Repertoires of G protein-coupled receptors for *Ciona*-specific neuropeptides

Akira Shiraishi[a], Toshimi Okuda[a], Natsuko Miyasaka[a], Tomohiro Osugi[a], Yasushi Okuno[b], Jun Inoue[c], and Honoo Satake[a,1]

[a]Bioorganic Research Institute, Suntory Foundation for Life Sciences, 619-0284 Kyoto, Japan; [b]Department of Biomedical Intelligence, Graduate School of Medicine, Kyoto University, 606-8507 Kyoto, Japan; and [c]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, 904-0495 Okinawa, Japan

Edited by Thomas P. Sakmar, The Rockefeller University, New York, NY, and accepted by Editorial Board Member Jeremy Nathans March 11, 2019 (received for review September 26, 2018)

Neuropeptides play pivotal roles in various biological events in the nervous, neuroendocrine, and endocrine systems, and are correlated with both physiological functions and unique behavioral traits of animals. Elucidation of functional interaction between neuropeptides and receptors is a crucial step for the verification of their biological roles and evolutionary processes. However, most receptors for novel peptides remain to be identified. Here, we show the identification of multiple G protein-coupled receptors (GPCRs) for species-specific neuropeptides of the vertebrate sister group, *Ciona intestinalis* Type A, by combining machine learning and experimental validation. We developed an original peptide descriptor-incorporated support vector machine and used it to predict 22 neuropeptide–GPCR pairs. Of note, signaling assays of the predicted pairs identified 1 homologous and 11 *Ciona*-specific neuropeptide–GPCR pairs for a 41% hit rate: the respective GPCRs for Ci-GALP, Ci-NTLP-2, Ci-LF-1, Ci-LF-2, Ci-LF-5, Ci-LF-6, Ci-LF-7, Ci-LF-8, Ci-YFV-1, and Ci-YFV-3. Interestingly, molecular phylogenetic tree analysis revealed that these receptors, excluding the Ci-GALP receptor, were evolutionarily unrelated to any other known peptide GPCRs, confirming that these GPCRs constitute unprecedented neuropeptide receptor clusters. Altogether, these results verified the neuropeptide–GPCR pairs in the protochordate and evolutionary lineages of neuropeptide GPCRs, and pave the way for investigating the endogenous roles of novel neuropeptides in the closest relatives of vertebrates and the evolutionary processes of neuropeptidergic systems throughout chordates. In addition, the present study also indicates the versatility of the machine-learning–assisted strategy for the identification of novel peptide–receptor pairs in various organisms.

machine learning | peptide descriptor | deorphanization | neuropeptide | G protein-coupled receptor

Ascidians (tunicates) are invertebrate chordates and the phylogenetically closest living relatives of vertebrates (1–3). Such a critical phylogenetic position sheds light on the significance of investigating the evolutionary process and diversity of biological systems throughout the chordates, including the nervous, neuroendocrine, and endocrine systems (1, 4). Neuropeptides play various pivotal roles in these systems as multifunctional signaling molecules, and the majority of cognate receptors for neuropeptides belong to the G protein-coupled receptor (GPCR) superfamily (5, 6). Thus, the elucidation of specific neuropeptide–GPCR pairs is a primary step in the investigation of the biological roles of neuropeptides, their underlying regulatory mechanisms, and their evolutionary history. In the cosmopolitan species of ascidians, *Ciona intestinalis* Type A (*Ciona robusta*), many major neuropeptides (~40) have so far been characterized by purification, cDNA cloning, and peptidomic approaches (7–13). These neuropeptides are classified into two groups. The first group includes homologs or prototypes of vertebrate neuropeptides: cholecystokinin, calcitonin, gonadotropin-releasing hormones (GnRHs), galanin-like peptides (GALP), tachykinin, and vasopressin (7–13). The molecular characterization of *Ciona* neuropeptides substantiated that this invertebrate chordate conserves a greater number of neuropeptide homologs than protostomes (e.g., *Caenorhabditis elegans* and *Drosophila melanogaster*) and other invertebrate deuterostomes (7–13), confirming the evolutionary and phylogenetic relatedness of ascidians to vertebrates. The second group includes *Ciona*-specific novel neuropeptides, namely Ci-NTLPs, Ci-LFs, and Ci-YFV/Ls (*SI Appendix*, Fig. S1 and Table S1), which share neither consensus motifs nor sequence similarity with any other peptides (8, 9). The presence of both homologous and species-specific neuropeptides highlights this phylogenetic relative of vertebrates as a prominent model organism for studies of molecular and functional conservation and specialization in neuropeptidergic systems during chordate evolution. To date, ~160 GPCRs have been predicted and categorized into five major groups (glutamate, rhodopsin, adhesion, frizzled, and secretin) in *Ciona* (14). Furthermore, GPCRs for *Ciona* tachykinins (Ci-TKs) (10), GnRHs (11), cholecystokinin (12), and vasopressin (13) have been elucidated based on the similarity of their sequences to vertebrate homologs. These findings are in good agreement with the principle that GPCRs for homologous neuropeptides possess sequence similarity to homologous GPCRs conserved in other species. In contrast, GPCRs for novel neuropeptides cannot be predicted based on sequence similarity, which has hampered the identification

Significance

Elucidation of neuropeptide–receptor pairs is essential for the investigation of peptidergic signalling processes. Although sequence alignment and molecular phylogenetic analysis can easily predict G protein-coupled receptors for homologous neuropeptides, these methods cannot predict receptors for novel peptides, so many neuropeptide–receptor pairs remain to be identified. We used our original machine-learning system, peptide descriptor-incorporated support vector machine, to predict multiple neuropeptide–receptor pairs of the vertebrate sister group, *Ciona robusta*. The *Ciona*-specific neuropeptide–receptor pairs were validated with cell-based pharmacological assays, showing biological roles for the neuropeptides in a model protochordate. Because of the critical phylogenetic position of *Ciona*, the present study also elucidates the evolutionary processes underlying neuropeptidergic systems in chordates.

Author contributions: H.S. designed research; A.S., T. Okuda, N.M., T. Osugi, Y.O., J.I., and H.S. performed research; A.S. and H.S. contributed new reagents/analytic tools; A.S., T. Okuda, T. Osugi, Y.O., J.I., and H.S. analyzed data; and A.S., T. Osugi, Y.O., J.I., and H.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. T.P.S. is a guest editor invited by the Editorial Board.

[1]To whom correspondence should be addressed. Email: satake@sunbor.or.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1816640116/-/DCSupplemental.

Published online April 1, 2019.

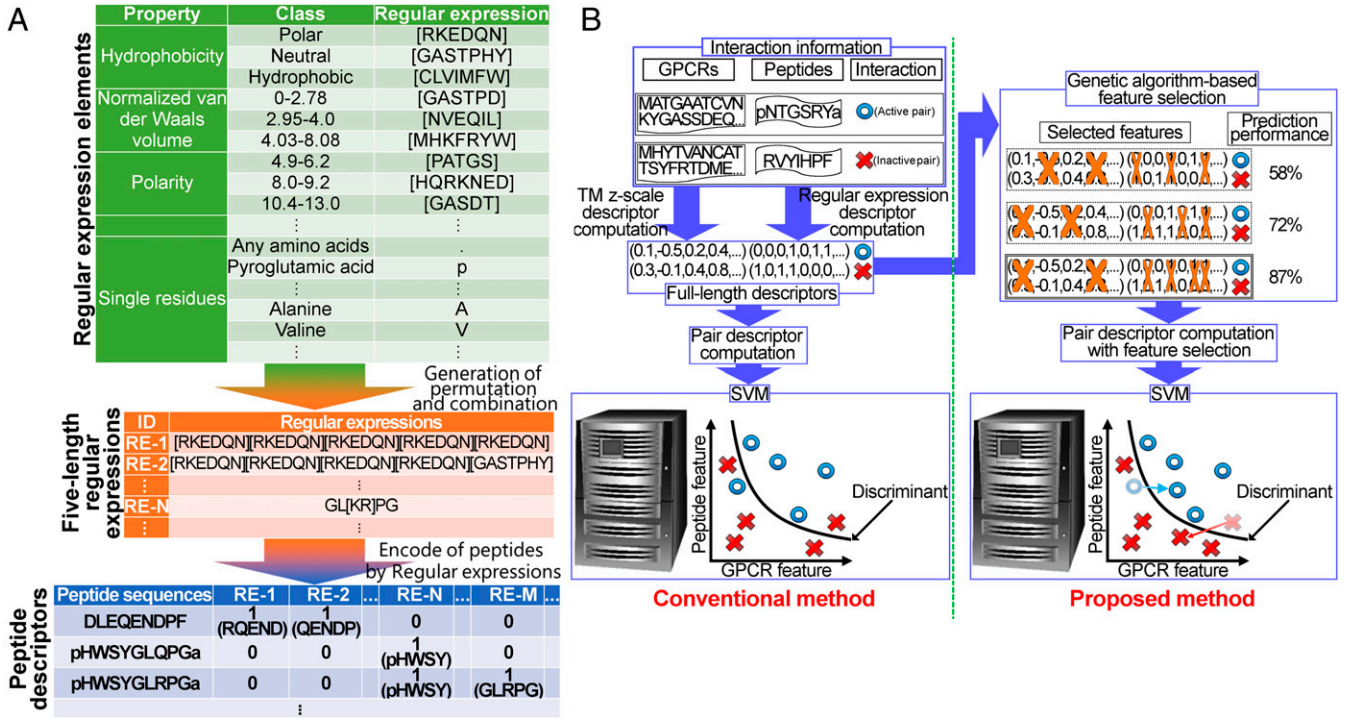

Fig. 1. Overview of PD-incorporated SVM. (*A*) Regular expression-based PDs were generated by concatenating five elements, which represent amino acid categories, and regular expressions for these fragments were calculated. Ambiguous residue categories for regular expressions are listed in *SI Appendix*, Table S2. (*B*) Conventional prediction methods include computation of kernels, which are necessary for SVM to learn and predict interaction pairs, from all of the elements of a GPCR and ligand descriptors. In contrast, the proposed method includes a step for GAFS. In GAFS, descriptor sets were selected to improve the prediction performance with the AUC, which was measured by LOSO cross-validation.

of GPCRs for these neuropeptides. Indeed, no GPCRs for the aforementioned novel neuropeptides (Ci-NTLPs, Ci-LFs, and Ci-YFV/Ls) have ever been identified because these neuropeptides share neither consensus motifs nor sequence similarity with any other peptide. Thus, their cognate GPCRs cannot be predicted by multiple-sequence alignment-based molecular phylogenetic analyses. Similarly, although recent advances in transcriptomes and peptidomes have led to the discovery of numerous putative highly conserved and novel neuropeptides and their cognate receptor candidates (8, 15), many novel GPCRs still remain to be deorphanized.

