JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

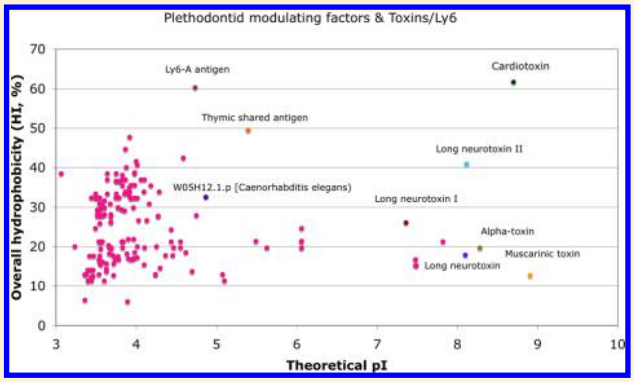
Multidimensional Drift of Sequence Attributes and Functional Profiles in the Superfamily of the Three-Finger Proteins and Their Structural Homologues

Andrzej Galat*

Commissariat à l'Energie Atomique, Direction des Sciences du Vivant, Institut de Biologie et de Technologies de Saclay, Service d'Ingénierie Moléculaire des Protéines, F-91191 Gif sur Yvette, France

Supporting Information

ABSTRACT: Functional diversity of the three-finger-protein domain (TFPD) had been acquired via hypervariability of some sequence positions and extensive insertion/deletion of short AA-segments that caused multidimensional drift of several sequence attributes such as the overall (HI) and local hydrophobicity levels, the isoelectric point (pI), distribution of charges, and local polarizability potentials. In consequence, the pIs of various TFPDs vary from 3 to 10, and the HIs span from highly hydrophilic to extreme hydrophobic levels. Our analyses of diverse genomic databases suggest that a primordial TFP-like fold could have been adapted to extracellular N-terminal domain (ECD) of the TGFβ series of receptors that are crucial for morphogenesis of multicellular organisms with elaborated body plans. It seems plausible that from the exons coding for the primordial ECD had radiated some ORFs that gave rise to small monodomain and multidomain TFPs that may be either membrane-bound or soluble entities and whose repertoire expanded in the genomes of vertebrates. We show that the above-mentioned attributes have been restrained within narrow ranges in several functionally related groups of TFPs that are expressed in miscellaneous organisms.



INTRODUCTION

Overall geometrical features of the Three-Finger-Protein (TFP) were established from the X-ray structure of neurotoxins isolated from sea snake venoms.[1,2] Subsequent X-ray studies revealed that the sequence of the N-terminal extra-cellular domain (ectodomain, ECD) of transforming growth factor-β (TGFβ) receptors despite its low similarity with various sequences of snake neurotoxins displays significant structural homology with the latter.[3] The ECDs of the TGFβ family of receptors bind different combinations of growth factors having cystein knot fold, namely bone morphogenetic proteins (BMPs), activins, inhibins, and TGFβ factors.[4−6] Binding of those diverse combinations of growth factors to the ECDs of the TGFβ family of receptors, which are linked via a trans-membrane (TM) segment with a serine kinase domain, known as the Activin receptor-Like Kinase (ALK), activates some intracellular signalization pathways.[5,7] These proteins together with their various extracellular ligands are vital in the embryogenesis and development of multicellular organisms with elaborated body plan.[8] The BMP and Activin Membrane-Bound Inhibitor pseudoreceptor (BAMBI) lacking ALK domain has the capacity to limit TGFβ signaling.[8] BAMBI is encoded in the genomes of diverse vertebrates and some insects. A survey of the X-ray structures and sequences of various TFPDs including the ECDs of the TGFβ family of receptors has revealed that even if the domain may contain from four to six disulfide bridges, only three of them are fully conserved and are essential for the stability of the fold.[9,10] Thus, even if the average overall sequence similarity in the group of diverse TFPDs is low ($ID_{ave} \leq 15\%$), the domain retains the same fold.[10] It is largely due to the peculiar arrangement of the three highly conserved disulfide bridges, which form a tightly interacting network at the palm of the TFPD, and which keep the fingers roughly pointing at the same direction. Majority of AA substitutions and different insertions and deletions (indels) events took place within the sequences of the fingers. Hypervariability of some sequence positions and diverse indel events had created a multitude of fine spatial features, which in consequence gave rise to a rich spectrum of binding specificities and biological functions.[10]

