm body). After finding the background level (b), a 5x5 moving window was scanned over the trimmed image, and the mean (m) and standard deviation (s) of the pixels inside the window were computed at every pixel position. If m was less than 0.7b or s was larger than 0.3m, then the pixel was considered to be a pixel of the worm body and was assigned a value 1. In order to clean up the spots inside the worm body, a morphological closing operator (binary dilation followed by erosion) was applied (Gonzalez and Woods 2002). Next, the sequential algorithm for component labeling was used to remove unwanted isolated objects (Jain *et al*, 1995). The connected components were labeled by scanning the image in x and y directions sequentially, and the largest component was selected to guarantee that there will be only one object, the worm, in the binary image.

Image Feature Extraction: All of the software for binarization, skeletonization, and feature extraction was coded in C and implemented on a UNIX machine. Some features (e.g., the area of the worm, that is, the number of pixels which make up the single binary object in the frame) could be computed on a single frame; these were computed for all 600 frames in the sequence. The average value, the maximum value and the minimum value were then computed for these 600 measurements. Some of the maximum and minimum values are outliers introduced by noise or errors during image capture and processing. To avoid using these extreme values, it was more useful to summarize the group statistics with such quantities as the 90^{th} and 10^{th} percentile values out of the population of 600 numbers. Hereafter we use max and min to denote the 90^{th} and 10^{th} percentile values. Other features could not be extracted from a single frame, for example, the movement between two frames, or the movement within 10 seconds (20 frames). Since there are approximately 600 frames total in a sequence, the movement between two frames could be computed 300 times if we take pairs of frames in a non-overlapping fashion, or it could be calculated 599 times taking pairs of frames in a sliding window or overlapping fashion. Likewise, for the movement within 20 frames, we could compute 581 values for overlapping 20-frame intervals. Quantities of this type were calculated in a sliding window fashion. As before, the average, max, min, and other order statistics can be computed from this set of numbers. Features that describe worm body transparency (median pixel value) , and head and tail movement relative to centroid were also measured (W. Geng, unpublished). A complete list of features used in classification, along with their mean and variance for each genotype, is included as supplemental data.

RESULTS

Collection and normalization of behavioral feature data

To explore the natural clustering of behavioral phenotypes using defined quantitative parameters, we collected digital image data from 8 representative genotypes: the standard wild-type strain N2, and 7 loss-of-function mutants affecting different molecules involved in nervous system function. For each genotype, 100 five-minute recordings (98 for *unc-29*, 99 for *unc-2*) were made of individual adult hermaphrodites, with images captured at a frequency of 2 Hz. For each recording, 253 parameters describing aspects of the animal's movement, body texture, or body posture were measured; the feature measurements for a single recording were designated as a single multidimensional data point. We then analyzed the clustering of these 797 data points with the goal of determining the optimal substructure of the behavioral data. In particular, we sought to determine how the feature data clustered in multidimensional space and to then correlate the clustering pattern of the feature data with the known biology of the mutant types in the study.

Standardizing inputs on a set of carefully selected features plays an important role in pattern recognition. Since our features were measured in different units, it was necessary to normalize them on a common scale to avoid one feature dominating others. The outliers introduced by noise and errors during the feature extraction process tend to give false clusters in clustering analysis; thus, the scaling method also needs to be carefully selected to suppress outliers. We evaluated three standard normalization methods: Min-max (linear transformation of the original input range into [-1,1]), Zscore (defined as $x = \frac{f - mean(f)}{stdev(f)}$, where f is the original input feature), and sigmoidal method (Grossman 2002). The Sigmoidal method is defined as $y = \frac{1 - e^{-x}}{1 + e^{-x}}$, where x is the output of Zscore scaling. Figure 1b shows a comparison among different scaling methods. The Sigmoidal method was chosen because it obtains a better balance of limiting outliers and equalizing feature variance on our dataset given our goal of natural clustering.

Representation of phenotypic patterns in multidimensional feature space

To visualize the phenotypic patterns as defined by the selected parameters, we used principal component analysis (PCA) (Duda *et al*. 2001) to obtain a two-dimensional projection of our 253-dimensional data. We observed (Figure 2a) that the data points for each mutant type formed a data cloud that occupied a specific region of feature space. To investigate the distribution of these clouds, we computed the centroid for each mutant type (i.e., the center of the data cloud as measured by Euclidean distance), and considered this to be the prototype for that mutant type (Table 2). Consistent with our expectation, the majority of the worm samples for each type were closer to its respective prototype than were samples from other mutant types (Table 3). Interestingly, the distances between the centers of the mutant data clouds also showed a strong correspondence to the

similarities between the described mutant phenotypes. For example, the clouds for the 4 mutants (*unc-2, unc-36, unc-29,* and *unc-38*) described in the literature as "kinkers" mapped close together in feature space, whereas the wild-type, *goa-1, nic-1* and *egl-19* clouds were more widely separated from the other types and from each other. Moreover, the closest two clusters were *unc-29* and *unc-38* (3.5), encode nicotinic receptor subunits with overlapping functional expression. *unc-2* and *unc-36* (distance 3.6), the next closest clusters, respectively encode α-1 and α-2 voltage-gated calcium channel subunits with nearly coincident expression patterns. This indicates that a simple Euclidean distance in feature space can be used to quantify the relative similarity between different mutant types.

Feature selection and classification of phenotypes

Since one of our main objectives is to identify parameters that define particular mutant types, we wished to identify a small number of features that provide discriminative information. A variance plot (Figure 1a) shows that the top 43 principal components (17% of total PCs) capture over 94% of total variance. This gives a strong indication that a few carefully selected features would represent the data well.