To date, reverse-pharmacology techniques have been employed for the elucidation of novel ligand–GPCR pairs (16). However, the reverse-pharmacology strategy for deorphanization of GPCRs is analogous to gambling and not systematic: it is time-consuming, costly, and serendipitous. Additionally, limited information regarding GPCR tertiary structures and variations in ligand-receptor binding modes has hampered tertiary structure-based prediction or virtual screening of peptide ligands for orphan GPCRs, including homology modeling. Indeed, only a few low molecular-weight molecules, but not peptides, have been characterized as novel ligands for GPCRs (17–20). These shortcomings indicate the need for a new general and systematic approach for the identification of various novel peptide–GPCR pairs.

Statistical machine learning has been used to predict various ligand–receptor pairs (21–24). In the chemical genomics-based strategy, known ligand–receptor pair information is encoded as numerical vectors (descriptors) or kernels representing amino acid sequences or physicochemical properties, which are input to a machine-learning system, such as a support vector machine (SVM). Indeed, machine-learning systems were used to predict multiple novel ligand–protein pairs using integrated pattern recognition of chemical properties and sequence information of ligands and receptors (25). We previously predicted low molecular-weight drug candidates for human GPCRs using this machine learning system (21, 26). These findings demonstrate the potential of machine learning in the prediction of *Ciona* peptide–GPCR pairs. However, no peptide descriptors (PDs) are available for machine learning for the reliable and efficient prediction of neuropeptide–GPCR pairs (21, 26).

In this study, we identified 12 (11 *Ciona*-specific and 1 homologous) neuropeptide–GPCR pairs by a combination of an originally developed machine-learning system, PD-incorporated SVM, and experimental evidence for specific signaling by the predicted neuropeptide–GPCR pairs, and verified unprecedented phylogenetic relatedness of GPCRs for neuropeptides.

Results

CPI Data Collection. A total of 1,352 compound–protein interactions (CPIs) were collected from IUPHAR, GPCR-SARfari, and UniProt annotations and literature and used as the training dataset. These were composed of 531 human, 310 mouse, 379 vertebrate (vertebrates other than humans and mice), and 132 invertebrate (nonascidian invertebrates) CPIs (Dataset S1). Subsequently, collected GPCRs or peptides were converted into descriptors for machine learning (21, 26, 27). Molecular descriptors for low molecular-weight compounds (28, 29) and proteins (30, 31) have been available for machine-learning–based prediction of CPIs (21, 26). However, chemical descriptors are limited to low molecular-weight compounds due to the computational burden imposed by larger compounds, and protein descriptors cannot be used with short amino acid peptides due to the sparse information available for them for machine learning. To develop PDs possessing peptide physicochemical and biological properties, we initially designed PDs composed of regular expressions (Fig. 1*A* and *SI Appendix*, Table S2), which are 1- to 5-aa sequences comprising any amino acid and their physicochemical properties defined by PROFEAT categories (32). The PDs generated 25,935,478-dimensional bit (0, 1) vectors, which represent the absence and presence of subsequences matching the regular expressions. GPCRs were encoded with a

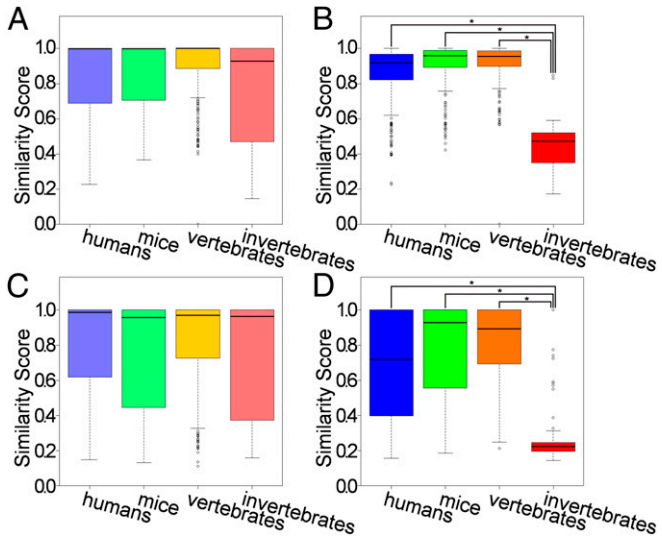

Fig. 2. Boxplots of SSs for GPCRs and peptides. (*A*) SSs for GPCRs against other GPCRs in the same subsets (humans, mice, other vertebrates, and nonascidian invertebrates), (*B*) SSs for GPCRs against other GPCRs in other subsets, (*C*) SSs for peptides against other peptides in the same subsets, and (*D*) SSs for peptides against other peptides in other subsets. $*P < 0.05$.

transmembrane (TM) *z*-scale descriptor according to our previous study (21).

Subsequently, we estimated the distribution of similarity scores (SSs) between each peptide or GPCR and the samples most similar to themselves or other subsets, as previously described (21). Tanimoto coefficients (33) of TM *z*-scale descriptors and the aforementioned original PDs were used for the estimation of the SSs of the GPCRs and peptides, respectively. The average GPCR SSs of humans, mice, vertebrates, and invertebrates with themselves were 0.849 ± 0.020, 0.860 ± 0.020, 0.905 ± 0.016, and 0.762 ± 0.026, respectively (Fig. 2*A* and *SI Appendix*, Fig. S2*A*). The average GPCR SSs of humans, mice, vertebrates, and invertebrates with other subsets were 0.855 ± 0.015, 0.906 ± 0.012, 0.912 ± 0.012, and 0.436 ± 0.014, respectively (Fig. 2*B* and *SI Appendix*, Fig. S2*B*). Although GPCR SSs of humans, mice, and vertebrates were all higher than 0.8, the SSs of invertebrates were less, at 0.436, indicating the dissimilarity of our collected invertebrate GPCRs. The average peptide SSs of humans, mice, vertebrates, and invertebrates with themselves were 0.800 ± 0.025, 0.751 ± 0.029, 0.823 ± 0.024, and 0.314 ± 0.031, respectively (Fig. 2*C* and *SI Appendix*, Fig. S2*C*). The average peptide SSs of humans, mice, vertebrates, and invertebrates with other subsets were 0.676 ± 0.029, 0.778 ± 0.027, 0.822 ± 0.021, and 0.268 ± 0.017, respectively (Fig. 2*D* and *SI Appendix*, Fig. S2*D*). Similar to GPCR SSs, the invertebrate peptide SS (0.268) was extremely small compared with the SSs of humans, mice, and vertebrates (>0.6). This invertebrate-specific distribution of SSs (Fig. 2 and *SI Appendix*, Fig. S2) represents the sequence varieties of invertebrate GPCRs and peptides, but vertebrate GPCRs and peptides contain more orthologs than those in other species. To estimate prediction performance of species-specific CPIs, we evaluated the performance using leave-one-species-out (LOSO) validation (21, 26, 28).

PD-Incorporated SVM Prediction of *Ciona* Neuropeptide–Receptor Pairs. PDs encoding peptides and a TM *z*-scale descriptor encoding GPCRs (Fig. 1*B*) were utilized for the encoding of 1,352 CPIs (Dataset S1) and the same number of generated noninteraction pairs (*Materials and Methods*). The resulting CPIs and noninteraction pairs were in turn utilized as training sets for SVMs (Fig. 1*B*). The prediction performances of trained SVMs were evaluated by LOSO internal validation using the predicted CPIs and noninteraction pairs as test sets (21, 26, 28). Because the CPI datasets partitioned into respective subsets for peptide–GPCR interactions in humans, mice, other vertebrates, and invertebrates were predicted using models containing the other datasets in a LOSO analysis, species-wide prediction performance was evaluated by LOSO cross-validation. The LOSO analysis using the PD-incorporated SVM produced values for leave-humans-, mice-, vertebrates-, and invertebrates-out of 0.949 ± 0.003, 0.977 ± 0.001, 0.988 ± 0.001, and 0.592 ± 0.032 for the area under the receiver operating characteristic curve (AUC) and 0.884 ± 0.010, 0.937 ± 0.010, 0.971 ± 0.003, and 0.501 ± 0.101 for accuracy (ACC) (Fig. 3*A* and *SI Appendix*, Table S3).

To confirm the prediction performance of the present PDs, the peptide–receptor prediction performance using other descriptors—specifically, 5–0, 5–1, and 5–2 mismatch descriptors (30, 34), a class of string kernels that compare sequence strings representing *k*-mer subsequences—were also evaluated by LOSO. LOSO analysis using the 5–0 mismatch descriptors for leave-humans-, mice-, vertebrates-, and invertebrates-out yielded 0.800 ± 0.011, 0.875 ± 0.003, 0.921 ± 0.012, and 0.436 ± 0.006 for the AUC and 0.743 ± 0.031, 0.852 ± 0.005, 0.861 ± 0.037, and 0.443 ± 0.020 for ACC (*SI Appendix*, Fig. S4*A* and Table S3). LOSO analysis using the 5–1 mismatch descriptors for leave-humans-, mice-, vertebrates-, and invertebrates-out yielded 0.867 ± 0.011, 0.925 ± 0.004, 0.962 ± 0.003, and 0.473 ± 0.012 for the AUC and 0.820 ± 0.021, 0.890 ± 0.012, 0.924 ± 0.017, and 0.496 ± 0.022 for ACC (*SI Appendix*, Fig. S4*A* and Table S3). LOSO analysis using the 5–2 mismatch descriptors for leave-humans-, mice-, vertebrates-, and invertebrates-out yielded 0.792 ± 0.008, 0.848 ± 0.011, 0.898 ± 0.009, and 0.497 ± 0.010 for the AUC and 0.737 ± 0.031, 0.815 ± 0.018, 0.861 ± 0.024, and 0.493 ± 0.020 for ACC (*SI Appendix*, Fig. S4*A* and Table S3). These data indicate that the scores of our developed PDs were higher than those of 5–0, 5–1, and 5–2 mismatch descriptors,