Small monodomain proteins having TFPD fold are expressed either as soluble forms, such as PATE-B, PATE-P,[11] or Lynx1C isoform which are modulators of some receptors,[12] secreted Ly-6/uPAR-related protein 1 (SLURP1),[10] or as entities that are anchored to the extracellular membrane via different linkers such as CD59 which inhibits complement membrane attack complex[13] and diverse members of the Ly-6 group of

Received: May 27, 2015
Published: September 2, 2015

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

proteins.[14] Proteins consisting either two or three TFPDs are expressed in vertebrates, namely C4.4A protein and urokinase-plasminogen associated receptor (uPAR), respectively.[15]

Analyses of the recently sequenced genomes of several marine organisms with elaborated body plans, whose ancestors probably appeared about 600 to 700 millions years ago, have revealed that they contain coding segments for the primordial forms of the TGFβ family of receptors and their ligands.[16−18] In this communication we have made thorough analyses of several groups of proteins containing the TFPD fold. Our analyses show that some crucial sequence attributes of the ECDs and ALK domains of the TGFβ family of receptors encoded in diverse marine organisms and those encoded in the human genome retain a high conservation level and display coherent functional profiles across the metazoan species. Moreover, besides the TFGβ series of receptors that are present at the genomes of primordial forms of sponges and other primitive multicellular organisms, the genomes of some marine invertebrate organisms contain few ORFs encoding small-size three-finger proteins (TFPs) whose repertoire had considerably expanded in the vertebrates. Analyses of diverse genomic databases suggest that the sequences of the exons coding for TFPDs underwent a considerable genetic drift, which gave rise to several groups of genes coding for various TFPs with distinct functional features. Our analyses have shown that the sequence attributes of some groups of small TFPs are contained within relatively narrow ranges. For example, the plethodontid modulating factors (PMF, also known as courtship pheromones in salamanders),[19] which are negatively charged proteins, have a fold that is similar to the TFPD fold,[9] whereas toxic components of snake venoms (neurotoxins and cardiotoxins) in the great majority are basic proteins.[9] We also discussed how the conservation of some sequence attributes in the TFPDs of diverse groups of TFPs is crucial for their functional features.

1. METHODS

1.1. Databases. In our analyses we used several genomic sequence databases of miscellaneous marine organisms, insects, and some vertebrates and the nonredundant protein sequence database (NrPSD) assembled in the NCBI (http://ncbi.nlm.nih.gov).[20] The following genomic databases were analyzed: the placozoan *Trichoplax adhaerens*,[21] the sponge native to the Great Barrier Reef *Amphimedon queenslandica*,[22] the fresh water polyp *Hydra magnipapillata*,[23] the starlet sea anemone *Nematostella vectensis*,[24] the sea squirt *Ciona intestinalis*,[25] the Florida lancelet *Branchiostoma floridae*,[26] the sea urchin *Strongylocentrotus purpuratus*,[27] the hemichordate acorn warm *Saccoglossus kowalevskii*,[28] the honey bee *Apis mellifera*,[29] the fly *Drosophila melanogaster*,[20,32] and the *Homo sapiens* genome.[20,30] The genomic databases were scanned with the BLAST program[31] accessible via the NCBI server and via FlyBase site (http://flybase.org/blast/)[32] using as templates the entire sequences of the TGFβ family of receptors expressed in *H. sapiens* and *D. melanogaster*, respectively.

