Gene discovery by e-genetics: *Drosophila* odor and taste receptors

Junhyong Kim[1,2] and John R. Carlson[1,*]

[1]Department of Molecular, Cellular and Developmental Biology, Yale University, P.O. Box 208103, New Haven, CT 06520-8103, USA
[2]Department of Ecology and Evolutionary Biology and Department of Statistics, Yale University, P.O. Box 208106, New Haven, CT 06520-8106, USA
*Author for correspondence (e-mail: john.carlson@yale.edu)

Journal of Cell Science 115, 1107-1112 (2002)

Summary

A new algorithm that examines DNA databases for proteins that have a particular structure, as opposed to a particular sequence, represents a novel 'e-genetics' approach to gene discovery. The algorithm has successfully identified new G-protein-coupled receptors, which have a characteristic seven-transmembrane-domain structure, from the *Drosophila* genome database. In particular, it has revealed novel families of odor receptors and taste receptors, which had long eluded identification by other means. The two new gene families, the *Or* and *Gr* genes, are expressed in neurons of olfactory and taste sensilla and are highly divergent from all other known G-protein-coupled receptor genes. Modification of the algorithm should allow identification of other classes of multitransmembrane-domain protein.

Key words: Receptor, Olfaction, Taste, *Drosophila*, G-protein-coupled receptor

Introduction

Drosophila is an excellent organism in which to study the cell biology of olfactory and taste system function and development (Sengupta and Carlson, 2000). These chemosensory systems are composed of relatively few cells, their responses can be conveniently measured either behaviorally or through physiological measurements of individual cells and *Drosophila* offer the advantages of powerful genetics and a sequenced genome.

The fly has two pairs of olfactory organs: the antennae and maxillary palps. Each antenna is covered with ~500 sensilla (sensory hairs), and the maxillary palps each contain 60 such hairs. These sensilla are innervated by up to four neurons. The major taste organ of the fly is a mouthpart called the labellum, which bears ~70 sensilla. These are sensitive to a variety of compounds and endow the flies with the ability to detect a broad range of tastants. Most of the labellar sensilla contain four chemosensory neurons.

To understand the remarkable sensitivity of the fly's olfactory system, its ability to distinguish among odors and the developmental mechanisms by which the system generates a rich diversity of olfactory receptor neurons, enormous efforts were made to isolate odor receptors. Many laboratories have tried to isolate odor receptor genes from a variety of insects over the years, using a host of strategies. The approaches have included genetic screens, subtractive cDNA screens, enhancer trap screens and biochemical approaches. After the isolation of odor receptor genes from vertebrates (Buck and Axel, 1991) and *Caenorhabditis elegans* (Troemel et al., 1995), researchers sought insect orthologs by using a plethora of low-stringency hybridization and PCR techniques, all of which were unsuccessful. Following the accumulation of substantial amounts of genomic sequence by the *Drosophila* genome project, however, we successfully used a bioinformatics screen to identify these receptors. The approach we have employed may be useful in searches for genes encoding other receptors and for genes encoding other classes of protein. It may also be useful in subclassifying proteins of a particular type and in making predictions about the ligands they bind.

e-Genetics as a means to identify novel multitransmembrane-domain protein genes

Our approach, which we term 'e-genetics', is based on an electronic search through DNA databases – the fly genomic sequence in this case – for genes encoding proteins that have multiple transmembrane domains (Clyne et al., 1999b). An underlying assumption in our search was that fly odor receptors are G-protein-coupled receptors (GPCRs). This assumption was based on available data concerning insect olfactory transduction (Boekhoff et al., 1990a; Boekhoff et al., 1990b; Hildebrand and Shepherd, 1997) and on an analogy to odor receptors in other organisms. Members of the GPCR superfamily are extremely divergent in sequence, but they have a common structure, each containing seven transmembrane domains. We therefore designed a search for proteins that have this particular structure rather than a particular sequence.