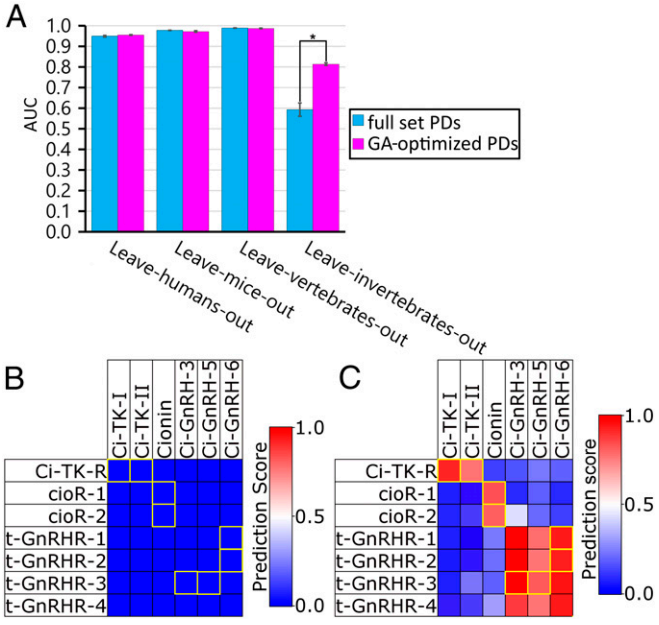

Fig. 3. The original PD-incorporated SVM showed sufficient prediction performance for neuropeptide–GPCR pairs of various species. (*A*) AUCs of model composed from original PDs and the resultant model from the second-round GAFS are shown with error bars representing the SEM of five repeated experiments with independently generated negative data. $*P < 0.05$. (*B* and *C*) Prediction results for *C. intestinalis* CPIs using (*B*) the model resulting from original PDs and (*C*) the model resulting from the second-round GAFS are shown as a heat map. The color gradient represents predicted values for individual peptide–GPCR interactions. The known pairs are outlined in yellow.

confirming high prediction performance of the developed PDs. Consequently, we employed our PDs for the following analysis. However, the prediction performance for leave-invertebrates-out was still lower (0.592 ± 0.032 for the AUC) than that for vertebrates (leave-humans-, mice-, and vertebrates-out). To improve the prediction performance, we optimized the PDs using two rounds of genetic algorithm-based feature selection (GAFS) (Fig. 1*B*; also see *SI Appendix*, *Supplemental Methods*).

After GAFS, the optimized PDs displayed leave-humans-, mice-, and vertebrates-out of 0.955 ± 0.002, 0.972 ± 0.004, and 0.986 ± 0.002 for the AUC and 0.889 ± 0.011, 0.926 ± 0.013, and 0.959 ± 0.006 for ACC. Furthermore, the AUCs and ACCs for leave-invertebrates-out improved to 0.813 ± 0.006 and 0.847 ± 0.076, respectively. These scores confirmed high prediction performance of the neuropeptide–GPCR pairs for any species by the PD-incorporated SVM. Subsequently, we examined the prediction accuracy of the PD-incorporated SVM, trained with all 1,352 CPIs (Dataset S1) and the noninteraction pairs, for eight known CPIs for *Ciona* peptide and their cognate receptor pairs (Dataset S1) that were not included in the LOSO analysis. As shown in Fig. 3*C*, Ci-TK-I, Ci-TK-II, and cionin (*Ciona* cholecystokinin homolog) were predicted to interact specifically with cognate receptors Ci-TK-R, CioR1, and CioR2, respectively, by machine learning (Fig. 3*C*). These outputs were completely consistent with the previous experimental evidence for their specific interactions (10, 12). Similarly, t-GnRH-3, t-GnRH-5, and t-GnRH-6 were predicted to interact with their cognate receptors with somewhat low ligand selectivity, as previously reported (11). Thus, the present PD-incorporated SVM was found to predict all eight known *Ciona* peptide–GPCR pairs with an accuracy of 80.95%. In contrast, no positive *Ciona* peptide–GPCR pairs were predicted with machine-learning models with 5–0, 5–1, and 5–2 mismatch descriptors (30, 34), which agrees with low leave-invertebrates-out validation (Fig. 3*B* and *SI Appendix*, Fig. S4 *B–E*). Collectively, the LOSO evaluation (Fig. 3*A*) and prediction accuracy for datasets of *Ciona* neuropeptide–receptor pairs (Fig. 3*C*) demonstrate that the PD-incorporated SVM model detects neuropeptide–GPCR pairs in both vertebrates and invertebrates. To the best of our knowledge, this is unique as a machine-learning model that can predict peptide-GPCR pairs of any animal species with high accuracy.

Using the PD-incorporated SVM trained with all 1,352 CPIs (rows 2–1,353 in Dataset S1) and the noninteraction pairs, we predicted the interactions between 19 *Ciona* neuropeptides (*SI Appendix*, Fig. S1 and Table S1) identified by our previous peptidomics study of the central nervous system (8) and 140 putative *Ciona* GPCRs (Dataset S2) extracted from the Ghost database (35) by GPCRalign (36). Each GPCR ID was abbreviated by omitting the splicing variant information (*SI Appendix*, Table S4). The prediction values for each pair ranged from 1 (absolute interaction) to 0 (absolute noninteraction). PD-incorporated SVM analysis of a total of 2,660 *Ciona* peptide–GPCR pairs [19 *Ciona* peptides (*SI Appendix*, Table S1) × 140 *Ciona* GPCRs (Dataset S2)] were subjected to PD-incorporated SVM prediction and a total of 13 putative peptide–GPCR pairs were produced with prediction scores higher than 0.7 for *Ciona* galanin-like peptide (Ci-GALP), Ci-NTLP-2, Ci-NTLP-3, Ci-LF-2, Ci-LF-3, Ci-LF-8, Ci-YFV-1, and Ci-YFV-3 (Fig. 4*A* and *SI Appendix*, Table S5).

Identification of 12 Neuropeptide–GPCR Pairs by Experimental Validation of the Predicted Pairs. We predicted and evaluated neuropeptide-GPCR pairs in two stages using a self-training strategy for semisupervised learning (37). For the first-stage evaluation, we experimentally assessed seven pairs (Ci-GALP-KH.C3.660; Ci-NTLP-2-KH.C9.683 and KH.C3.920; Ci-LF-2-KH.C2.127, KH.L172.28, and KH.C2.1132; and Ci-YFV-1-KH.C1.745) that had high prediction values (*SI Appendix*, Table S5) in the aforementioned model. Each promiscuous Gαq16-fused GPCR was transiently expressed in Sf9 cells, and intracellular Ca^{2+} mobilization was assessed in the presence of various concentrations of the peptide ligands. The cell-based signaling assay demonstrated that Ci-GALP, Ci-NTLP-2, Ci-LF-2, and Ci-YFV-1 induced Ca^{2+} mobilization in cells transfected with KH.C3.660 (Fig. 4*B*), KH.C9.683 (Fig. 4*C*), KH.C2.1132 (Fig. 4*D*), and KH.C1.745 (Fig. 4*E*), respectively, with nanomolar efficacy (Table 1). In contrast, dose-dependent responses were not observed with cells expressing other receptors. Furthermore, the PD-incorporated SVM was provided with data for the four experimentally validated *Ciona* GPCR–neuropeptide pairs as positive examples and three other pairs as negative examples for the second-stage validation, using a self-training strategy for semisupervised learning (37). The feature set for training and prediction was not changed from the PD-incorporated feature set used above, and the additional datasets were expected to update the discriminant functions (weight vectors) for the possible estimation of peptide–receptor interactions, leading to the prediction of more peptide–receptor pairs. As shown in Fig. 5*A*, the updated PD-incorporated SVM with additional training data output 22 putative peptide–GPCR pairs for Ci-NTLP-4, Ci-LF-1, Ci-LF-2, Ci-LF-5 to -8, Ci-YFV-1 to -3, and Ci-YFL-1 (*SI Appendix*, Table S6). Ca^{2+}-mobilization assays also verified specific (nanomolar efficacy) interactions of KH.C4.122 with Ci-LF-1 and Ci-LF-6 (Fig. 5 *B* and *C*); of KH.C2.1037 with Ci-LF-1, Ci-LF-5, and Ci-LF-6 (Fig. 5 *D–F*); of KH.C2.878 with Ci-LF-7 (Fig. 5*G*); of KH.C2.212 with Ci-LF-8 (Fig. 5*H*); and of KH.C8.781 with Ci-YFV-3 (Fig. 5*I* and Table 1). In contrast,

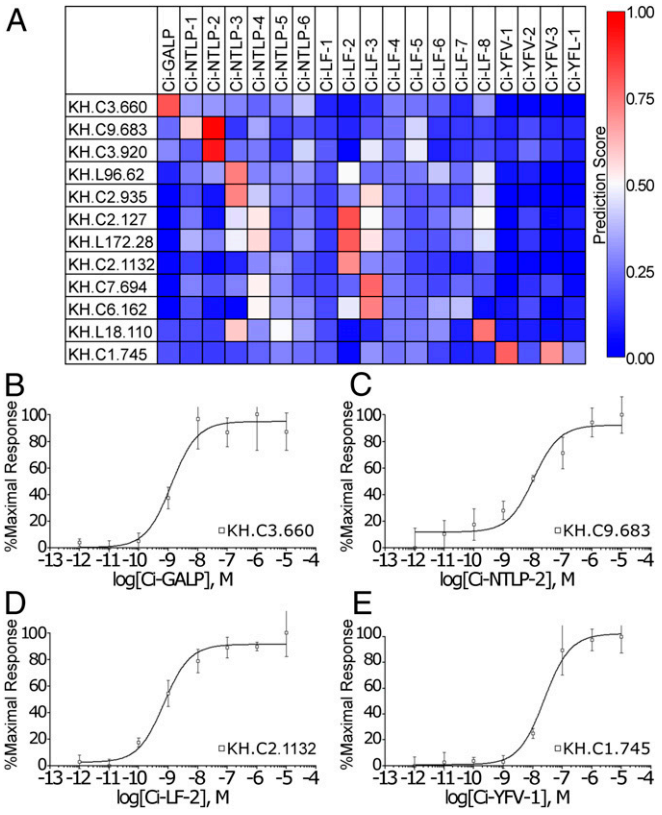

Fig. 4. GPCRs for Ci-GALP, Ci-LF-2, Ci-NTLP-2, and Ci-YFV-1 were identified by signaling assays based on the prediction by PD-incorporated SVM. (*A*) Prediction results for *C. intestinalis* CPIs are shown as a heat map. The color gradient represents the predicted value for each interaction between peptide and GPCR. Only GPCRs that were predicted to interact with at least one peptide with prediction scores higher than 0.7 are shown. Dose-dependent responses of (*B*) Ci-GALP, (*C*) Ci-NTLP-2, (*D*) Ci-LF-2, and (*E*) Ci-YFV-1 in Sf-9 cells expressing each receptor were assessed with intracellular Ca^{2+} mobilization, and sigmoid curves were calculated using Prism 3.03. Error bars represent the SEM of more than three experiments.