1.2. Transformation of Protein Sequence Genomic Databases into Metadata. The Data_Base program was used to transform genomic databases downloaded from the NCBI server in the GenPept flat format into input files to the Seq_Pro program, which was used for formatting sequence streams to ClustalW and calculation of various sequence attributes. Multiple sequence alignments (MSAs) were made with the ClustalW1.83 program,[33] whereas dendrograms were created with the use of Dendroscope.[34] Hydrophobicity and amino acid (AA) bulkiness profiles, distribution of charges, the isoelectric points (pIs), and variability of the above sequence attributes in the MSAs[35] of analyzed groups of proteins were calculated with an upgraded version of the Hela_MSA program.[36] The program uses MSA created by ClustalW as input from which it generates several files with explicitly indicated blocks of hydrophobic segments (Figures S1−S2 in the Supporting Information). Some blocks of the hydrophobic sequences were used as inputs to an upgraded version of the Pars_DB program[36] that explored the content of the Hydrophobic Sequence Space (HSS) of proteins encoded in the genomes of various species (Figure S3 in the Supporting Information). The program performed comparisons of the input block consisting of sequences of TGFβ series of receptors and their physical-chemical attributes with the attributes of the hydrophobic segments that were detected in proteins of the above-mentioned genomic databases. The sequences of the hydrophobic segments whose similarity with their counterparts in given MSA were higher than the threshold value that was set up to 65% of amino acid sequence similarity were stored for further processing (e.g., Figures S4−S5 in the Supporting Information). During analysis of each genomic database, Pars_DB sorts out the proteins according to their overall hydrophobicity indexes (HIs) into ten sectors, namely hydrophilic proteins have the HIs from 0.0 to 30%, moderately hydrophobic proteins have the HIs from 30 to 50%, whereas hydrophobic proteins have their HIs $\geq$ 50%; the HI is defined as the number of AAs that are present in the hydrophobic patches versus their total. For example, sector A in Figure 1B contains the proteins whose HIs vary from 0.0 to 10, while sector B has the proteins with the 10% < HIs $\leq$ 20% and so on. Computing of hydrophobicity profiles was based on the methodology related to self-organized criticality (SOC) theorem[37,38] and the scales that were recently described.[39] For example, Pars_DB estimated that the highest value in the hydrophobic segment reached about 25.0, whereas the lowest hydrophilic value was set at about −18.0 if the hydrophobicity scale established by Kyte-Doolittle[40] was used. In all our analyses presented herein, hydrophobicity values that exceed 5.0 (threshold ∼20% of the maximal value) in a given hydrophobic segment of a protein were assigned to high hydrophobicity peaks (HHP).

1.3. Some Statistical Measures Used for Description of MSAs and Sequence Attributes. The following statistical measures were used for the description of some sequence attributes calculated from various MSAs: 1° the standard deviation was calculated from the following formula

$$\sigma = ([1/N \sum_{i=1}^{N} (Y_i - Y_{ave})^2])^{1/2}$$

where Y_i are the calculated values of the pIs or the HIs, Y_{ave} is their mean value, and N is the number of analyzed sequences; asymmetry and peakness of data distribution were estimated from the skewness and kurtosis using the following formulas:

$$\text{Skewness} = \sum_{i=1}^{N} (Y_i - Y_{ave})^3/[(N-1)\sigma^3]$$

$$\text{Kurtosis} = \sum_{i=1}^{N} (Y_i - Y_{ave})^4/[(N-1)\sigma^4]$$

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

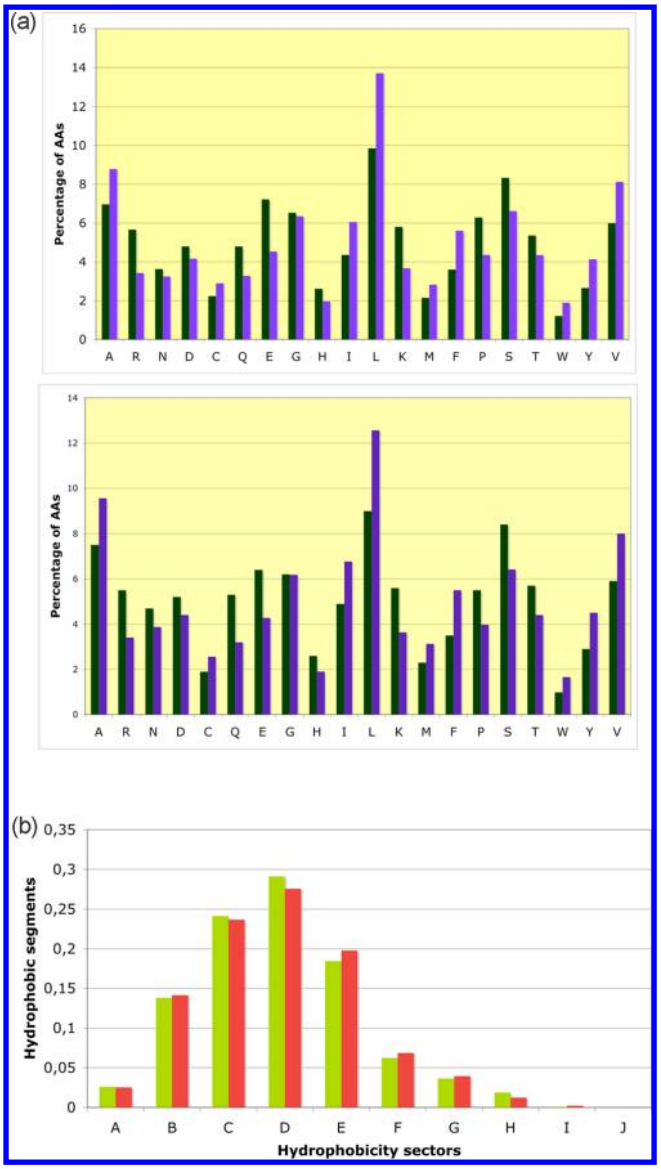


Figure 1. (A) Distributions of amino acid residues (AAs) in putative hydrophobic patches (violet bars) versus their overall contents (green bars) in the sequences of proteins encoded in the genomes of *Homo sapiens* (upper panel) and *D. melanogaster* (lower panel); (B) Distributions of hydrophobic segments in proteins that have different overall hydrophobicity; the values of the HIs were divided into 10 sectors and are shown as light green and light red bars for the proteins expressed in *H. sapiens* (35,586 proteins) and *D. melanogaster* (18,218 proteins), respectively (see the Methods).

We used the following strategy for the quantification of AAs changes in diverse MSAs: 1° each MSA was rearranged according to the descending values of the IDs that were supplied by the ClustalW program; 2° each reordered set of sequences was realigned with the ClustalW program; 3° the generated MSA was analyzed with a program that creates a transformed MSA, in which the repeated appearance of any AA at given sequence position in the MSA was replaced with a dot (Figure S11 in the Supporting Information). The output file keeps two series of sequences arranged according to 1° differences calculated for consecutive entries arranged according to descending IDs supplied by ClustalW, and 2° differences calculated between the arbitrarily chosen reference sequence versus all the remaining sequences arranged according to the descending values of the IDs. Conservation of sequence positions was assessed with Shannon's entropy $H_j(X)$, which is also known as the measure of uncertainty of e.g., sequence position j in given MSA. The $H_j(X)_s$ were calculated from the following formula[41]

$$H_j(X) = -\sum_{X=1}^{20} p_{Xj}\mathrm{Log}(p_{Xj})$$

where p_{Xj} is the probability of the appearance of amino acid residue of given type X, e.g. Phe, at position j. The fully conserved sequence position in given MSA has $H_i(X)$ = 0.0 (uncertainty of information I_e = 0.0), whereas the maximal value of $H_i(X)$ may exceed 3 if scaled down to the binary logarithm and remains ≤1 if scaled down to the logarithm of base 20.