To identify proteins by virtue of their structure, we developed a computer algorithm that examines DNA databases, identifies open reading frames (ORFs) and maps the predicted proteins into an n-dimensional protein space (Clyne et al., 1999b; Kim et al., 2000). The construction of the space is critical: our aim was to design a space in which GPCRs would map to one particular region of the space (Fig. 1). A fundamental principle of the design is that the space should support interpolation, that is, a newly identified protein that maps to a region of the space inhabited by previously identified GPCRs should be likely to encode a GPCR.

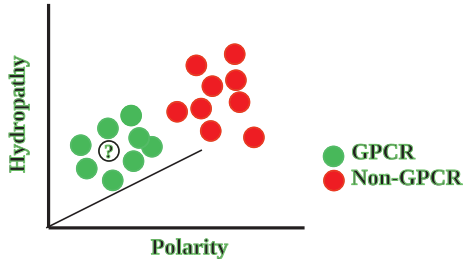

Fig. 1. Construction of an n-dimensional protein space that allows interpolation. Each dimension represents a tested variable. Only three dimensions are shown. Tested variables included hydropathy, polarity, pI, pKa, molecular weight, and amino acid composition. Adapted from Warr et al. (Warr et al., 2001).

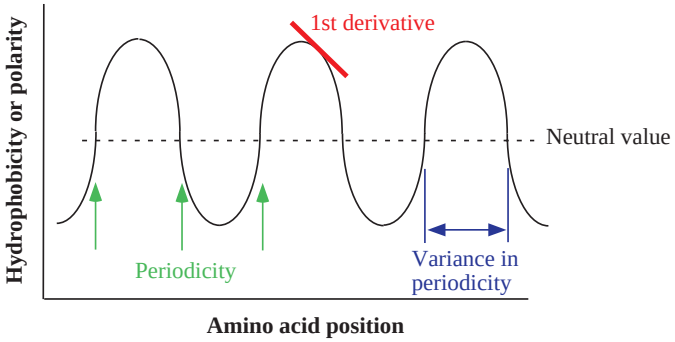

Fig. 2. Refined parameters useful in distinguishing GPCRs from non-GPCRs. A sliding window recognizer is used to characterize the structure of a protein. A portion of an idealized GPCR is shown. Parameters selected as being particularly useful were (1) average periodicity of the hydrophobicity function; (2) average periodicity of the polarity function; (3) variance in the periodicity of the polarity function; (4) variance in the first derivative of the polarity function; and (5) amino acid usage index. Adapted from (Warr et al., 2001); parameters are described in more detail in Kim et al. (Kim et al., 2000).

To develop the algorithm, we used a training set of 750 GPCRs and 1000 non-GPCRs extracted from the SwissProt protein database. To construct the space we tested various parameters to determine whether they were useful for distinguishing GPCRs from non-GPCRs. The space shown in Fig. 1 contains three dimensions, for ease of illustration, but in fact we used an n-dimensional space and tested a number of parameters, including hydropathy, polarity, pI, pKa, molecular weight and amino acid composition. Moreover, we tested numerous refinements of these parameters in an effort to describe the multitransmembrane-domain structure of GPCRs more precisely and to identify structural characteristics of GPCRs that distinguish them from non-GPCRs. We paid particular attention to physical properties of GPCRs that alternate periodically. For example, with respect to hydropathy, GPCRs contain alternating stretches of hydrophilic and hydrophobic residues. We used a sliding-window recognizer to describe the pattern of alternation (von Heijne, 1992; von Heijne, 1994). As illustrated in Fig. 2, local hydropathy can be seen to alternate as a function of residue position within the protein, ranging from regions of high hydrophobicity to low hydrophobicity. The sliding-window recognizer allowed us to quantify the linear organization of such physical and chemical properties along the length of the protein. Characteristics such as periodicity were quantified statistically (e.g. the average derivative of the sliding-window profile).

After testing 70 variables, we selected a set of five that together were particularly useful for distinguishing GPCRs from non-GPCRs: (1) the average periodicity of the hydrophobicity function, which describes the frequency with which the hydrophobicity function crosses a neutral value; (2) the average periodicity of a polarity function, which is related to the hydrophobicity function; (3) the variance in the periodicity of the polarity function; (4) the variance in the first derivative of the polarity function; and (5) an amino acid usage index. (Note that this parameter set differs slightly from that described by Kim et al. (Kim et al., 2000), who report a later version of the algorithm than that used by Clyne et al. (Clyne et al., 1999b).)