To identify best features for distinguishing any two worm types, we screened the entire feature set using a backward elimination process based on the linear Lagrangian Support Vector Machine classifier (Mangasarian and Musicant 2001; Model *et al* 2001). The support vector machine classifier was used because it generalizes well. The process started from the full feature set. In each iteration, one feature was eliminated from the remaining feature set by evaluating all the possible subsets (n subsets, each containing n-1 features) and selecting the subset that achieves the smallest training error as our next feature set. We used a low training error as an approximation of the importance of that feature. All the features can thus be ranked according to when they are eliminated from the backward elimination process. We repeated this process for all 8 mutant types in a pairwise fashion and generated 28 sequences of ranked features.

Feature subsets that are effective to distinguish all worm types were then selected progressively by choosing the most frequent features that appear on the top of all 28 sequences. For example, the first feature was selected as the feature that appeared most frequent as the No. 1 feature in all 28 sequences. The second feature was selected as the feature appears most frequently as the No. 1 or No. 2 features in all 28 sequences besides the feature that was already in the subset. A simple 1-nearest neighbor (1-NN) classifier with 10-fold cross-validation (Duda *et al*. 2001) was used to evaluate subset performance. To avoid over-fitting, a 10-fold cross validation technique was used. For each feature subset in each trial, we divided data from each worm type randomly into 10 sections. One section (80 worms) was held out for testing and the other 9 sections (720 worms) were used as training data. In subsequent steps in the trial, different testing and training sections were chosen. The classification error was calculated as the average of the 10 iterations for each of the 28 class pairs. For each subset, 50 trials were performed to give an aggregated classification error rate for that subset. We also compared the classification error of the first few principal components using the three scaling methods (Figure 1b).

A small set of features can be readily identified to approximate the dataset by following the cross-validation error curve. Table 4 shows the classification results by using all 253 and a subset of 39 features. The 39-feature subset was selected at the first significant dip location (at k=39) on the error curve. The data were well represented using a subset of 39 features for discriminating phenotypes. These features included several measurements of speed and reversals averaged over different time periods, and worm head and tail width and brightness information (Table 1).

Natural clustering of phenotypic data

To further investigate the clustering of the data points, we applied the k-means clustering algorithm to find the natural clusters in the behavioral data. For this analysis, each data point was treated individually without regard to mutant type. The k-means algorithm is an elementary but very popular clustering method. It enjoys the benefits of making no assumptions about the underlying data probability distributions, and is thus applicable to many problems. Suppose there are to be k clusters with respective centers $C = \{c_1, \ldots c_k\}$ and their corresponding non-overlapping divisions of feature space defined as $D = \{D_1, \ldots D_k\}$. Let $\|.\|^2$ denote "squared Euclidean distance". Our data are x_i : i=1,2,….797. We would like to choose $C = \{c_1, \ldots c_k\}$ so that $C = \arg\min_C \sum_{j=1}^{k} \sum_{x_i \in D_k} \|x_i - c_j\|^2$. While there is no closed form solution to the minimization, Lloyd (1957) demonstrated that an alternating descent algorithm will always converge. The Lloyd algorithm for k-means clustering is an iterative descent algorithm. Starting with an initial set of k representative points, all the points in the data set are assigned to whichever of the k points is closest according to some distance measure, usually Euclidean distance. Next, each of the k representative points is relocated to be the centroid of the data points which just got assigned to it. At this point, we have a new set of k representative points, and can go back to the assignment step. The algorithm iterates between these steps of data point assignment and cluster centroid calculation, until convergence is reached. The final convergence, in general, depends on the initial choice of k representative points. The algorithm does not necessarily find the global optimum, and so often many random initialization seeds are used. We generated sufficiently many (10,000) random initializations for each k and tracked the error at the convergence to be reasonably confident that the global minimum was found. Figures 3a-b show the cluster centers identified by the k-means algorithm; for each case, the centers are marked by black squares. Although the actual k-means clustering was done using all 253 selected features, the data were visualized by showing the first two principal components.

A key issue in k-means clustering is to determine the optimal number of clusters for the data set. We used two algorithms to determine the optimal cluster number for our behavioral data: the gap statistic (Tibshirani et al. 2001) and the information theoretic method (Sugar and James 2003).

The idea of the Gap Statistic is to standardize the graph of $\log(W_k)$ by comparing it to its expectation under an appropriate null reference distribution of the data. W_k is the

total within-cluster sum of squares around the cluster centers, when there are k clusters. Since we have 797 points in our data set, the null reference distribution is generated by drawing 797 samples from a distribution that is uniform along each feature data dimension. This is repeated B times. The expectation of the null reference $E\{\log(W_{kb}^*)\}$ can be estimated as $1/B\sum_{b=1}^{B}\log(W_{kb}^*)$, where W_{kb}^* is the within-cluster sum of squares of the b^{th} reference dataset, and B is the number of reference datasets. The distance between these two curves is defined as the Gap, $Gap(k) = 1/B\sum_{b=1}^{B}\log(W_{kb}^*) - \log(W_k)$, for k=1,…K, where K is the maximum number of clusters defined by the user according to the expected range of clusters. We use a maximum of 10 centers (K =10) and 5 reference datasets (B = 5). The sampling distribution can be measured by $s_k = sd_k\sqrt{1+1/B}$, where sd_k is the standard deviation of the reference null distribution. The formula to calculate the optimal number of clusters k_{opt} can be obtained as the first location where the gap curve starts to drop or level off. That is the first k that satisfies $gap(k) \geq gap(k+1) - as_{k+1}$, where a is a multiplier adjusted to reject null mode. Here it is set to 3.

The Information Theoretic approach tries to find the optimal number of clusters by fitting the within-cluster sum of squares curve (distortion curve) with two hyperbolic curves breaking at the location of the optimal k. The location of the break can be measured in a transformed domain when applying a negative power to the distortion curves. The magnitude of the power is controlled by the dimensionality of the data . Here it is set to –7. The transformed distortion curve usually can be approximated reasonably well by a piecewise linear function consisting of two straight lines with a break, or elbow, at the location of the optimal k. The optimal number of clusters can be easily obtained by finding the biggest jump, which is the difference between the successive points on the transformed distortion curve. The paper provides theoretic justification and points out that this method can also provide suboptimal solutions by finding smaller jumps in the curve. This is particularly appealing given our objective of exploring the substructure of the data.