1.4. Analyses of Structures. The Proten3_2 program[42] was used for analyses of X-ray structures downloaded from the RCSB protein data bank (www.rcsb.org).[43] Two-dimensional (2D) maps of intramolecular interaction clusters (IMICs) were calculated for interatomic distances 2.7 ≤ d ≤ 4.5 Å.[42] The sequence signature of Ly-6 domain, also known as LU-domain [(Ly-6 antigen)/(uPA receptor like domain)] and TFP-domain, is defined as follows: C_1-$(X)_i$-C_2-$(X)_j$-C_3-$(X)_k$-C_4-$(X)_l$-C_5-$(X)_m$-C_6C_7-$(X)_n$-C_8N, where the following disulfides are formed: C_1−C_3, C_2−C_4, C_5−C_6, and C_7−C_8. Diagrams of the HIs versus the theoretical isoelectric points of the aligned sequences were drawn as previously shown.[10] Compiled versions for Macintosh of all the programs for analyses of MSAs (written by the author) are available upon request.

2. RESULTS AND DISCUSSION

2.1. Distribution of Hydrophobic Segments in Proteins. In Figure 1 (upper part) are shown two panels containing the amino acid compositions (AACs) of the entire genomic databases of *H. sapiens* and *D. melanogaster* in comparison with the AACs of the hydrophobic segments calculated with Pars_DB.[36] These data illustrate that each of these databases contains the proteins, which have similar overall AACs. There is however a net increase of the contents of A, I, L, F, V, and Y in the putative hydrophobic patches (violet) in comparison to their contents in a given database (green). It is noticeable that the putative hydrophobic patches contain a sizable quantity of the AAs with polar side-chains or those with the high solubility profiles in aqueous solution, namely G and P.[39] Likewise, distributions of the HIs versus nominal masses and the calculated pIs remain comparable (Figures S4−S5 in the Supporting Information). For example, distributions of hydrophobic segments in the proteins expressed in humans and *D. melanogaster* are shown in Figure 1B. In both cases these distributions are similar even if the genomic database of *H. sapiens* contains nearly twice as many proteins as that of the fly. A similar set of data was obtained from the analysis of the *H. sapiens* genomic database (edition 2015 containing 99,739 proteins).

The distributions of high hydrophobicity peaks (HHPs) in the sequences of human proteins are shown in Figures 2A and 2B, respectively. The proteins were divided into five arbitrarily chosen sectors according to their nominal masses and the calculated pIs. On 35,586 analyzed human sequences, the HSS algorithm found 28098 putative HHPs in 13841 (39%) proteins. The group of proteins whose overall hydrophobicity is relatively small (HIs ≤ 30%) contains 2592 HHPs (19%),