These parameters together define a space in which GPCRs and non-GPCRs can be resolved (see below). We next sought a quantitative means of optimally distinguishing the two classes of proteins. For this purpose we used a non-parametric variant of a linear discriminant function (Gnandesikan, 1977; Kim et al., 2000), which is a means of separating two classes of entities. In three dimensions the function can be represented as the plane that best separates GPCRs from non-GPCRs (Fig. 3). Thus this function is used to classify an individual protein as a GPCR or a non-GPCR and can thus predict whether a given ORF encodes a GPCR.

To determine how well the algorithm works, we tested it on a set of 100 GPCRs and 100 non-GPCRs extracted from SwissProt. These proteins were then mapped into the space and classified with the discriminant function. The algorithm correctly classified 96% of the GPCRs as GPCRs, and none of the non-GPCRs was incorrectly classified as a GPCR (i.e. the percentage of false positives was 0) (Fig. 4).

Although these results were very encouraging, the proteins in the test set were all full-length proteins, and we needed to determine how well the algorithm performs with portions of protein sequence. Most *Drosophila* genes contain introns, and therefore the majority of the ORFs encode only a portion of a protein; likewise the ORFs from EST databases generally encode small regions of proteins. We therefore tested 100-amino-acid stretches of both GPCRs and non-GPCRs from the test set and found that the algorithm correctly classified >90% of GPCRs as GPCRs and incorrectly classified only 4% of non-GPCRs as GPCRs (Fig. 4). Moreover, when protein fragments of various lengths (50-400 amino acid residues) were tested, the algorithm consistently performed better than all sequence-

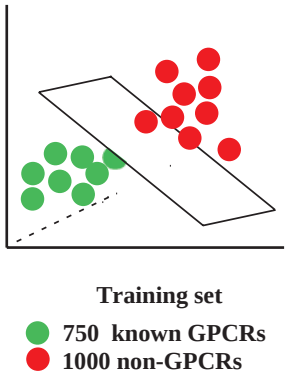

Fig. 3. Setting a discriminant function to maximally separate GPCRs from non-GPCRs in protein space. In a three-dimensional space, the function appears as a plane. The function was established using the training set of 750 known GPCRs and 1000 non-GPCRs. The function is used to classify novel proteins as either GPCRs or non-GPCRs, according to which side of the plane they map.

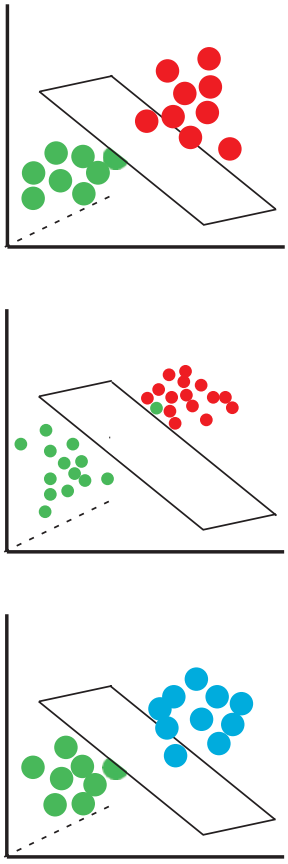

Fig. 4. Testing the algorithm. The top panel shows that the algorithm correctly identified 96% of a test set of 100 known GPCRs and produced no false positives. The middle panel shows the performance of the algorithm on 100 amino-acid stretches of GPCRs and non-GPCRs. The bottom panel shows the performance of the algorithm, following retraining, with a set of GPCRs and ion channels.

based algorithms with which we compared it (pattern matching searches using PROSITE and PRINTS databases, a profile matching search using pfscan and a hidden Markov-Model search using Pfam; see Kim et al. (Kim et al., 2000) for details.