As shown in Figure 3c-d, both methods identified 6 clusters as the optimal number (Table 4). In this optimal classification, the calcium channel mutants *unc-36* and *unc-2* were grouped into a single cluster and the nicotinic receptor mutants *unc-29* and *unc-38* into another cluster. In addition, the information theoretic approach identified an additional suboptimal solution of 8 clusters with each cluster composed primarily of a single mutant type (Figure 3d and Table 5). Together, these results demonstrated that worms of the same mutant type tend to exhibit similar behavioral patterns and further showed that cluster analysis can be used to assess phenotypic similarities between different mutant classes.

DISCUSSION

Quantitative definition of behavioral mutant phenotypes

We have shown here that quantitative morphological and locomotion features obtained from digital video recordings can be used to distinguish the behavioral phenotypes of *C. elegans* mutants. As shown in Table 3, a reduced set of approximately 40 features is sufficient to identify visibly dissimilar mutant types with very high reliability. Furthermore, these features can often be used to distinguish between types with highly similar phenotypes (e.g. unc-2 and unc-36) that can not be reliably identified even by an experienced human observer. Thus, the parameters in the reduced feature set are likely to have great utility in assessing subtle or modest abnormalities in behavior caused by hypomorphic mutant alleles or by incompletely penetrant dsRNA inhibition.

These studies have also provided insight into the nature of specific mutant phenotypes. For example, *unc-36, unc-29, unc-38* and *unc-2* have all been categorized as "weak kinkers", a term that has been difficult to define precisely. From Table 1, it is apparent that these mutants share many common effects on the variables used in our classification; in particular, all have a substantially higher angle change rate and substantially lower centroid movement and global speed parameters than wild-type. This combination of characters (increased body bending and a decreased rate of movement) thus provides an operational definition of the "kinker" phenotype. Likewise, the combination of increased centroid movement and increased angle change rate provides a functional definition of *goa-1*'s "hyperactive loopy" phenotype, while increased length and length/eccentricity and decreased angle change rate and speed define the "long, slow and floppy" phenotype of *egl-19*. In some cases, significant phenotypic differences were identified that were unnoticed (or unreported) in previous observer-based studies. For example, both *goa-1* and *unc-36* mutants showed particularly large reductions in the ratio of head-to-tail movement, an abnormality whose neural basis could be investigated in future studies. Thus, it has been possible not only to obtain precise quantitative descriptions of phenotypic classes whose definitions had previously been subjective and qualitative, but also to resolve subtle differences within broad classes such as kinker Uncs.

With the collection of larger data sets, it should be possible to use this approach to define and subdivide other widely-cited phenotypic classes of *C. elegans*. For example, it should be possible to obtain precise definitions for other classes of uncoordinated mutants, such as coilers, shrinkers, and loopy mutants. In addition, although we have focused here on the analysis of phenotypes associated with abnormal locomotion, the image parameters we have used in this study could also be used to categorize other classes of behavioral or developmental mutants that involve alterations in body morphology. Such studies would provide valuable insight into the nature of these additional phenotypic types; in addition, it would be interesting from an informatics perspective to learn how the inclusion of genes whose focus of action is outside the neuromuscular system would impact the importance of features used in classification.

Prospects for using behavioral phenotypes for bioinformatic analysis

The application of machine-based pattern recognition methods also allowed us to probe the similarities between different behavioral patterns based on their clustering in multidimensional feature space. In general, the pattern of phenotypic clustering mirrored

the known similarities in molecular function and cellular site of action of the mutant gene products. For example, the *unc-29* and *unc-38*, which respectively encode α and β nicotinic receptor subunits with overlapping expression patterns, formed a single cluster in the optimal clustering and had centers that were the closest together by Euclidean distance (Figure 3a). Likewise, *unc-2* and *unc-36* mutants, which are defective in the α-1 and α-2 subunits respectively of the neuronal N-type calcium channel, formed a single cluster in the optimal k-means clustering, and the centers of these two types' data clouds were relatively close in feature space. In fact, the centers for all four of these types (which have all been designated as kinker Uncs and all encode excitatory ion channels whose focus of action is primarily at body muscle neuromuscular junctions) were closer to one another than to the other Unc mutants or to wild-type. Thus, the quantitative phenotypic signature obtained through behavioral tracking appeared to correspond well to the underlying functional defects of the mutants we analyzed.

We anticipate that this type of comprehensive quantification of mutant behavioral phenotypes will have powerful applications in functional genomic studies. Clustering and pattern recognition analysis of microarray-derived gene expression profiles has provided important information about the likely functions of novel gene products in *C. elegans* and other organisms (Kim *et al*. 2001). In principle, a behavioral phenotype represents a similarly complex quantitative signature whose direct linkage to nervous system activity makes it particularly useful for classifying genes that function in excitable cells. In several genome-wide deletion and RNAi-based knockout surveys undertaken in *C. elegans*, the identification and classification of behavioral and other non-lethal phenotypes has been a crucial limiting factor (Fraser *et al*. 2000; Zipperlen *et al* 2001). Using the machine-based phenotyping approaches described here, it should be possible to record the behavior of an uncharacterized knockout strain, compare its phenotypic pattern to a database of known mutants, and make an informed initial hypothesis about the molecular pathways in which the mutant gene product participates.

LITERATURE CITED

Baek, J., P. Cosman, Z. Feng, J. Silver, and W. R. Schafer, 2002 Using machine vision to analyze and classify C. elegans behavioral phenotypes quantitatively. J. Neurosci Meth **118**: 9 –21.

Brenner, S., 1974 The genetics of Caenorhabditis elegans. Genetics **77**: 77-94.