Table 1. Identified GPCR–peptide pairs

Ghostdatabase ID for receptor gene	Receptor gene name	Ligand	EC_{50} (nM)
KH.C3.660	*Ci-GALP-R*	Ci-GALP	1.29
KH.C9.683	*Ci-NTLP-2-R*	Ci-NTLP-2	11.05
KH.C4.122	*Ci-LF-1-R*	Ci-LF-1	5.25
KH.C4.122	*Ci-LF-1-R*	Ci-LF-6	223.87
KH.C2.1037	*Ci-LF-5/6-R*	Ci-LF-1	141.25
KH.C2.1037	*Ci-LF-5/6-R*	Ci-LF-5	4.78
KH.C2.1037	*Ci-LF-5/6-R*	Ci-LF-6	1.55
KH.C2.1132	*Ci-LF-2-R*	Ci-LF-2	0.71
KH.C2.878	*Ci-LF-7-R*	Ci-LF-7	2.04
KH.C2.212	*Ci-LF-8-R*	Ci-LF-8	1.35
KH.C1.745	*Ci-YFV-1-R*	Ci-YFV-1	24.55
KH.C8.781	*Ci-YFV-3-R*	Ci-YFV-3	1.98

all of the above neuropeptides show no Ca^{2+} mobilization at other GPCRs with prediction scores higher than 0.7. Altogether, these results provided evidence for the identification of a Ci-GALP receptor (Ci-GALP-R), Ci-NTLP-2 receptor (Ci-NTLP-2-R), Ci-LF-1 receptor (Ci-LF-1-R), Ci-LF-2 receptor (Ci-LF-2-R), Ci-LF-5/6 receptor (Ci-LF-5/6-R), Ci-LF-7 receptor (Ci-LF-7-R), Ci-LF-8 receptor (Ci-LF-8-R), Ci-YFV-1 receptor (Ci-YFV-1-R), and Ci-YFV-3 receptor (Ci-YFV-3-R) (Table 1). Although Ci-LF-1-R and Ci-LF-5/6-R were weakly activated by Ci-LF-6 and Ci-LF-1 (Table 1), respectively, Ci-LF-1-R exhibited a 42-fold selectivity for Ci-LF-1 relative to Ci-LF-6, while Ci-LF-5/6 exhibited a 91-fold selectivity for Ci-LF-6 relative to Ci-LF-1. Consequently, Ca^{2+}-mobilization assays for a total of 29 predicted pairs (7 pairs from the first-stage evaluation and 22 from the second-stage evaluation) resulted in a 41% hit rate (12 experimentally validated pairs).

Molecular Phylogenetic Tree Analysis of Identified *Ciona* Neuropeptide GPCRs. To evaluate the presence of known receptors closely related to the identified *Ciona* GPCRs, gene trees were estimated by collecting similar bilaterian sequences (Fig. 6). We used the Ci-GALP-R sequence as a query with the Basic Local Alignment Search Tool (BLAST) to demonstrate that similar sequences were detected in genome data representing all deuterostome lineages. Among them, Ci-GALP-R displayed 37–42% sequence identity (*SI Appendix*, Fig. S5*A*) to eight vertebrate galanin or GALP receptors (38) and 35–44% sequence identity (*SI Appendix*, Fig. S5*A*) to nine putative cephalochordate galanin or GALP receptors, indicating sequence identity of Ci-GALP-R to those of other galanin/GALP receptor family GPCRs. Molecular phylogenetic tree analysis demonstrated that urochordate GALP-Rs were positioned outside of either vertebrates or cephalochordate galanin/GALP receptors (*SI Appendix*, Fig. S5*A*), indicating that urochordate GALP-Rs evolved in unique ways. However, the deuterostome GALP-R clade including Ci-GALP-R was consistently supported by both neighbor-joining (NJ) and maximum-likelihood (ML) analysis (Fig. 6*A* and *SI Appendix*, Fig. S5 *A*, *1–3*), revealing that Ci-GALP-R shares a common ancestor with the vertebrate galanin receptor proteins.

A BLAST search using the Ci-NTLP-2-R sequence as a query identified similar deuterostome sequences, including eight vertebrate adhesion GPCRs (20–24% identity) (*SI Appendix*, Fig. S5*B*). However, among these BLAST hits, phylogenetic analyses did not identify any nonurochordate sequence similar to the Ci-NTLP-2-R sequence (Fig. 6*B* and *SI Appendix*, Fig. S5*B*). In addition, the sequence alignment showed that the N terminus of Ci-NTLP-2-R is shorter than that of other GPCRs (*SI Appendix*, Fig. S5 *B*, *4*). Some adhesion GPCRs are known as receptors for high molecular-weight protein ligands, such as collagen (adhesion GPCR G6; ENST00000394143.5) (39, 40) and neurexins (adhesion GPCR L1; ENST00000340736.10) (41). Notably, the amino acid length of the ligand of Ci-NTLP-2-R, Ci-NTLP-2 (8 aa, MMLGPGIL) (*SI Appendix*, Table S1), is far shorter than those of collagens (>1,000 aa) and FLRT3 (>600 aa). Given that a significant sequence identity was not found between Ci-NTLP-2 and these proteins, Ci-NTLP-2-R is considered to be a GPCR for a short neuropeptide but not an adhesion-related protein.

A BLAST search using the Ci-LF-1-R sequence as a query identified similar deuterostome sequences (*SI Appendix*, Fig. S5*C*). Ci-LF-Rs showed 21–26% sequence identity to 10 vertebrate GPCRs, including class A small molecular-weight transmitters GPCRs (cannabinoid receptors, adenosine receptors, and adrenergic receptors) (*SI Appendix*, Fig. S5*C*). Phylogenetic analyses indicated that all Ci-LF-Rs belong to the *Ciona*-specific clade (Fig. 6*C*) and this clade is deeply nested within the urochordate LF-R clade consisting of presumable LF-R sequences of *Botryllus schlosseri* and *Oikopleura dioica*. The urochordate LF-R clade, however, did not have any closely related sequence of nonurochordate deuterostomes (*SI Appendix*, Fig. S5*C*). This result suggests that, after a split of other tunicate lineages, the Ci-LF-Rs evolved within the *Ciona* lineage as paralogs via gene multiplication and are in good agreement with the finding that the Ci-LF-Rs share little sequence homology with any hitherto known GPCR for peptides.

A BLAST search using the Ci-YFV-1-R sequence as a query identified similar sequences from urochordates but not from other deuterostomes (Fig. 6*D*). Phylogenetic analyses (*SI Appendix*, Fig. S5*D*) demonstrated that Ci-YFV-Rs were grouped with sequences of probable YFV-Rs of *Ciona* and *B. schlosseri*. This result suggests that Ci-YFV-Rs were generated within the urochordate lineage. Combined with the experimental evidence for specific neuropeptide–GPCR pairs (Figs. 4 and 5), these molecular phylogenetic tree analyses suggest that Ci-NTLP-Rs, Ci-LF-Rs, and Ci-YFV-Rs are not closely related to any other known GPCRs.

Expression of Ci-GALP-R, Ci-NTLP-2-R, Ci-LF-R, and Ci-YFV-R Genes in Various Tissues. Real-time PCR revealed the expression patterns of the identified GPCRs. For example, Ci-LF-Rs, except Ci-LF-8-R, were shown to be expressed specifically in the oral and atrial siphons (Fig. 7), suggesting some biological roles of Ci-LF-1 to -7 in feeding behavior. Ci-GALP-R, Ci-YFV-1-R, and Ci-YFV-3-R were more highly expressed in the neural complex, compared with other identified GPCRs (Fig. 7). These results demonstrate the unique expression profile of these GPCRs and suggested that their peptide ligands produce diverse biological functions.

Discussion

Neuropeptides play multiple biological roles upon binding to their cognate receptors expressed in various tissues and cells. Thus, identification of neuropeptide–GPCR pairs, namely, deorphanization of GPCRs, is a crucial step in the elucidation of their endogenous roles. Moreover, both novel and homologous neuropeptides have been characterized in various organisms, highlighting the significance of neuropeptidergic signaling systems in molecular and functional evolution and diversification in the animal kingdom. However, elucidation of nonhomologous neuropeptide–receptor pairs remains a severe bottleneck in a wide range of biological sciences, because prediction and identification of the receptors for novel peptides is one of the most time-consuming and serendipity-dependent tasks in biology due to low sequence similarities and poor molecular phylogenetic correlations, even in human and model organisms. Although reverse-pharmacological strategies generally require multiyear trial-and-error testing to elucidate one ligand-receptor pair, the identification of receptors for novel ligands, including species-specific nonhomologous peptides, still depends on this strategy (42). A large-scale combinatorial reverse-pharmacological method identified 19 invertebrate neuropeptide GPCRs (16), but this strategy requires multistep experiments for numerous peptide–receptor pair candidates, and most of the identified peptide–receptors were homologs of other species. In this study, we efficiently and systematically identified multiple neuropeptide–GPCR pairs of the phylogenetically closest relative of vertebrates, *C. intestinalis* Type A, with the assistance of an original machine learning-based approach. Of particular significance