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

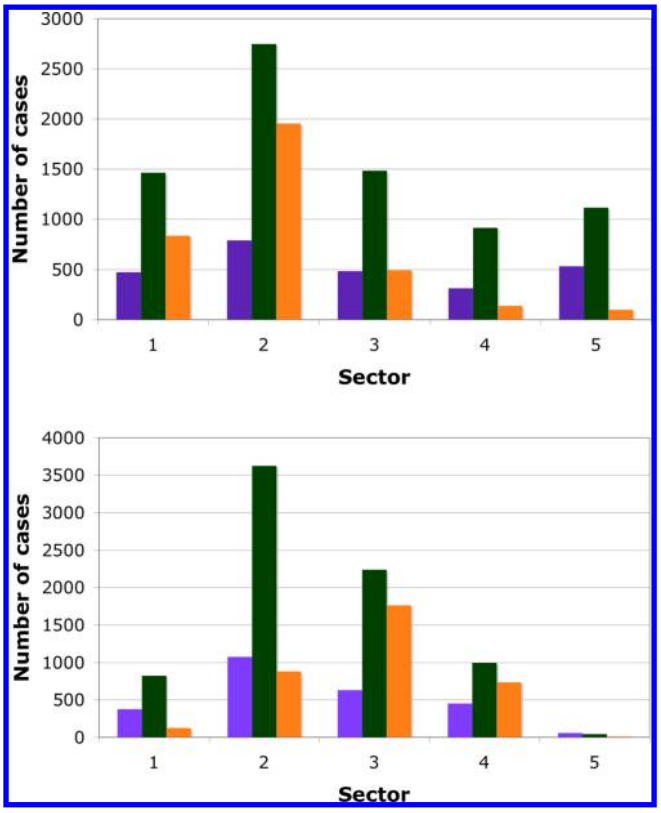


Figure 2. Distributions of proteins in the following three ranges of the HIs and five arbitrarily chosen sectors of nominal masses (upper panel) and the calculated pIs (lower panel). The proteins with the HIs ≤ 30% (violet rectangles), the HIs from 30 to 50% (green rectangles), and the proteins with the HIs > 50% (orange rectangles). The sectors in the upper panel contains proteins whose nominal masses are below 30 kDa (sector 1), from 30 to 60 kDa (sector 2), from 60 to 90 kDa (sector 3), from 90 to 120 kDa (sector 4) and with masses >120 kDa (sector 5). The sectors in the lower panel have proteins whose nominal pI are below 5 (sector 1), the pIs from 5 to 7 (sector 2), the pIs from 7 to 9 (sector 3), the pIs from 9 to 11 (sector 4), and the pIs > 11 (sector 5).

whereas the group of proteins with the high overall hydrophobicity (HIs ≥ 50%) has 3520 HHPs (25%). It is noticeable that even the highly charged proteins with the pIs above 9 or below 5 contain sizable quantities of the HHPs.

2.2. Conservation of Sequence Attributes and Hydrophobic Segments in the TGFβ Superfamily of Receptors. The sequences of the TGFβ superfamily of receptors from diverse marine genomes were aligned with their 12 human counterparts, and the resulting MSA was analyzed with the Hela_MSA program, which delineated variations of their hydrophobicity profiles, AA bulkiness variability along the polypeptide chain, changes in amino acid composition (AAC), and distribution of charges (pIs). Although the sequences of the TGFβ receptors studied here have the moderate HI values, they contain several HHPs. The TFPD at the N-terminus contains two small hydrophobic segments and is linked via a highly hydrophobic TM segment (an HHP) with a hydrophilic GSGS-rich domain (TGFβ RI type only) that associates with FKBP12 that is terminated with ALK domain (Figure S6 in the Supporting Information). It is noticeable that several α-helices in the C-terminal lobe of the ALK domain are hydrophobic, especially the internal α-helix (violet, HHP, known as α_F-helix),[44] which form the hydrophobic core of the helix bundle. The 12 ECDs of the human members of the TGFβ family of receptors plus the ECD of the BAMBI pseudoreceptor have an average sequence similarity ID_{ave} = 21%, whereas the 12 ALK domains have the average ID = 46%. Each of the two series of TGFβ-receptors (types I and II) have a high level of sequence conservation, namely analyses of the MSA containing 476 sequences of type I receptors that are expressed in different species gave the ID_{ave} = 68% (Table S1 in the Supporting Information).

It has been shown that 2D tandem maps of intramolecular interaction clusters (IMICs) have similar patterns for the canonical TFPDs such as snake neurotoxins and the ECDs of various TGFβ-receptors despite the generally low ID_{ave} for their sequences.[3,6,9] Likewise, relatively high levels of conservation of the IMICs were found on the tandem 2D maps of diverse kinase domains, namely in Figure 3 is shown a tandem 2D map containing the IMICs calculated for the kinase domains of the human TGFβ receptor I (ALK5, 1PY5)[45] and aurora kinase A (3UNZ).[46] The secondary structure in the N-terminal lobes of these two kinases (indicated in the red dotted lines at the upper left corner) and the large C-terminal lobe containing several α-helices show for a similar distribution of IMICs in these two kinase domains despite the low sequence similarity (ID = 17%).