Our algorithm, called the quasi-periodic feature classifier (QFC), may have worked well with short sequences of GPCRs because the discriminating variables, such as amino acid usage, tended to yield consistent values even with short sequences. Since transmembrane domains are short, and in many cases the intervening loops of GPCRs are also short, even a stretch as short as 100 residues of a GPCR may contain two or even three transmembrane domains. By contrast, some other algorithms, especially those based on hidden Markov models, may require longer sequences than are available in individual ESTs or exons. It is possible that the relative success of a structure-based search such as ours may in some respect simply reflect a higher degree of conservation for certain structural properties than for the linear sequence of amino acids.

The protein space was designed to distinguish GPCRs from other proteins on the basis of the multitransmembrane-domain structure of GPCRs. One might ask how other kinds of multitransmembrane-domain proteins, such as ion channels, are classified. Early versions of the algorithm did in fact identify a number of ion channels and transporters, along with GPCRs, from DNA databases. However, by altering the choice of variables we trained the algorithm to distinguish between GPCRs and ion channels. In a test of this retrained algorithm, 96% of full-length GPCRs were correctly identified, and only 3% of ion channels were misidentified as GPCRs (Fig. 4).

Use of the algorithm to discover odor receptor genes

An early version of the algorithm was used to scan an initial release of sequence from the Berkeley *Drosophila* Genome Project (Clyne et al., 1999b). Although only a small portion of the genome had been sequenced, we expected that if the entire genome contained a large number of odor receptor genes, then even this small portion of the genome would be likely to contain a few. The algorithm took ~1 minute of computer time to identify a list of ORFs encoding candidate GPCRs. The list included some previously identified GPCRs, as well as some known ion channels and transporters. Most of the ORFs encoded only a small number of transmembrane domains, in many cases two or three. To determine which of these ORFs represented exons that were spliced to neighboring exons so as to encode seven transmembrane domains, we used the *Drosophila* consensus splice site sequences to identify flanking exons that might be spliced to the identified ORF (Mount, 1992). In this manner we identified full-length candidates for GPCR genes.

To test whether any of these GPCR genes encode odor receptors, we designed PCR primers that spanned introns of these genes and used RT-PCR to determine whether any were expressed in olfactory organs. One of the first genes tested is in fact found to be expressed in the antenna but in no other tested tissue. A second gene on a short list of tested candidates has sequence similarity to the first, is also expressed in the antenna and lies within 500 bp of the first on chromosome 2. We subsequently found by BLAST searching (Altschul et al., 1990) that these two genes define a large family of genes, the *Or* genes, that are widely dispersed through the chromosomes and surprisingly divergent in sequence. Different members were found to be expressed in different subsets of olfactory receptor neurons, as expected of odor receptors. In fact, the number and distribution of neurons expressing an individual odor receptor gene were reminiscent of the pattern of neurons exhibiting a particular odor-response spectrum, as determined in physiological measurements of individual neurons (de Bruyne et al., 1999; de Bruyne et al., 2001). The identity of the *Or* genes as odor receptor genes was further supported by analysis of mutants of the *acj6 (abnormal chemosensory jump)* gene, which encodes a POU-domain transcription factor (Clyne et al., 1999a). A subset of olfactory receptor neurons in these mutants shows alterations in odor specificity, which suggested that the odor receptor genes are improperly regulated in these neurons. A subset of *Or* genes was then found to be improperly regulated in olfactory receptor neurons of *acj6* mutants, which is consistent with the hypothesis that they encode odor receptors (Clyne et al., 1999b). The *Or* genes were independently identified (Gao and Chess, 1999; Vosshall et al., 1999) and characterized (Gao et al., 2000; Vosshall et al., 2000) by others, and recent functional evidence from overexpression of one *Or* gene,

Or43a, in the antenna and in *Xenopus* oocytes has confirmed its identity as an odor receptor gene (Stortkuhl and Kettler, 2001; Wetzel et al., 2001).