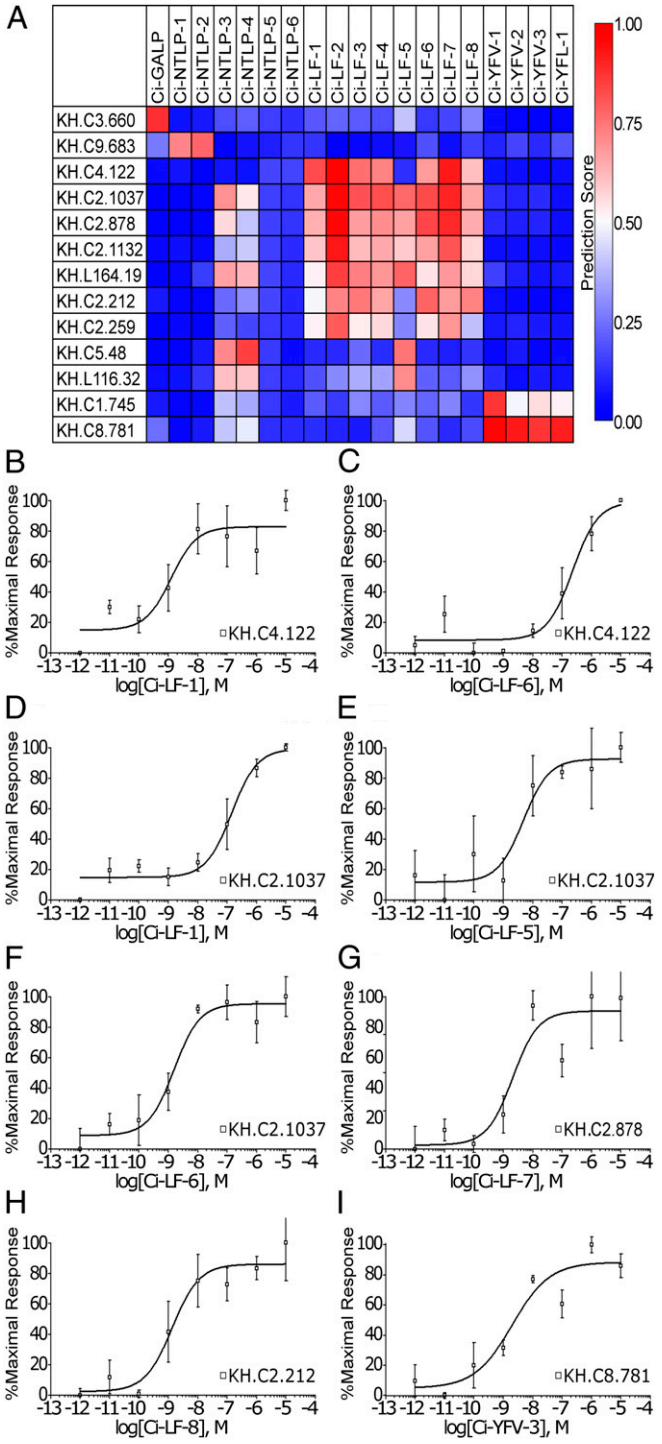

Fig. 5. Data feedback of experimentally validated results of four *Ciona* neuropeptide–GPCR pairs, leading to the identification of eight additional pairs. (*A*) Prediction results for *C. intestinalis* CPIs are shown as a heat map. The color gradient represents the predicted value for each interaction between peptide and GPCR. Only GPCRs that were predicted to interact with at least one peptide with prediction scores higher than 0.7 are shown. Dose-dependent responses of (*B* and *D*) Ci-LF-1, (*E*) Ci-LF-5, (*C* and *F*) Ci-LF-6, (*G*) Ci-LF-7, (*H*) Ci-LF-8, and (*I*) Ci-YFV-3 in Sf-9 cells expressing each receptor were assessed with intracellular Ca^{2+} mobilization, and sigmoid curves were calculated using Prism 3.03. Error bars represent the SEM of more than three experiments.

is that we succeeded in elucidating 1 homologous and 11 *Ciona*-specific neuropeptide–GPCR pairs during validation of 29 predicted peptide–receptor pairs. This represents a 41% hit rate using only 1,352 CPIs, namely, data for known endogenous peptide–GPCR pairs. Examination of these 29 predicted interactions and elucidation of 12 (11 *Ciona*-specific and 1 homologous) neuropeptide–GPCR pairs were completed within only 9 mo after the first-round prediction of *Ciona* neuropeptide–GPCR pairs (Fig. 4*A*). This is an obviously higher throughput than that of reverse-pharmacological strategies. Consequently, the present study illustrates the effectiveness of combining PD-incorporated SVM with cell-based experimental validation for the identification of neuropeptide–GPCR pairs.

Combined with previously identified homologous neuropeptide–GPCR pairs, this study led to the elucidation of a total of 26 neuropeptide–GPCR pairs in *Ciona*, which is comparable to those of conventional protostomian model organisms, such as *Drosophila* and *C. elegans* (5). Previously, only a few biological roles of *Ciona* neuropeptides had been elucidated: regulation of vitellogenic follicles by Ci-TK (9, 43) and metamorphosis by GnRH (9, 44). Thus, the present identification of multiple neuropeptide–GPCR pairs (Figs. 5 and 6) and localization of the GPCR gene expression surely facilitates the elucidation of neuropeptidergic molecular mechanisms (Fig. 7) and networks underlying various biological events regulated by the nervous, neuroendocrine, and endocrine systems in *Ciona*. Furthermore, because *Ciona* is the closest living relative of vertebrates, this study is also expected to contribute a great deal to the exploration of the common and species-specific evolution of the nervous, neuroendocrine, and endocrine systems throughout the Chordata phylum.

We verified that Ci-LF-1, -2, -5, -6, -7, and -8 and Ci-YFV-1 and -3 exhibited prominent selectivity to their receptors (Figs. 3 and 4), whereas receptors for Ci-LF-3 and -4, Ci-YFV-2, and Ci-YFL-1 have yet to be elucidated. This is mainly due to the failure of expression of the most probable receptor candidate proteins in expression systems, including mammalian cells, insect cells, and *Xenopus* oocytes, rather than implicit prediction of peptide–receptor systems. Replacement of the N-terminal regions of *Ciona* GPCRs with those of mammals or insects should result in functional expression, leading to the experimental validation of predicted peptide–GPCR pairs.

Recently, molecular phylogenetic approaches have provided some insight into evolutionary aspects and classification of invertebrate peptides, GPCRs, and peptide–GPCR pairs (5, 45). For example, integrative molecular phylogenetic analyses identified 29 categories of peptide and GPCR subfamilies based on position-specific scoring matrices of GPCRs and peptide precursors, followed by prediction of peptide–GPCR pairs (5, 45). However, these methods were limited to the prediction of known homologous peptide–GPCR pairs. Of particular significance is that Ci-NTLP-2-R, Ci-LF-Rs, and Ci-YFV-Rs constitute unique clades with orphan GPCRs or GPCRs for nonpeptide endogenous ligands, not with hitherto known GPCRs for peptides, indicating that these genes were generated in a species-specific lineage (Fig. 6 and *SI Appendix*, Fig. S5). The existence of such *Ciona*-specific evolutionarily unrelated neuropeptide GPCR genes is compatible with a rapid evolutionary rate of the *Ciona* genome and species-specific gene multiplication (46). In other words, the present molecular phylogenetic trees (Fig. 6 and *SI Appendix*, Fig. S5) strongly suggest that novel neuropeptide GPCRs also constitute unique clades with GPCRs for non-peptidic ligands in other species, including humans, supporting the view that methods based on sequence similarity or molecular phylogenetic relatedness have not been useful for predicting novel peptide–GPCR pairs. In contrast, unprecedented molecular mechanisms and evolutionary processes of peptide–GPCR interactions have a high likelihood of being recognized by the PD-incorporated SVM, suggesting that the present machine-learning approach will lead to the exploration of new

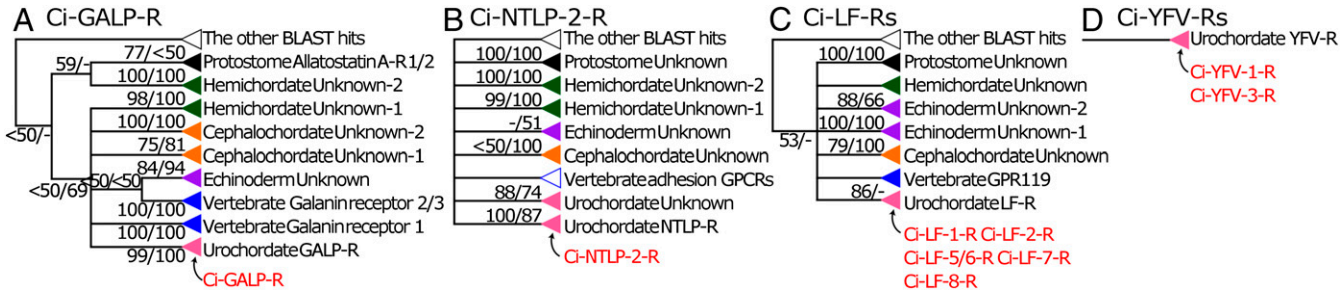

Fig. 6. Demonstration that Ci-NTLP-2-R, Ci-LF-Rs, and Ci-YFV-Rs are not evolutionarily related to any known neuropeptide GPCRs using gene trees of (*A*) Ci-GALP-R, (*B*) Ci-NTLP-2-R, (*C*) Ci-LF-Rs, and (*D*) Ci-YFV-Rs. Phylogenic trees of ligand-identified *Ciona* GPCRs were constructed using the ML and NJ methods (*SI Appendix*, Fig. S5 *A–D*, 612 sites, 450 sites, 492 sites, and 394 sites, respectively) and resultant topologies were confirmed by estimating ML trees based solely on TM domains. Each schematic of gene trees was constructed based on the three molecular phylogenetic trees using ORTHOSCOPE 1.0.1 (58). The monophyly-supported and -unsupported gene clades were indicated by closed triangles and open triangles, respectively. Clade names indicate inferred functions of ancestral genes based on clade members with experimental data. The number at each branch node represents the percentage given by 100× bootstrap replicates (ML/NJ). *Ciona* GPCRs characterized herein are shown in red.

phylogenetically unrelated GPCR repertoires in a wide range of species, including humans.

Machine-learning methods have provided predictive models or simulations of ligand–receptor interactions (24, 47–50). However, the experimental evidence for these has been limited to nonendogenous small compounds (26). Moreover, to the best of our knowledge, this prediction of peptide–receptor pairs using machine learning enabled by the development of original PDs is unique (Fig. 1 and *SI Appendix*, Table S2). Collectively, the present study shows identification of cognate endogenous peptide–receptor pairs using a sequential combination of machine learning and experimental validation. Additionally, the aforementioned hit rate of the PD-incorporated SVM (41%) was much higher than those for the elucidation of GPCRs for small nonendogenous compound prediction using in silico virtual screening, such as structure-based (20) and other chemical genomic models (26).