A set of hydrophobic sequences from the ALK domains from human TFGβ series of receptors was simultaneously used as input to Pars_DB in order to find all proteins in various genomic databases that have similar sequence patterns to ALK (see Methods). The searches revealed that variable numbers of the TGFβ family of receptors are expressed in those diverse marine organisms. For example, in the database of *C. intestinalis* (phylum: chordate) the algorithm detected significant sequence similarity between the TGFβ series of receptors expressed in this archetypal deuterostoma with some segments from the human ALK domains. Sequence segments from several kinases also have some similarity to human ALK domains such as centromere/kinetochore protein (ZW10 homologue; XP_002127662) and mixed lineage kinase (MLK) domain-like protein (XP_002122121). Likewise, analyses of the *Hydra magnopiliata* (phylum: cnidarian) genomic database have revealed that this "perennial" organism expresses several TGFβ receptors whose ALK domains are similar to their counterparts in the hTGFβ series of receptors. Analyses of the *S. purpuratus* (phylum: echinodermata), *A. queenslandica* (phylum: porifera), and *S. kovaleskii* (phylum: hemichordata) genomes have revealed that they encode several members of the TGFβ family of receptors (Figure S7 in the Supporting Information).

In the *D. melanogaster* genome (phylum: arthropod) are encoded five members of the TGFβ family of receptors, namely three receptors type I [thickveins (*tkv*), saxophone (*sax*), and baboon (*babo*)] and two receptors type II [wishful thinking (*wit*) and punt (*put*)], which interact with secreted growth factors (morphogens) such as decapentaplegic (*dpp*), screw (*scr*), glass bottom boat (*gbb*), activin-β (*act*), maverick (*mav*), myoglanin (*myg*), and dawdle (*daw*). Binding of a growth hormone to receptor type II involves recruiting and trans-phosphorylation of receptor type I, which in turn phosphorylates one of the mothers-against-dpp (MAD) transcription factors depending on the functional feature of the bound growth factor.[47] Structural homologues of the above receptors are encoded in the *Homo sapiens* genome, namely there are 12 members of the TGFβ family of receptors and one pseudoreceptor (BAMBI). Using 15 hydrophobic sequence