Discovery of taste receptor genes

Another sensory system whose receptors had remained elusive was taste. Despite a great deal of effort, insect taste receptor genes had not been cloned. We were therefore very interested to find that another gene identified by the algorithm as a GPCR is expressed in the labellum, the major taste organ of the fly (Clyne et al., 2000). This gene proved to be a member of another large multigene family widely dispersed in the genome: the *Gr* genes (for gustatory receptor). The family members encode proteins that have ~7 predicted transmembrane domains and are even more divergent in sequence than the *Or* genes. Among the first 19 full-length *Gr* genes identified, 18 were found to be expressed in the labellum by RT-PCR. Moreover, their expression is highly specific, in the sense that virtually none are expressed in heads from which taste organs have been removed; likewise, only a small fraction are expressed in the thorax or abdomen. To test whether these genes are expressed in taste neurons, we used the *pox-neuro* mutant, in which taste sensilla are transformed into mechanosensory sensilla (Awasaki and Kimura, 1997; Dambly-Chaudiere et al., 1992; Nottebohm et al., 1992; Nottebohm et al., 1994). In wild-type flies, most taste sensilla contain four taste neurons and one mechanosensory neuron; wild-type mechanosensory sensilla contain a single, mechanosensory, neuron. In the labellum of *pox-neuro*, expression of nearly every *Gr* gene tested was abolished, which is consistent with their expression in the taste neurons of wild-type flies (Clyne et al., 2000).

These results have been extended by others (Scott et al., 2001), who carried out in situ hybridization with 56 members of the *Gr* family and were able to detect expression of six *Gr* genes in subpopulations of labellar neurons. Although 47 of the genes showed no detectable expression in adult head tissue by in situ hybridization, three genes were shown to be expressed in the antenna, which suggests that although most members of the family encode taste receptors, some encode odor receptors. Additional evidence for the expression of several *Gr* genes in taste neurons was obtained by using *Gr* promoters to drive reporter gene expression (Scott et al., 2001; Dunipace et al., 2001).

The simplest interpretation of all these results is that many, if not all, *Gr* genes encode taste receptors. Functional evidence that a *Gr* gene, *Gr5a*, encodes a taste receptor for the sugar trehalose was recently obtained from physiological and behavioral studies of mutants and from transgenic rescue experiments (Dahanukar et al., 2001). The identity of *Gr5a* as a trehalose receptor is also supported by a correlation between trehalose reception and sequence polymorphisms in the *Gr5a* gene and by a correlation between trehalose reception and *Gr5a* expression (Ueno et al., 2001).

e-Genetics as a tool for gene discovery

All methods of gene discovery rest on assumptions. Many genetic approaches assume limited functional redundancy, which is a risky assumption in the case of large multigene families. Many molecular approaches make assumptions about levels and specificity of gene or protein expression and can be difficult to apply to genes expressed in very small subsets of cells. A new approach to gene discovery, which we call e-Genetics, has been generated by the vast expansion of sequence data from genomics projects. This approach, in which genes are identified electronically by specially designed algorithms, also rests on assumptions, however.

A critical assumption made by conventional *in silico* searching with the BLAST algorithm (Altschul et al., 1990) and similar tools is that the target protein resembles a previously identified protein in its linear sequence of amino acids. This assumption in turn rests on three postulates: (i) that proteins are related by evolutionary descent; (ii) that evolutionarily related proteins contain regions of conserved function; (iii) that regions of conserved function contain conserved sequences. The third postulate, linking sequence and function, is basic to both computational and experimental gene searches; however, it is subject to limitations. First, the sequences of some proteins evolve rapidly, especially in the case of a gene family such as chemosensory receptors whose members duplicate and diverge rapidly to meet new environmental challenges or opportunities. A second example of functionally related proteins whose sequence similarity may be especially limited is proteins whose function is carried out largely by general structural features rather than by numerous specific interactions critically dependent on particular amino-acid residues. For example, a chemosensory receptor may be able to tolerate a relatively wide variety of mutations in its transmembrane domains and still be functional by virtue of its ability to bind and signal the presence of at least some hydrophobic ligands. There may be more numerous sequence constraints on the ability of many enzymes to bind a particular substrate and cofactors tightly and to catalyze a particular chemical reaction.

Conventional sequence searching may still be effective in identifying proteins whose sequence has diverged markedly. For example, it may be possible to reduce the evolutionary distance between the target protein and the known protein by ‘phylogenetic walking’, that is, by traversing the distance in a series of steps, each from one organism to a related organism. However, such a strategy is useful only if the size of individual steps is sufficiently reduced. Note that, although odor receptor genes had been isolated from *C. elegans* (Troemel et al., 1995), the phylogenetic distances between vertebrates, *C. elegans*, and *Drosophila* are comparable, and sequence information from the worm was not useful in identifying odor receptors in the fly. Thus there are cases in which an entirely different strategy may be more effective, for example, one like our protein-structure-based approach.