The LOSO validation, which enabled the evaluation of species-wide prediction performance, contributed to estimating the prediction performance of the species-specific CPIs. We also estimated the prediction performances using fivefold cross-validation (5-CV) (24, 26). As shown in *SI Appendix*, Fig. S6, 5-CV showed prediction performance of AUCs higher than 0.85 for all descriptors, including 5–0, 5–1, and 5–2 mismatch descriptors and PDs, whereas no known *Ciona* peptide–GPCR pairs were predicted (*SI Appendix*, Fig. S4 *B–D*). In contrast, despite low performance of the original SVM with any descriptors validated by leave-invertebrates-out analysis (AUC < 0.6), GAFS-optimized PD-incorporated SVM validated by leave-invertebrates-out analysis showed higher prediction performance (AUC of 0.813) and, indeed, output complete prediction of all known *Ciona* peptide–GPCR pairs (Fig. 3 *B* and *C*) and led to the elucidation of 12 *Ciona* peptide–GPCR pairs (Figs. 4 and 5). Collectively, these results proved that 5-CV overestimated the prediction performance compared with leave-invertebrates-out analysis. These gaps between validation scores and actual prediction accuracy are likely to result from the difference in distribution of orthologous GPCRs among species. As shown in Fig. 2, a total of 1,220 human, mouse, and vertebrate CPIs include numerous orthologous peptides and receptors with high SS from a single phylum (Vertebrata), whereas invertebrate CPIs include various species-specific peptides and GPCRs with low SS from a wide range of phyla (e.g., Nematoda, Arthropoda, and Mollusca) regardless of the small number (132 invertebrate CPIs). These features of CPIs are thought to cause the overestimation of the prediction performance by the 5-CV (*SI Appendix*, Fig. S6) and leave-humans-, mice-, and vertebrates-out methods (Fig. 3*A* and *SI Appendix*, Fig. S6).

Also of interest is that self-training using experimentally validated data (CPIs) (Fig. 4) facilitated the identification of additional

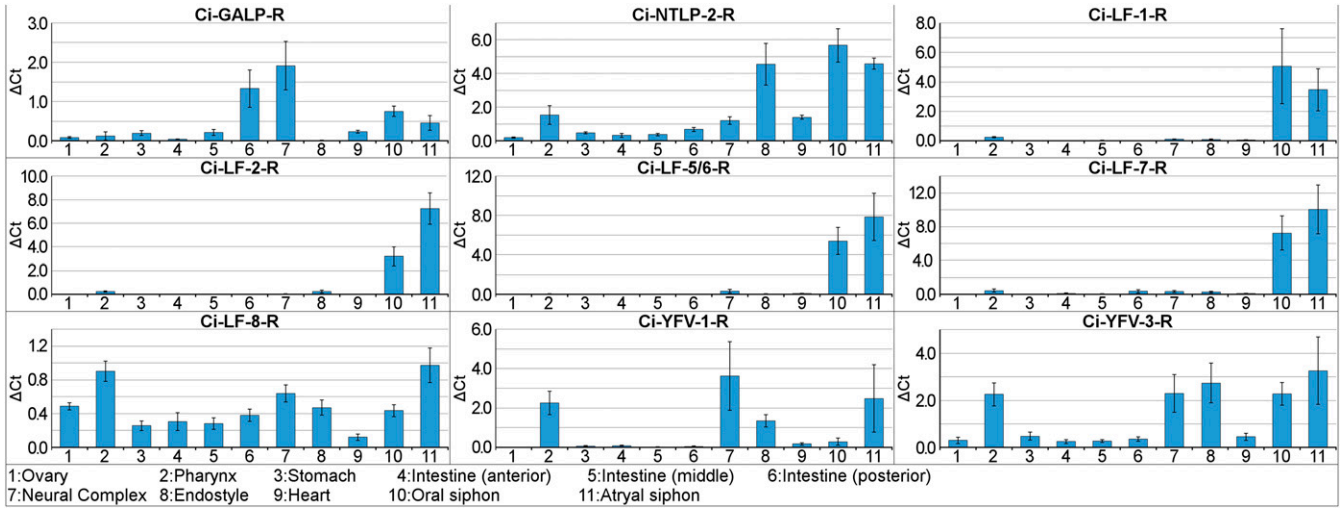

Fig. 7. Gene expression profiles for *Ciona* neuropeptide GPCRs. Relative expression of the *Ciona* neuropeptide receptors to *Ciona* GAPDH in the indicated tissues was confirmed by real-time PCR. Data are shown as the means of three independent experiments ± SE.

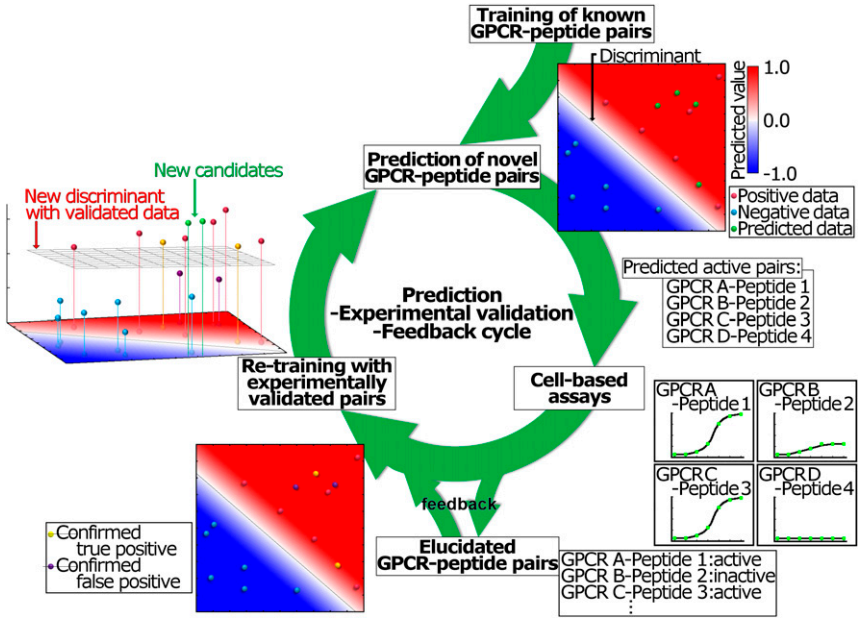

Fig. 8. PD-incorporated SVM cycle: the prediction-experimental validation-data feedback is a powerful procedure for deorphanization of GPCRs for novel neuropeptides. The original prediction model was constructed by learning positive data (red dots) and negative data (blue dots) for known neuropeptide–GPCR pairs, followed by cell-based signaling assays of each predicted pair; predicted GPCR–peptide pairs are green dots, positive matches are yellow dots, and false positives are purple dots. The feedback of experimentally validated neuropeptide–GPCR pairs updated the prediction model, which enabled the prediction of more positive GPCR–peptide pairs. This improved prediction model indicated that repeated prediction-experimental validation-feedback cycles make the PD-incorporated SVM more "intelligent" and improve the prediction performance.

peptide–GPCR pairs (Fig. 8), and some negative CPIs were also generated (Fig. 5). These results provide evidence that validated data feedback to the PD-incorporated SVM improves the prediction accuracy and then verifies an unprecedented mode of ligand–GPCR interaction; in brief, the SVM has become more "intelligent" by acquiring new knowledge. Novel GPCRs have also been found in other species using next-generation sequencer-based genome or transcriptome analyses (51, 52), whereas the cognate ligands of most of such GPCRs have yet to be identified. In this context, the present study indicates that our PD-incorporated SVM (Fig. 8) can identify numerous peptide–GPCR pairs in various organisms via self-training, leading to the elucidation of molecular mechanisms underlying peptide–GPCR recognition and net evolutionary processes of peptide–GPCR interactions. Overall, these findings highlight the current prediction ability of the PD-incorporated SVM using limited amounts of CPI data and indicate the potential for further prediction system development for novel human peptide–GPCR pairs, including artificial peptidic drug candidates.

In conclusion, we have efficiently and systematically elucidated multiple neuropeptide–GPCR pairs in a phylogenetically critical invertebrate chordate, *C. intestinalis* Type A, using a combination of machine learning and experimental validation. This study not only contributes to the investigation of molecular mechanisms for various nervous, neuroendocrine, and endocrine systems of *Ciona*, but also sheds light on the versatility of PD-incorporated SVM in the identification of multiple peptide–receptor pairs.

Materials and Methods

CPI Data. CPI pairs with peptide ligands were collected from the IUPHAR Database (53) and UniProtKB knowledge base (54). From these databases, we utilized 261, 183, 169, 1, 13, and 10 CPI pairs for humans, mice, rats, opossums, zebrafish, and chickens, respectively. The information about the GPCR and peptide sequences was obtained from the UniProtKB (54). Additionally, we collected data for noninteraction pairs and invertebrate peptide–GPCR interaction pairs from the literature. All of the collected interactions and references are listed in Dataset S1. The 531 human interactions (rows 2–532 in Dataset S1), 310 mouse interactions (rows 533–842 in Dataset S1), 379 vertebrate interactions (rows 843–1,241 in Dataset S1), and 132 invertebrate interactions (rows 1,242–1,353 in Dataset S1) were used for training datasets as positive pairs. To generate the same number of negative pairs, we collected the reported noninteraction pairs and generated the randomly selected negative pairs as previously reported (21, 29). A total of 3 reported noninteraction pairs (rows 1,354–1,356 in Dataset S1) and 528 randomly selected negative pairs for humans, 310 randomly selected negative pairs for mice, 7 reported noninteraction pairs (rows 1,357–1,363 in Dataset S1) and 372 randomly selected negative pairs for vertebrates, and 82 reported noninteraction pairs (rows 1,364–1,445 in Dataset S1) and 50 randomly selected negative pairs for invertebrates were used for training datasets.