Duda, R., P. Hart, and D. Stork, 2001 *Pattern Classification*, Second Edition, Wiley, New York, Wiley.

Fraser, A. et al., 2000 Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. Nature **408**: 325-330.

Gonzalez R and R. Woods, 2002 *Digital Image Processing*, Second Edition, Prentice Hall, New Jersey.

Grossman, D., 2002 *Short Course in Data Warehousing and Data Mining*, (online material http://www.ir.iit.edu/~dagr/DataMiningCourse/Spring2001/Notes/Data_Preprocessing.pdf).

Hodgkin, J., 1983 Male phenotypes and mating efficiency in Caenorhabditis elegans. J. Genetics **103**: 43-64.

Jain, R., K. Rangachar, and B. Schunck, 1995 *Machine Vision*, New York, McGraw-Hill.

Kim, S., D. Poole, L. Waggoner, A. Kempf, D. Ramirez, A. Treschow, W. R. Schafer, 2001 Genes affecting the activity of nicotinic receptors involved in C. elegans egg-laying behavior. Genetics **157**:1599-1610.

Mangasarian, O., and R. Musicant, 2001 Lagrangian Support Vector Machines. J. Machine Learning Research **1**: 161-177.

Model, F., P. Adorjan, A. Olek, and C. Piepenbrock, 2001 Feature Selection for DNA methylation based cancer classification. Bioinformatics. **17** Suppl, S1: 57-64.

Sugar, C., and G. James, 2002 Finding the number of clusters in a data set: An information theoretic approach. J. American Statistical Assoc. To appear

Sulston, J. and H. Horvitz, 1977 Post-embryonic cell lineages of the nematode *Caenorhabditis elegans*. Dev. Biol. **56**: 110-156.

Sulston, J., E. Schierenberg, J. White, and J. Thomson, 1983 The embryonic cell lineage of the nematode *Caenorhabditis elegans.* Dev. Biol. **100**: 64-119.

Tibshirani, R., G. Walther, and T. Hastie, 2001 Estimating the number of clusters in a dataset via the Gap statistic. J. Royal Statistical Society Series B. **63**: 411-423.

White, J., E. Southgate, N. Thomson, and S. Brenner 1986 The structure of the Caenorhabditis elegans nervous system. Philos. Trans. R. Soc. Lond. (Biol.) **314**: 1-340.

Zipperlen, P., A. Fraser, R. Kamath, M. Martinez-Campos and J. Ahringer 2001 Roles for 147 embryonic lethal genes on C. elegans chromosome I identified by RNA interference and video microscopy EMBO J. **20**: 3984-3992.

TABLE 1

Features used in mutant characterization

Variable	Statistics	Worm Type							
		w.t.	*goa-1*	*nic-1*	*unc-36*	*unc-38*	*unc-29*	*egl-19*	*unc-2*
CNTMVAVG	Mean	0.05	0.05	0.02	0.02	0.02	0.03	0.02	0.01
(centroid movt. avg)	Std	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
CNTMVMAX	Mean	0.22	0.24	0.11	0.09	0.12	0.14	0.14	0.08
(centroid movt. max)	Std	0.04	0.04	0.03	0.02	0.04	0.04	0.04	0.03
LNECRAVG	Mean	299.92	262.80	220.26	283.35	282.80	301.24	337.51	301.30
(length/eccent. avg)	Std	14.23	14.16	22.38	14.95	17.95	18.37	16.56	17.15
LNECRMIN	Mean	285.58	243.60	205.50	265.65	262.86	277.03	317.83	278.26
(length/eccent. min)	Std	13.67	12.89	19.60	10.87	12.20	15.91	18.34	11.51
LNMFRMAX	Mean	1633.88	1206.41	807.41	1451.85	1231.37	1346.92	2077.69	1383.64
(length/MER. max)	Std	140.95	131.67	199.91	147.69	137.48	178.26	215.69	208.13
ANCHRMAX	Mean	3.89	6.74	7.44	6.35	6.02	5.85	3.47	6.90
(angle change max)	Std	0.40	0.92	1.77	0.80	0.83	0.87	0.62	1.00
ANCHSMAX	Mean	2.69	4.46	5.16	3.88	3.98	3.91	2.42	4.17
(angle change std max)	Std	0.24	0.62	1.12	0.43	0.51	0.58	0.35	0.52
RV20MAX	Mean	4.84	4.95	0.63	2.32	2.32	3.05	2.30	1.42
(max reversal rate in 20s)	Std	1.16	1.05	0.66	0.53	0.82	1.00	0.92	0.61
RV20AVG	Mean	1.22	1.66	0.05	0.54	0.40	0.57	0.41	0.27
(reversal rate 20s average)	Std	0.54	0.51	0.07	0.22	0.21	0.27	0.30	0.13
RV40MAX	Mean	7.13	7.37	0.72	3.24	3.15	4.18	3.12	1.93
(max reversal rate in 40s)	Std	2.03	1.83	0.82	0.81	1.19	1.40	1.47	0.73
RV60MAX	Mean	8.92	9.36	0.74	4.04	3.76	5.00	3.85	2.27