The search for proteins on the basis of structural similarity assumes that proteins that have similar functions have similar structures. Although there is support for this assumption (e.g. Wilson et al., 2000), the extent of structural similarity may be limited. However, there are many cases in which structure is conserved more prominently than sequence (Murzin et al., 1995), as illustrated by the superfamily of GPCRs. GPCRs and other multi-membrane-spanning proteins are particularly well suited for the analysis described here in that many of their structural features can be detected and described quantitatively without a complete *ab initio* structure prediction.

All computational search strategies entail a trade-off between specificity (which produces a low frequency of false positives) and sensitivity (which increases the fraction of targets identified). In the case of sequence-based searches, one can often estimate the specificity by calculating the probability that a particular motif of n residues will occur by chance in a given sequence. In the case of structure-based searches, it is difficult to predict how best to set the search parameters, and attempts to increase the sensitivity of detection lead to an increase in the frequency of false positives. However, it is possible to tolerate a high false-positive rate if there are efficient secondary screens available to validate candidates or if there is auxiliary information, such as genetic map positions of target genes. A high false-negative rate can be tolerated if the target belongs to a large multigene family whose members can be identified by BLAST searches once a single founding member is identified.

What other genes might be discovered through this form of e-genetics? In addition to discovering other novel GPCRs, the algorithm described here might be modified to identify other kinds of protein. We have shown that the algorithm can be trained to distinguish ion channels from GPCRs, suggesting a use in identifying new ion channel genes from a variety of organisms. There is a wide diversity of ligand-gated, voltage-gated and gap junction channels, many of which contain four or six transmembrane domains, and it seems likely that additional types of channel genes remain to be discovered. In fact, connexins, components of gap junction channels that have four transmembrane domains, provide another example of genes whose invertebrate homologs remained elusive for many years: many unsuccessful efforts were made to isolate invertebrate homologs by sequence similarity. When their apparent homologs were finally identified by genetic means (Krishnan et al., 1993; Phelan et al., 1998b), they were found also to have a four-transmembrane-domain structure like those of vertebrates, but their primary sequences are unrelated to those of vertebrates (Phelan et al., 1998a). It also seems likely that through selection of appropriate variables the algorithm might be modified to recognize transporters, which have multiple transmembrane domains, or perhaps receptors related to Frizzled, which has seven transmembrane domains and which has been implicated in pattern formation in *Drosophila* and other species.

Table 1. Subclassification of GPCRs with the QFC algorithm (variant) and a quadratic discrimination function

Classification problem	Training data set % correctly identified		Test data set % correctly identified	
Class A versus Class BC GPCRs	A	BC	A	BC
	99.6	97.6	99.6	84.1
Large ligand versus small ligand Class A GPCRs	Large	Small	Large	Small
	92.0	96.0	86.4	93.8

The algorithm was used to distinguish Class A vs Class BC GPCRs and to distinguish large ligand versus small ligand Class A GPCRs. Results are shown for the training data sets (1163 Class A, 126 Class BC, 413 large ligand and 592 small ligand GPCRs). Test set data are from cross-validation tests in which each sequence is deleted from the training set, the algorithm is trained with the reduced data set and then the deleted sequence is classified.

In addition to discovering genes, our algorithm and others that analyze features of protein structure (e.g. Jakkola et al., 1999; Kim, 2001) may be useful in classifying and characterizing known genes. Toward this end we have been extending our work by using different kinds of statistical method to describe the quasi-periodical alternation of GPCR structure and different kinds of discriminant functions to classify GPCRs. We have been able to distinguish class A (rhodopsin-like) and class BC (secretin-like and metabotropic glutamate) GPCRs (GPCRDB; http://www.gpcr.org/7tm/) in preliminary experiments using a quadratic, as opposed to linear, discriminant function (Table 1). Moreover, this preliminary analysis has allowed subclassification of large ligand-binding versus small ligand-binding class A GPCRs. Further development of this approach may provide finer subtyping of multi-transmembrane-domain proteins and a novel method of ligand prediction, all without direct reliance on the linear amino-acid sequence.