Peptide Kernels. We constructed the PDs with regular expression-based high-resolution representations, which encode the existence or absence of regular expression-represented 5-aa motifs. The descriptors were calculated in three steps (Fig. 1*A*). First, we collected the 51 regular expression elements to match amino acids, which consist of 21-bit representations of PROFEAT (32), 3 repeats, N-terminus and C-terminus marks of peptide sequences, and 25 single residues (*SI Appendix*, Table S2). For example, the regular expression element of [KR] (13th element of *SI Appendix*, Table S2) matches a single residue of lysine or arginine. In the second step, all of the permutations and combinations of 5 of these 51 regular expression elements were generated. For example, pHW[GASDT]Y matches the peptide sequences possessing pyroglutamic acid, followed by histidine; followed by glycine, alanine, threonine, aspartic acid, or serine; and followed by tyrosine. The expression ^N.Y{1,5} matches the peptide sequences possessing asparagine at the N terminus, followed by any amino acid, and followed by one- to five-length repetitions of tyrosine. Third, the peptide sequences were encoded with bit (0, 1) vectors, which represent each regular expression match (= 1) or nonmatch (= 0). Then, to unify redundant regular expressions, if there was a pair of regular expressions appearing in the same compound set, the regular expression showing the narrower range was removed. The inner products of these bit vector pairs were calculated as the kernels for each peptide pair.

We also calculated the mismatch descriptor to compare with our proposed regular expression-based descriptors, which is a class of string kernels that compares sequence strings representing *k*-mer subsequences. The mismatch kernel allows for mutations between the subsequences. Specifically, the mismatch kernel is calculated based on shared occurrences of (*k*-*m*)-patterns in the data, where the (*k*-*m*)-patterns consist of all *k*-length subsequences that differ from a fixed *k*-length sequence pattern by at most *m* mismatches.

The inner products of bit vector pairs were calculated as the mismatch kernels for each peptide pair.

GPCR Kernels. TM *z*-scale descriptors were employed for representations of GPCRs, as previously described (21). Briefly, seven TM sequences were directly substituted with the *z*-scale vectors that represent five leading principal components obtained from 26 measured and computed physicochemical properties of amino acids. The 935-dimensional descriptors were generated by concatenating 5D vectors (z1–z5) for each of the 187 residues of TMs in GPCRs. The inner products of GPCR descriptor pairs were calculated as GPCR kernels for each GPCR pair.

Similarity Scores. The SSs of the GPCRs and peptides were defined as Tanimoto coefficients (33) of their top 1% most similar GPCRs and peptides, respectively, as described in our previous study (21). For calculation of Tanimoto coefficients, we utilized TM *z*-scale descriptors and regular expression-based descriptors for GPCRs and peptides, respectively.

CPI Pair Kernels and SVM Prediction. We utilized kernel methods to incorporate CPI data into SVMs (55) for constructing prediction models, as previously described (21). Here, kernels for CPI pairs were represented as the products of linear kernels for PDs and GPCR descriptors. Parameters of the SVM regularization were optimized using a grid search. All of the training and test CPIs are included in Dataset S1.

Performance Evaluation. The prediction performance of our proposed model was evaluated using LOSO internal validation, as previously reported (21, 26, 28). In the present LOSO validation, the CPIs and noninteraction pairs partitioned into human data, mouse data, vertebrate data, and invertebrate data were predicted using models containing the other CPIs and noninteraction pairs. For example, for the leave-humans-out validation, mouse, vertebrate, and invertebrate CPIs (rows 533–1,353 in Dataset S1) and noninteraction pairs were used for SVM training, and prediction performances (ACC and AUC) were calculated using the prediction results for human CPIs and noninteraction pairs. The performance of the internal validation was measured by ACC = (TP+TN)/(TP+TN+FP+FN), where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. To further confirm the prediction performance of CPIs using LOSO analysis, we also measured the performance of internal validation using the AUC (56), which is an index independent of the decision threshold of the prediction model and class probability distributions of predicted data. SEMs of ACCs and AUCs were estimated by five repeated experiments with independently generated negative data. Differences between AUCs and ACCs were evaluated using a Student's *t* test as appropriate, with $P < 0.05$ considered as significant.

Peptide Synthesis. The peptide sequences we utilized are listed in *SI Appendix*, Table S1. All peptides were synthesized using an ABI 430A solid-phase peptide synthesizer (Applied Biosystems) and the Fast Moc method, according to the manufacturer's instruction.

Gαq16-Fused *C. intestinalis* GPCRs. Each GPCR ID in the Ghost database (35) was indicated by abbreviated IDs without splicing variant information. The full-length IDs are listed in *SI Appendix*, Table S4. *C. intestinalis* putative full-length GPCRs—KH.C3.660, KH.C9.683, KH.C4.122, KH.C2.1132, KH.C2.1037, KH.C2.878, KH.C2.212, KH.C1.745, and KH.C8.781—were cloned from the central nervous system and were C-terminally fused with human Gαq16 protein, which was coupled with GPCRs and triggered intracellular calcium mobilization upon binding of a specific ligand (57). The human Gαq16 ORF clone (OriGene) was amplified (*SI Appendix*, Table S7) and ligated into the XbaI site of a pFastbacI plasmid (Invitrogen). Then, KH.C3.660, KH.C9.683, KH.C4.122, KH.C2.1132, KH.C2.1037, KH.C2.878, KH.C2.212, KH.C1.745, and KH.C8.781 were cloned into the NotI/XbaI site of the Gαq16-ligated pFastbacI plasmids, respectively. Transformation of competent cells with the *Ciona* GPCR-Gαq16-pFastbacI plasmid and the resulting bacmid isolation was performed according to the manufacturer's instructions for the Bac-to-Bac system (Thermo Fisher Scientific).

Calcium Accumulation Assay. Sf9 cells (Thermo Fisher Scientific) were grown in Sf900 II (Thermo Fisher Scientific) containing 10% FBS (Sigma) at 28 °C. *Ciona* GPCR-Gαq16-recombinant baculoviruses were generated in Sf9 cells transfected with the above bacmids using Cellfectin II, titrated, isolated, and transiently transfected into Sf9 cells using the Bac-to-Bac system according to the manufacturer's instruction (Thermo Fisher Scientific). Forty-eight hours after transfection, Sf9 cells were loaded for 30 min with 2.5 μM of Fluo-8 AM (AAT Bioquest) diluted in loading buffer [HBSS supplemented with 1.25 mM of probenecid and 0.04% (wt/vol) of pluronic F-127]. Each *Ciona* GPCR-fused human Gαq16 expression at cell membrane was confirmed by immunostaining using the anti-Gαq16 antibody (Ori Gene TA318890). Various concentrations of peptides were administrated to Sf9 cells in a FlexStation II-automated apparatus (Molecular Devices). Real-time fluorescent kinetics for Fluo-8 were observed at excitation/emission wavelengths of 490/514 nm. The calcium accumulation data were analyzed using Prism v6 (GraphPad) to fit to a sigmoidal concentration-response curve, and the means ± SEMs of EC_{50} were calculated.

Real-Time PCR. Total RNA (2 μg) extracted from various tissues of *Ciona* was reverse-transcribed using SuperScript III (Invitrogen) and oligo (dT) 20 primer. Real-time PCR was performed using the CFX96 Real-time System and SsoAdvanced Universal SYBR Green Supermix (Bio-Rad Laboratories). Total volume of reaction mixtures was 20 μL, consisting of 100-ng template cDNA, each 500-nM primer, and 10 μL SYBR Green Master Mix solution. PCR was performed for initial steps at 95 °C for 30 s, followed by 44 cycles at 95 °C for 15 s and at 60 °C for 30 min. A melting-curve analysis was performed to confirm the absence of primer dimers. Ct values for GAPDH and identified GPCR genes were calculated according to the manufacturer's instruction. The mean ± SEM of GAPDH-normalized ΔCt values were estimated from three replicates. Sequences of the primers used for the real-time PCR are listed in *SI Appendix*, Table S8.

Molecular Phylogenetic Analysis for *Ciona* GPCRs. GPCR sequences similar to the identified *Ciona* GPCR sequences were extracted by ORTHOSCOPE 1.0.1 (58). To implement a BLAST search in ORTHOSCOPE, coding sequences from Ci-GALP-R, Ci-NTP-2-R, Ci-LF-1-R, and Ci-YFL-1-R were used as queries against gene models of vertebrates (*Homo sapiens* and *Gallus gallus*), urochordates (*C. intestinalis*, *Ciona savignyi*, *B. schlosseri,* and *O. dioica*), cephalochordates (*Branchiostoma floridae* and *Branchiostoma belcheri*), echinoderms (*Acanthaster planci* and *Strongylocentrotus purpuratus*), hemichordates (*Ptychodera flava* and *Saccoglossus kowalevskii*), and protostomes (*D. melanogaster*, *C. elegans*, and *Lingula anatina*). The BLAST hit sequences were screened using an E-value cut-off of $<10^{-3}$, and the top five hits were used for the subsequent phylogenetic analyses. The protein sequences retrieved by the ORTHOSCOPE analyses were aligned using MAFFT (59). Multiple sequence alignments were trimmed by removing poorly aligned regions using TRIMAL 1.2 (60) with the option "gappyout." Corresponding coding sequences were forced onto the amino acid alignment using PAL2NAL (61) to generate nucleotide alignments for following analyses.

Gene phylogenetic trees were estimated using ML and NJ methods with the first and second codon positions and bootstrap analyses of genes encoding full-length sequences (for NJ and ML analyses) and TM domains (for ML analysis) of GPCRs based upon 100 replicates. Codon-partitioned ML analyses were performed with RAxML 8.2.12 (62), which invokes a rapid bootstrap analysis and searches for the best-scoring ML tree with the general time-reversible with gamma (GTRGAMMA) (63, 64) model. NJ analyses were conducted using the software package Ape in R using the TN93 model (65) with γ-distributed rate heterogeneity (64). The sequences for ligand-identified GPCRs in this paper are presented in Dataset S3. The molecular phylogenetic trees of full-length sequences (for NJ and ML analyses) and TM domains (for ML analysis) of GPCRs were constructed using the MEGA software (v7) (66). Each schematic of gene trees was constructed by focusing on gene clades consistently supported by the three molecular phylogenetic trees (*SI Appendix*, Fig. S5) using the ORTHOSCOPE, as previously reported (58).

ACKNOWLEDGMENTS. We thank Prof. Shigetada Nakanishi for his fruitful comments on the manuscript. *Ciona intestinalis* was raised and supplied by the National Bio-resource Project of *Ciona* (MEXT, Japan). This work was supported in part by the Japan Society for the Promotion of Science Grant 16K07430 (to H.S.).

1. Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439:965–968.
2. Denoeud F, et al. (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* 330:1381–1385.
3. Satoh N, Rokhsar D, Nishikawa T (2014) Chordate evolution and the three-phylum system. *Proc Biol Sci* 281:20141729.
4. Satoh N, Levine M (2005) Surfing with the tunicates into the post-genome era. *Genes Dev* 19:2407–2411.

5. Mirabeau O, Joly JS (2013) Molecular evolution of peptidergic signaling systems in bilaterians. *Proc Natl Acad Sci USA* 110:E2028–E2037.
6. Hewes RS, Taghert PH (2001) Neuropeptides and neuropeptide receptors in the *Drosophila melanogaster* genome. *Genome Res* 11:1126–1142.
7. Satake H, Kawada T (2006) Neuropeptides, hormones, and their receptors in ascidians: Emerging model animals. *Invertebrate Neuropeptides and Hormones: Basic Knowledge and Recent Advances*, ed Satake H (Transworld Research Network, Kerala, India), pp 253–276.
8. Kawada T, et al. (2011) Peptidomic analysis of the central nervous system of the protochordate, *Ciona intestinalis*: Homologs and prototypes of vertebrate peptides and novel peptides. *Endocrinology* 152:2416–2427.
9. Matsubara S, et al. (2016) The significance of *Ciona intestinalis* as a stem organism in integrative studies of functional evolution of the chordate endocrine, neuroendocrine, and nervous systems. *Gen Comp Endocrinol* 227:101–108.
10. Satake H, et al. (2004) Tachykinin and tachykinin receptor of an ascidian, *Ciona intestinalis*: Evolutionary origin of the vertebrate tachykinin family. *J Biol Chem* 279:53798–53805.
11. Tello JA, Rivier JE, Sherwood NM (2005) Tunicate gonadotropin-releasing hormone (GnRH) peptides selectively activate *Ciona intestinalis* GnRH receptors and the green monkey type II GnRH receptor. *Endocrinology* 146:4061–4073.
12. Sekiguchi T, Ogasawara M, Satake H (2012) Molecular and functional characterization of cionin receptors in the ascidian, *Ciona intestinalis*: The evolutionary origin of the vertebrate cholecystokinin/gastrin family. *J Endocrinol* 213:99–106.
13. Kawada T, Sekiguchi T, Itoh Y, Ogasawara M, Satake H (2008) Characterization of a novel vasopressin/oxytocin superfamily peptide and its receptor from an ascidian, *Ciona intestinalis*. *Peptides* 29:1672–1678.
14. Kamesh N, Aradhyam GK, Manoj N (2008) The repertoire of G protein-coupled receptors in the sea squirt *Ciona intestinalis*. *BMC Evol Biol* 8:129.
15. Hauser F, Cazzamali G, Williamson M, Blenau W, Grimmelikhuijzen CJ (2006) A review of neurohormone GPCRs present in the fruitfly *Drosophila melanogaster* and the honey bee *Apis mellifera*. *Prog Neurobiol* 80:1–19.
16. Bauknecht P, Jékely G (2015) Large-scale combinatorial deorphanization of platynereis neuropeptide GPCRs. *Cell Rep* 12:684–693.
17. Reynolds KA, Katritch V, Abagyan R (2009) Identifying conformational changes of the beta(2) adrenoceptor that enable accurate prediction of ligand/receptor interactions and screening for GPCR modulators. *J Comput Aided Mol Des* 23:273–288.
18. Kobilka BK, Deupi X (2007) Conformational complexity of G-protein-coupled receptors. *Trends Pharmacol Sci* 28:397–406.
19. Schwartz TW, Frimurer TM, Holst B, Rosenkilde MM, Elling CE (2006) Molecular mechanism of 7TM receptor activation—A global toggle switch model. *Annu Rev Pharmacol Toxicol* 46:481–519.
20. Huang XP, et al. (2015) Allosteric ligands for the pharmacologically dark receptors GPR68 and GPR65. *Nature* 527:477–483.
21. Shiraishi A, Niijima S, Brown JB, Nakatsui M, Okuno Y (2013) Chemical genomics approach for GPCR-ligand interaction prediction and extraction of ligand binding determinants. *J Chem Inf Model* 53:1253–1262.
22. Klabunde T (2007) Chemogenomic approaches to drug discovery: Similar receptors bind similar ligands. *Br J Pharmacol* 152:5–7.
23. Weill N, Rognan D (2009) Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J Chem Inf Model* 49:1049–1062.
24. Hamanaka M, et al. (2017) CGBVS-DNN: Prediction of compound-protein interactions based on deep learning. *Mol Inform,* 36.
25. Jacob L, Vert JP (2008) Protein-ligand interaction prediction: An improved chemogenomics approach. *Bioinformatics* 24:2149–2156.
26. Yabuuchi H, et al. (2011) Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol Syst Biol* 7:472.
27. Niijima S, Yabuuchi H, Okuno Y (2011) Cross-target view to feature selection: Identification of molecular interaction features in ligand-target space. *J Chem Inf Model* 51:15–24.
28. Mauri A, Consonni V, Pavan M, Todeschini R (2006) Dragon software: An easy approach to molecular descriptor calculations. *Match (Mulh)* 56:237–248.
29. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50: 742–754.
30. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics* 20:467–476.
31. Saigo H, Vert JP, Akutsu T (2006) Optimizing amino acid substitution matrices with a local alignment kernel. *BMC Bioinformatics* 7:246.
32. Rao HB, Zhu F, Yang GB, Li ZR, Chen YZ (2011) Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* 39:W385–W390.
33. Martin EJ, et al. (1995) Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J Med Chem* 38:1431–1436.
34. Leslie C, Kuang R (2003) Fast kernels for inexact string matching. *Porceedings of the 16th Annual Conference on Learning Theory and Kernel Workshop*, eds Schölkopf B, Warmuth M (Springer, Heidelberg, Germany), pp 114–128.
35. Satou Y, Satoh N (2005) Cataloging transcription factor and major signaling molecule genes for functional genomic studies in *Ciona intestinalis*. *Dev Genes Evol* 215:580–596.
36. Bissantz C, Logean A, Rognan D (2004) High-throughput modeling of human G-protein coupled receptors: Amino acid sequence alignment, three-dimensional model building, and receptor library screening. *J Chem Inf Comput Sci* 44:1162–1176.
37. Triguero I, García S, Herrera F (2015) Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Knowl Inf Syst* 42:245–284.
38. Kim DK, et al. (2014) Coevolution of the spexin/galanin/kisspeptin family: Spexin activates galanin receptor type II and III. *Endocrinology* 155:1864–1873.
39. Luo R, Jin Z, Deng Y, Strokes N, Piao X (2012) Disease-associated mutations prevent GPR56-collagen III interaction. *PLoS One* 7:e29818.
40. Paavola KJ, Sidik H, Zuchero JB, Eckart M, Talbot WS (2014) Type IV collagen is an activating ligand for the adhesion G protein-coupled receptor GPR126. *Sci Signal* 7:ra76.
41. Boucard AA, Ko J, Südhof TC (2012) High affinity neurexin binding to cell adhesion G-protein-coupled receptor CIRL1/latrophilin-1 produces an intercellular adhesion complex. *J Biol Chem* 287:9399–9413.
42. Laschet C, Dupuis N, Hanson J (2018) The G protein-coupled receptors deorphanization landscape. *Biochem Pharmacol* 153:62–74.
43. Aoyama M, et al. (2008) A novel biological role of tachykinins as an up-regulator of oocyte growth: Identification of an evolutionary origin of tachykininergic functions in the ovary of the ascidian, *Ciona intestinalis*. *Endocrinology* 149:4346–4356.
44. Kamiya C, et al. (2014) Nonreproductive role of gonadotropin-releasing hormone in the control of ascidian metamorphosis. *Dev Dyn* 243:1524–1535.
45. Jékely G (2013) Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proc Natl Acad Sci USA* 110:8702–8707.
46. Satoh N (2016) *Chordate Origins and Evolution: The Molecular Evolutionary Road to Vertebrates* (Elsevier, Boston).
47. Gawehn E, Hiss JA, Schneider G (2016) Deep learning in drug discovery. *Mol Inform* 35:3–14.
48. Yuriev E, Holien J, Ramsland PA (2015) Improvements, trends, and new ideas in molecular docking: 2012-2013 in review. *J Mol Recognit* 28:581–604.
49. König C, Alquézar R, Vellido A, Giraldo J (2018) Systematic analysis of primary sequence domain segments for the discrimination between class C GPCR subtypes. *Interdiscip Sci* 10:43–52.
50. Sagawa T, et al. (2018) Logistic regression of ligands of chemotaxis receptors offers clues about their recognition by bacteria. *Front Bioeng Biotechnol* 5:88.
51. Li C, et al. (2013) Comparative genomic analysis and evolution of family-B G protein-coupled receptors from six model insect species. *Gene* 519:1–12.
52. Chen N, et al. (2005) Identification of a nematode chemosensory gene family. *Proc Natl Acad Sci USA* 102:146–151.
53. Southan C, et al.; NC-IUPHAR (2016) The IUPHAR/BPS guide to PHARMACOLOGY in 2016: Towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res* 44:D1054–D1068.
54. UniProt Consortium (2015) UniProt: A hub for protein information. *Nucleic Acids Res* 43:D204–D212.
55. Vapnik VN (1998) *Statistical Learning Theory* (John Wiley & Sons, New York).
56. Ling CX, Huang J, Zhang H (2003) AUC: A statistically consistent and more discriminating measure than accuracy. *Proc IJCAI* 3:519–524.
57. Tabata K, Baba K, Shiraishi A, Ito M, Fujita N (2007) The orphan GPCR GPR87 was deorphanized and shown to be a lysophosphatidic acid receptor. *Biochem Biophys Res Commun* 363:861–866.
58. Inoue J, Satoh N (2019) ORTHOSCOPE: An automatic web tool for phylogenetically inferring bilaterian orthogroups with user-selected taxa. *Mol Biol Evol* 36:621–631.
59. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518.
60. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.
61. Suyama M, Torrents D, Bork P (2006) PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34: W609–W612.
62. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
63. Yang Z (1994) Estimating the pattern of nucleotide substitution. *J Mol Evol* 39: 105–111.
64. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol* 39:306–314.
65. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10:512–526.
66. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874.