(max reversal rate in 60s)	Std	2.74	2.40	0.86	1.12	1.56	1.76	1.97	0.85
RV80MAX	Mean	10.50	11.31	0.79	4.67	4.33	5.77	4.51	2.63
(max reversal rate in 80s)	Std	3.41	2.99	0.96	1.47	1.88	2.18	2.48	1.06
RV100MAX	Mean	11.81	13.05	0.81	5.29	4.79	6.24	4.95	2.89
(max reversal rate in 100s)	Std	4.15	3.48	1.05	1.70	2.16	2.46	2.83	1.28
RV120MAX	Mean	12.93	14.84	0.85	5.88	5.20	6.94	5.37	3.20
(max reversal rate in 120s)	Std	4.86	3.97	1.11	1.94	2.39	2.79	3.20	1.40
TOTRV	Mean	29.20	39.76	1.08	12.50	9.27	13.28	9.77	6.29
(total reversal)	Std	12.63	12.34	1.64	4.98	4.94	5.98	7.07	3.13
TOTMOVE	Mean	13644.85	13594.72	576.09	3134.27	2807.24	3360.12	5853.53	1415.46
(dist. moved 5 min)	Std	3194.92	4142.15	206.39	1313.27	1197.43	1517.06	1883.90	795.85
PRP50MAX	Mean	1476.94	1140.73	30.04	415.05	437.87	507.94	700.22	269.31
(max disp., 25 sec)	Std	364.64	332.86	30.54	194.80	187.24	239.78	229.43	135.97
PRP40MAX	Mean	1275.91	1038.62	27.20	358.08	371.42	439.09	593.53	235.57
(max disp., 20 sec)	Std	290.06	285.34	26.03	156.10	151.63	199.79	176.45	109.61
PRP30MAX	Mean	1028.84	864.90	24.92	291.26	294.31	364.63	466.44	196.71
(max disp., 15 sec)	Std	217.09	218.06	23.19	114.60	115.89	155.88	133.21	85.91
PRP20MAX	Mean	740.83	673.40	20.53	211.72	214.07	273.01	340.68	143.19
(max disp., 10 sec)	Std	146.34	124.36	16.39	72.52	79.40	116.15	86.81	56.67
PRP10MAX	Mean	412.31	376.43	15.51	118.26	120.07	160.58	198.66	81.64
(max disp., 5 sec)	Std	74.39	55.69	10.37	33.93	39.28	64.60	46.41	27.71
MVHLFAVG	Mean	24.27	24.48	0.99	6.39	6.13	8.29	11.01	3.65
(avg speed, .5 sec)	Std	5.62	7.34	0.35	2.29	2.51	3.64	3.24	1.77
MVHLFMAX	Mean	59.56	53.11	5.66	18.14	19.69	25.74	35.65	12.15
(max speed, .5 sec)	Std	8.35	8.46	3.41	6.63	5.92	7.86	33.15	3.42

LNGTHAVG	Mean	288.94	243.39	207.58	266.00	262.89	276.15	320.12	276.03
(avg length)	Std	13.34	12.80	19.62	10.15	11.57	13.51	14.14	10.70
LNGTHMAX	Mean	299.11	254.46	217.35	276.24	273.91	288.61	331.20	288.04
(max length)	Std	13.68	13.30	20.36	10.88	12.42	13.74	14.49	11.01
LNGTHMIN	Mean	277.77	232.03	197.68	255.51	251.59	263.03	307.94	264.12
(min length)	Std	13.06	12.58	18.72	10.00	10.93	14.13	17.69	13.13
CNLNRAVG	Mean	0.09	0.10	0.13	0.09	0.10	0.09	0.08	0.09
(avg center width/length)	Std	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01
CNLNRMAX	Mean	0.10	0.11	0.14	0.09	0.11	0.10	0.08	0.10
(max center width/length)	Std	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
CNLNRMIN	Mean	0.09	0.09	0.12	0.08	0.09	0.09	0.07	0.08
(min center width/ length)	Std	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
HCTHRMAX	Mean	0.49	0.57	0.53	0.57	0.51	0.51	0.48	0.59
(max head to center thickness	Std	0.02	0.03	0.04	0.04	0.02	0.03	0.09	0.03
HEADBRAVG	Mean	79.48	81.96	81.12	84.10	81.80	82.50	88.86	84.37
(avg head brightness)	Std	5.04	7.63	7.22	8.12	5.86	6.29	5.88	7.25
TAILBRMIN	Mean	49.79	49.07	58.07	71.15	53.25	51.75	65.56	68.29
(min tail brightness)	Std	3.61	4.83	6.91	7.80	3.82	3.55	5.74	7.71
TAILBRMAX	Mean	67.29	64.17	73.20	92.53	70.82	68.14	85.53	87.92
(max tail brightness)	Std	4.90	6.82	7.42	8.64	5.32	5.47	7.42	8.81
HTBRRMAX	Mean	1.65	1.78	1.47	1.24	1.64	1.69	1.36	1.31
(head/tail brightness max)	Std	0.12	0.17	0.19	0.08	0.13	0.13	0.12	0.11
HANGCRMAX	Mean	11.09	11.97	14.24	12.40	11.32	10.61	10.00	13.98
(head angle change max)	Std	0.66	0.93	2.07	1.23	0.80	0.89	0.44	1.63
HDMVHFAVG	Mean	23.49	23.13	7.45	7.12	10.29	12.88	11.58	7.29

(head movt. .5s)	Std	4.93	5.93	2.41	1.78	2.61	4.19	3.53	2.04
HTMVRAVG	Mean	2.46	1.62	2.76	1.50	2.22	2.00	1.68	2.00
(head/tail movt.avg)	Std	0.33	0.18	0.44	0.15	0.37	0.33	0.24	0.26
HDHFTOTMV	Mean	13219.08	12843.83	4325.98	3388.97	4676.34	5148.65	6048.65	2755.74
(head movt. 5min)	Std	2844.91	3295.47	1524.75	836.87	1169.66	1606.87	1631.36	849.89
TLHFTOTMV	Mean	8379.06	11730.28	1737.28	3024.35	2971.15	3652.05	4747.09	1751.84
(tail movt. 5min)	Std	1931.21	3173.85	620.67	816.82	1016.01	1348.93	1337.85	556.15