Conclusion/perspectives

This e-genetic approach to gene discovery offers great opportunities for further development. We optimized our particular algorithm well enough to identify the receptors we were seeking, but more training might greatly increase its resolving power. Further development of this and similar algorithms is attractive in part because of the speed and efficiency of the e-genetic approach. In contrast to the many years of micropipetting in our laboratory and others in the quest for insect odor and taste receptor genes, it took only one minute to run the algorithm (followed of course by primer design and RT-PCR analysis). As more and more DNA sequence data accumulate from model organisms, disease-bearing organisms and organisms from the distant reaches of phylogeny, opportunities for gene discovery will only increase. The ability to search sensitively for new proteins on the basis of structure as well as sequence may increase the richness of the bounty.

Our work is supported by the Merck Genome Research Institute (J.K.) and by the NIH and a McKnight Investigator Award (J.C.).

References

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Awasaki, T. and Kimura, K. (1997). pox-neuro is required for development of chemosensory bristles in Drosophila. *J. Neurobiology* **32**, 707-721.

Boekhoff, I., Raming, K. and Breer, H. (1990a). Pheromone-induced stimulation of inositol-trisphosphate formation in insect antennae is mediated by G-proteins. *J. Comp. Physiol. B* **160**, 99-103.

Boekhoff, I., Strotmann, J., Raming, K., Tareilus, E. and Breer, H. (1990b). Odorant-sensitive phospholipase C in insect antennae. *Cell. Signal.* **2**, 49-56.

Buck, L. and Axel, R. (1991). A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* **65**, 175-187.

Clyne, P. J., Certel, S. J., de Bruyne, M., Zaslavsky, L., Johnson, W. A. and Carlson, J. R. (1999a). The odor specificities of a subset of olfactory receptor neurons are governed by Acj6, a POU-domain transcription factor. *Neuron* **22**, 339-347.

Clyne, P. J., Warr, C. G., Freeman, M. R., Lessing, D., Kim, J. H. and Carlson, J. R. (1999b). A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in *Drosophila*. *Neuron* **22**, 327-338.

Clyne, P., Warr, C. and Carlson, J. (2000). Candidate taste receptors in *Drosophila*. *Science* **287**, 1830-1834.

Dahanukar, A., Foster, K., van der Goes van Naters, W. and Carlson, J. R. (2001). A *Gr* receptor is required for response to the sugar trehalose in taste neurons of *Drosophila*. *Nat. Neurosci.* **4**, 1182-1186.

Dambly-Chaudiere, C., Jamet, E., Burri, M., Bopp, D., Basler, K., Hafen, E., Dumont, N., Spielmann, P., Ghysen, A. and Noll, M. (1992). The paired box gene pox neuro: a determinant of poly-innervated sense organs in *Drosophila*. *Cell* **69,** 159-172.

de Bruyne, M., Clyne, P. J. and Carlson, J. R. (1999). Odor coding in a model olfactory organ: the *Drosophila* maxillary palp. *J. Neuroscience* **19**, 4520-4532.

de Bruyne, M., Foster, K. and Carlson, J. (2001). Odor coding in the *Drosophila* antenna. *Neuron* **30**, 537-552.

Dunipace, L., Meister, S., McNealy, C. and Amrein, H. (2001). Spatially restricted expression of candidate taste receptors in the *Drosophila* gustatory system. *Curr. Biol.* **11**, 822-835.

Gao, Q. and Chess, A. (1999). Identification of candidate *Drosophila* olfactory receptors from genomic DNA sequence. *Genomics* **60**, 31-39.

Gao, Q., Yuan, B. and Chess, A. (2000). Convergent projections of *Drosophila* olfactory neurons to specific glomeruli in the antennal lobe. *Nat. Neurosci.* **3**, 780-785.

Gnandesikan, R. (1977). *Methods for statistical data analysis of multivariate observations.* New York: John Wiley & Sons.