The mean and standard deviation value of each feature for each worm type are given in the table. Variables used were: CNTMVAVG--average of centroid movement; CNTMVMAX--maximum centroid movement; LNECRAVG--average length/eccentricity ratio; LNECRMIN--minimum length/eccentricity ratio; LNMFRMAX--maximum length/MER (minimum enclosing rectangle) fill ratio; ANCHRMAX--maximum angle change rate; ANCHSMAX--maximum angle change rate standard deviation; RV20MAX,RV20AVG,RV40MAX,RV60MAX,RV80MAX,RV100MAX,RV120MAX --maximum reversal rate sampled at 20, 40, 60, 80, 100,and 120 sec; TOTRV – reversal, 5 min; TOTMOVE--distance moved, 5 min; PRP50MAX ,PRP40MAX, PRP30MAX, PRP20MAX, PRP10MAX--maximum distance moved, sampled at 50, 40, 30, 20 and 10 sec; MVHLFAVG, MVHLFMAX--average, maximum distance moved, 0.5 sec; LNGTHAVG, LNGTHMAX, LNGTHMIN--average, maximum, and minimum length; CNLNRAVG,CNLNRMAX,CNLNRMIN—average, maximum, and minimum center thickness/length ratio; HCTHRMAX – maximum head/center thickness ratio; HEADBRAVG – average head brightness; TAILBRMIN, TAILBRMAX – minimum and maximum tail brightness; HANGCRMAX – maximum angle change rate in head section; HDMVHFAVG – average head distance moved with regard to center , .5 sec; HTMVHFAVG – average head distance /tail distance moved with regard to center , 5 min; HDHFTOTMV – head distance moved with regard to center, 5 min; TLHFTOTMV – tail distance moved with regard to center, 5 min.

TABLE 2

Euclidean distance between prototype centers

	w.t.	***goa-1***	***nic-1***	***unc-36***	***unc-38***	***unc-29***	***egl-19***	***unc-2***
w.t.	**-**	**6.5**	**11.0**	**8.4**	**7.0**	**5.7**	**5.9**	**8.7**
goa-1		**-**	**9.0**	**6.6**	**6.9**	**5.8**	**8.5**	**7.1**
nic-1			**-**	**6.6**	**5.6**	**8.0**	**10.6**	**6.6**
unc-36				**-**	**5.2**	**5.1**	**6.1**	**3.6**
unc-38					**-**	**3.5**	**6.8**	**4.1**
unc-29						**-**	**5.2**	**4.2**
egl-19							**-**	**7.1**
unc-2								**-**

Euclidean distance between prototype centers (cluster centers) measured in 253-dimension feature space. Wild-type--*nic-1* are the furthest; *unc-29--unc-38* and *unc-2 --unc-36* are among the closest. This indicates a simple Euclidean distance in feature space can be used to quantify the relative similarity between mutant types.

TABLE 3

1-NN cross-validation results using 253 features

	wild	*goa-1*	*nic-1*	*unc-36*	*unc-38*	*unc-29*	*egl-19*	*unc-2*
wild	**1.00**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
goa-1	**0.01**	**0.94**	**0**	**0.01**	**0.02**	**0.01**	**0**	**0**
nic-1	**0**	**0**	**0.99**	**0**	**0**	**0**	**0**	**0.01**
unc-36	**0**	**0**	**0**	**0.84**	**0.05**	**0**	**0**	**0.11**
unc-38	**0**	**0**	**0.01**	**0**	**0.80**	**0.19**	**0**	**0**
unc-29	**0**	**0**	**0.01**	**0**	**0.37**	**0.60**	**0**	**0.02**
egl-19	**0**	**0**	**0**	**0.03**	**0.01**	**0.01**	**0.95**	**0**
unc-2	**0**	**0**	**0**	**0.08**	**0.04**	**0**	**0.01**	**0.87**

1-NN cross-validation results using 39 features

	wild	*goa-1*	*nic-1*	*unc-36*	*unc-38*	*unc-29*	*egl-19*	*unc-2*
wild	**1.00**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
goa-1	**0.01**	**0.95**	**0**	**0.01**	**0.02**	**0.01**	**0**	**0**
nic-1	**0**	**0**	**0.99**	**0**	**0**	**0**	**0**	**0.01**
unc-36	**0**	**0**	**0**	**0.87**	**0.03**	**0**	**0**	**0.09**
unc-38	**0**	**0**	**0.02**	**0**	**0.78**	**0.20**	**0**	**0**
unc-29	**0**	**0**	**0.01**	**0**	**0.36**	**0.62**	**0**	**0.01**
egl-19	**0**	**0**	**0**	**0.03**	**0.01**	**0**	**0.95**	**0**
unc-2	**0**	**0**	**0**	**0.09**	**0.04**	**0**	**0.01**	**0.86**

10-fold cross-validated classification result using 1-Nearest Neighbor classifier. The percentage number shows the probability the mutant type specified in the row is classified as being mutant type specified in the column by this classifier. A subset of 39 features achieved a similar performance to the full set.

TABLE 4

Data points classified into 6 clusters

Center	#1	#2	#3	#4	#5	#6
wild	**97**	**2**	**0**	**0**	**1**	**0**
goa-1	**2**	**94**	**0**	**3**	**1**	**0**
nic-1	**0**	**0**	**100**	**0**	**0**	**0**
unc-36	**0**	**0**	**0**	**90**	**10**	**0**
unc-38	**0**	**0**	**2**	**7**	**91**	**0**
unc-29	**1**	**0**	**1**	**9**	**82**	**5**
egl-19	**0**	**0**	**0**	**1**	**0**	**99**
unc-2	**0**	**0**	**2**	**74**	**22**	**1**

Data points were classified into 6 clusters (optimal number of clusters) based on their shortest distance to the cluster centers identified by the k-means algorithm. For the 6-cluster result, *unc-38* and *unc-29* are clustered together, as are *unc-2* and *unc-36*.

TABLE 5

Data points classified into 8 clusters

Center	#1	#2	#3	#4	#5	#6	#7	#8
wild	**97**	**2**	**0**	**0**	**1**	**0**	**0**	**0**
goa-1	**2**	**93**	**0**	**4**	**1**	**0**	**0**	**0**
nic-1	**0**	**0**	**97**	**1**	**2**	**0**	**0**	**0**
unc-36	**0**	**0**	**0**	**70**	**5**	**2**	**0**	**23**
unc-38	**0**	**0**	**1**	**4**	**69**	**24**	**0**	**2**
unc-29	**0**	**0**	**0**	**5**	**26**	**64**	**1**	**2**
egl-19	**0**	**0**	**0**	**2**	**0**	**1**	**97**	**0**
unc-2	**0**	**0**	**1**	**15**	**15**	**1**	**1**	**66**

Data points were classified into 8 clusters (suboptimal number of clusters) based on their shortest distance to the cluster centers identified by the k-means algorithm. For the 8-cluster result, the majority of the samples belong to the right clusters.

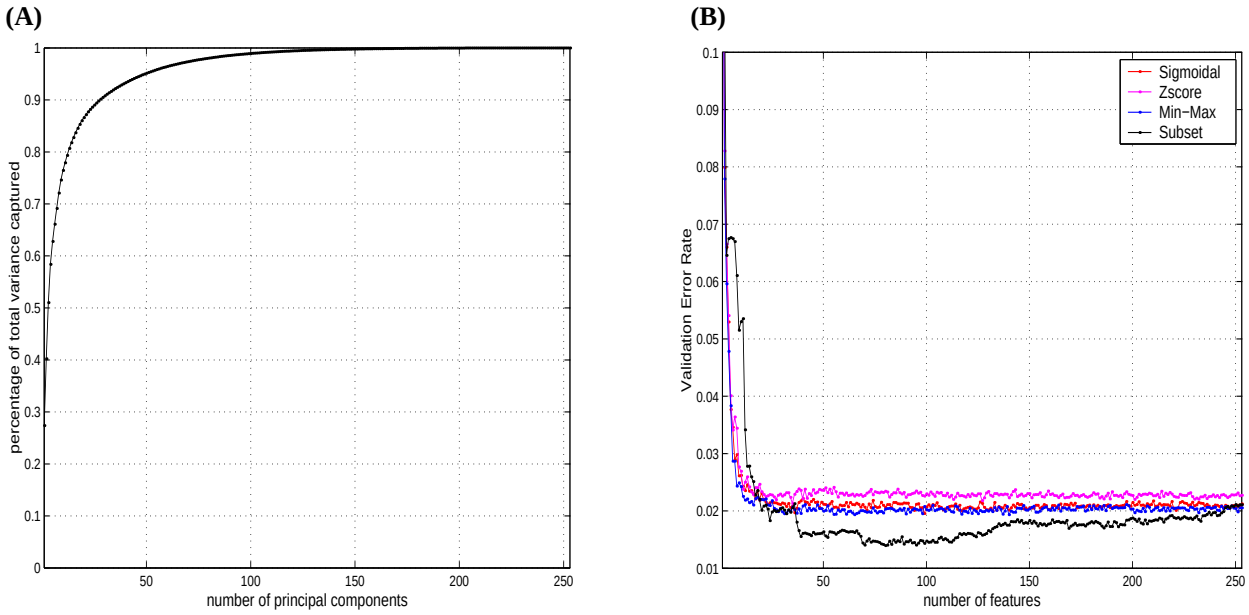

FIGURE 1: Feature data pre-processing and representation. **(A)** Percentage of the total variance captured by the first few principal components (PCs) shows the evidence that feature data may be represented in lower-dimensional space. The top 43 principal components (PCs) capture over 94% of the total variance. **(B)** Comparison between different scaling methods and feature subset. The blue, red, and magenta curves represent the 1 Nearest Neighbor (1-NN) classification error rate using Min-Max, Sigmoidal, and Zscore scaling respectively. The error was an average of 50 trials of 10-fold cross-validation result for each method. The features were selected from the first few Principal Components of the entire 253 input features. All three scaling methods achieved similar performance, with the sigmoidal method slightly outperforming the other two. The fact that the error curves level off indicates most of the useful information for classification is heavily concentrated in the very first few PCs. The black curve shows the same cross-validation test but with a subset of features selected by a backward elimination method. The black curve also shows the adverse effect of increasing error rate with more features added.