Hildebrand, J. G. and Shepherd, G. M. (1997). Mechanisms of olfactory discrimination: converging evidence for common principles across phyla. *Annu. Rev. Neurosci.* **20**, 595-631.

Jaakkola, T., Diekhans, M. and Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **4**, 149-158.

Kim, J. (2001). Descartes' fly: Geometry of genomic annotation. *Funct. Integr. Genomics* **1**, 241-249.

Kim, J., Moriyama, E., Warr, C., Clyne, P. and Carlson, J. (2000). Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. *Bioinformatics* **16**, 767-775.

Krishnan, S., Frei, E., Swain, G. and Wyman, R. (1993). Passover: a gene required for synaptic connectivity in the giant fiber system of *Drosophila*. *Cell* **73**, 967-977.

Mount, S. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* **20**, 4255-4262.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Nottebohm, E., Dambly-Chaudiere, C. and Ghysen, A. (1992). Connectivity of chemosensory neurons is controlled by the gene *poxn* in *Drosophila*. *Nature* **359**, 829-832.

Nottebohm, E., Usui, A., Therianos, S., Kimura, K., Dambly-Chaudiere, C. and Ghysen, A. (1994). The gene poxn controls different steps of the formation of chemosensory organs in *Drosophila*. *Neuron* **12**, 25-34.

Phelan, P., Bacon, J., Davies, J., Stebbings, L., Todman, M., Avery, L., Baines, R., Barnes, T., Ford, C., Hekimi, S. et al. (1998a). Innexins: a family of invertebrate gap-junction proteins. *Trends Genet.* **14**, 348-349.

Phelan, P., Stebbings, L., Baines, R., Bacon, J., Davies, J. and Ford, C. (1998b). *Drosophila* Shaking-B protein forms gap junctions in paired *Xenopus* oocytes. *Nature* **391**, 181-184.

Scott, K., Brady, R., Cravchik, A., Morozov, P., Rzhetsky, A., Zuker, C. and Axel, R. (2001). A chemosensory gene family encoding candidate gustatory and olfactory receptors in *Drosophila*. *Cell* **104**, 661-673.

Sengupta, P. and Carlson, J. (2000). Genetic models of chemoreception. In *The neurobiology of taste & smell* (ed. W. S. T. Finger, D. Restrepo), pp. 41-72. New York: Wiley Press.

Stortkuhl, K. and Kettler, R. (2001). Functional analysis of an olfactory receptor in *Drosophila* melanogaster. *Proc. Natl. Acad. Sci. USA* **98**, 9381-9385.

Troemel, E. R., Chou, J. H., Dwyer, N. D., Colbert, H. A. and Bargmann, C. I. (1995). Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans*. *Cell* **83**, 207-218.

Ueno, K., Ohta, M., Morita, H., Mikuni, Y., Nakajima, S., Yamamoto, K. and Isono, K. (2001). Trehalose sensitivity in *Drosophila* correlates with mutations in and expression of the gustatory receptor gene *Gr5a*. *Curr. Biol.* **11**, 1451-1455.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487-494.

von Heijne, G. (1994). Decoding the signals of membrane protein sequences. In *Membrane Protein Structure* (ed. S. H. White), pp. 27-40. New York: Oxford Univ. Press.

Vosshall, L., Wong, A. and Axel, R. (2000). An olfactory sensory map in the fly brain. *Cell* **102**, 147-159.

Vosshall, L. B., Amrein, H., Morozov, P. S., Rzhetsky, A. and Axel, R. (1999). A spatial map of olfactory receptor expression in the *Drosophila* antenna. *Cell* **96**, 725-736.

Warr, C., Clyne, P., de Bruyne, M., Kim, J. and Carlson, J. (2001). Olfaction in *Drosophila*: coding, genetics, and e-genetics. *Chem. Senses* **26**, 201-206.

Wetzel, C., Behrendt, H., Gisselmann, G., Stortkuhl, K., Hovemann, B. and Hatt, H. (2001). Functional expression and characterization of a *Drosophila* odorant receptor in a heterologous cell system. *Proc. Natl. Acad. Sci. USA* **98**, 9377-9380.

Wilson, C. A., Kreychman, J. and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233-249.