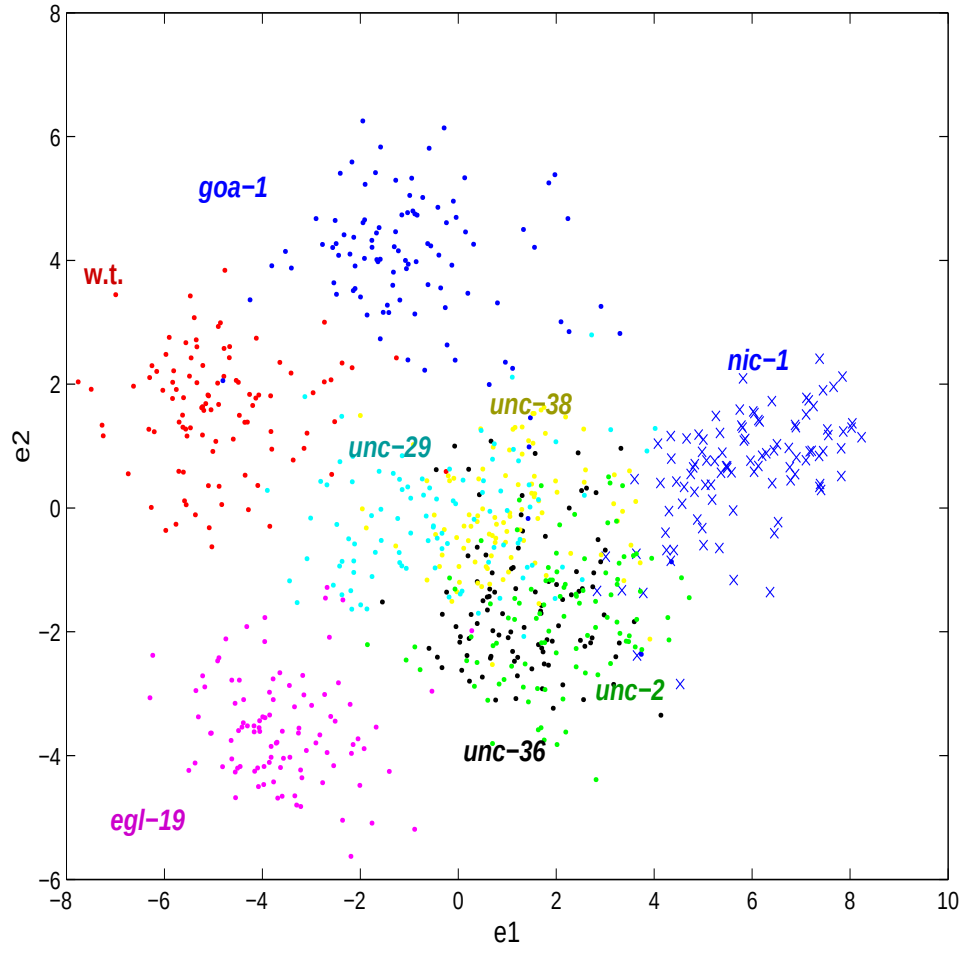

FIGURE 2: Distribution of behavioral data points in feature space. **(A)** The plot shows all 797 data points represented in their first two principal components space using sigmoidal scaling. The data points from the same mutant type are marked by the same color. The data points tend to form fairly tight data clouds for each worm type around each respective prototype, indicating a strong similarity within the mutant types.

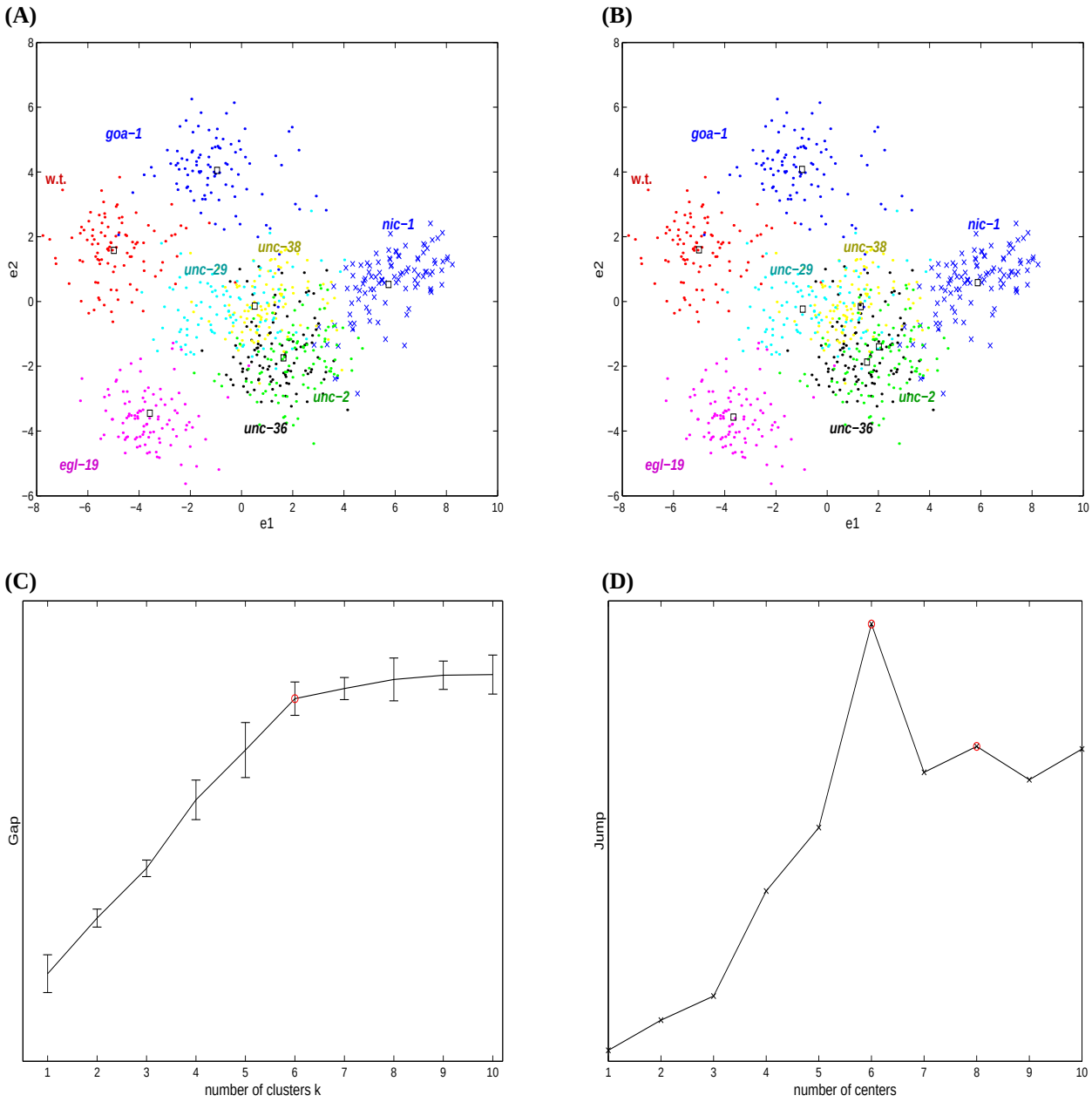

FIGURE 3: Natural clustering results. **A) and B)** Cluster centers found by k-means algorithm, k=6 and 8. The prototype centers were marked as black squares. **(C)** Gap plot by gap statistic method. The optimal number of clusters, marked by red circle, was identified as the gap curve first started to level off. **(D)** Jump plot by information theoretic method. The optimal and suboptimal number of clusters, marked by red circles, were identified as the most and second most significant peaks.