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

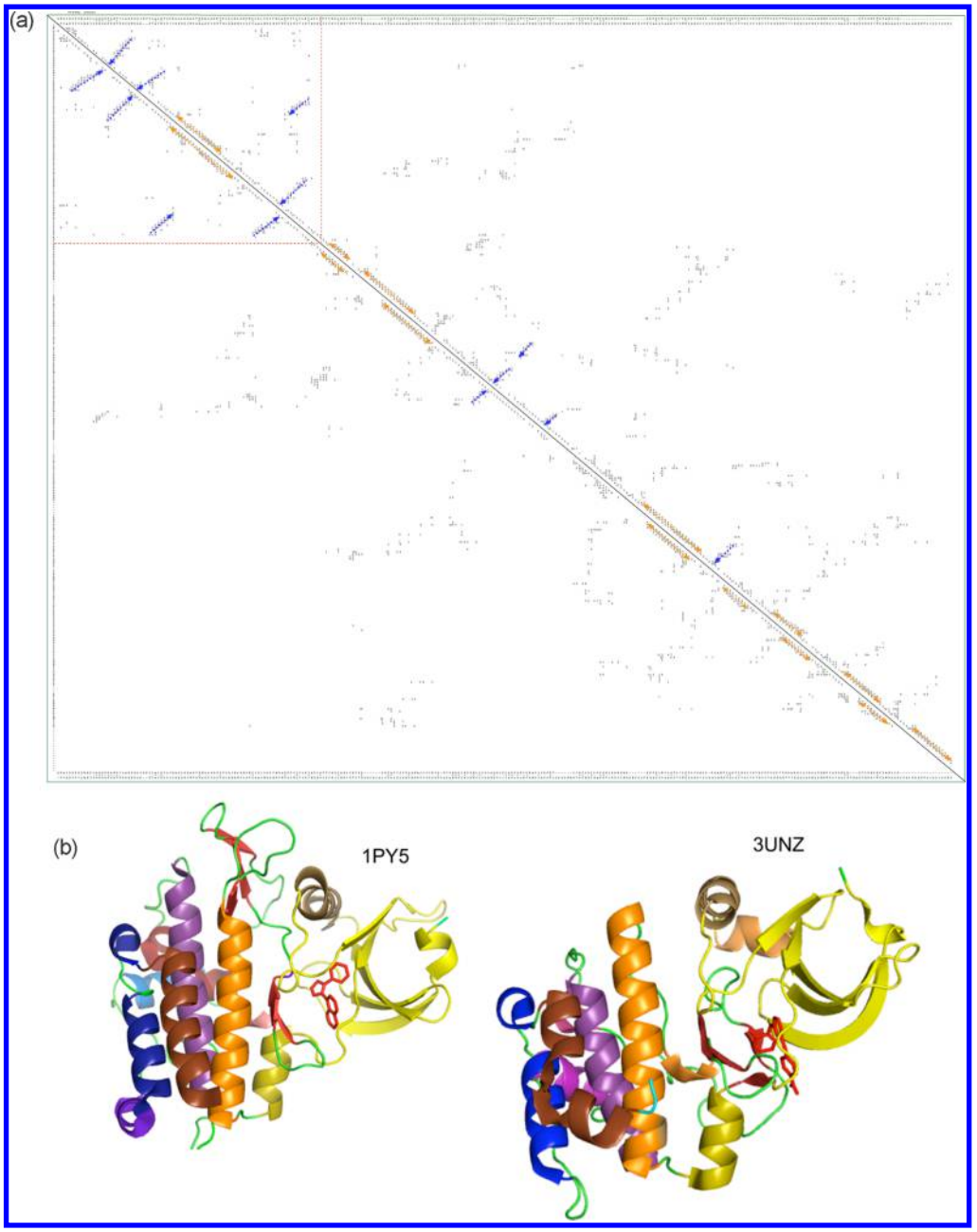


Figure 3. (A) 2D distance map calculated from the X-ray structures of the following two kinase domains: the hTGFβ receptor I (ALK5; 1PY5; upper panel) and aurora kinase A (3UNZ; lower panel); (B) X-ray structures of TGFβ receptor I kinase (ALK5) + 4-(3-pyridin-2-yl-1H-pyrazol-4-yl)quinoline that is in red sticks (1PY5, upper panel),[45] and Aurora kinase A (ARK-1) + 4-({4-[(2-fluorophenyl)amino]pyrimidin-2-yl}amino)benzoic acid that is in red sticks (3UNZ, lower panel).[46]

fragments from the ALK domains of the human TGFβ-family of receptors, Pars_DB found all the above-mentioned sequences in the *D. melanogaster* database. The Pars_DB searches also have revealed that several short sequence fragments belonging to various kinases possess some similarity to human ALK domains, namely stardust CG32717-PB and center divider CG6027-PA (Figure S8 in the Supporting Information).

2.3. Monodomain TFPs and Their Sequence Variants Encoded in Miscellaneous Organisms. The BLAST and Pars_DB programs revealed that the genomic databases of all the marine organisms, insects, and mammals that have been analyzed in this communication contain various numbers of the TGFβ receptors.[5−10,47] In contrast, small monodomain TFPs are sparsely expressed in different marine organisms such as a prostate and testis expressed (PATE)-like protein that is encoded in the *S. purpuratus* genome. Genomes of some phyla encode specific groups of proteins that contain one putative TFPD. For example, there are some large ORFs coding for fusion proteins, called low affinity cationic amino acid transporters 2-like (about 700 AAs), which contain N-terminal highly hydrophilic TFPD.[19] These types of ORFs were found in the genomes of different insects, such as the honeybee *Apis mellifera* (XP_393753.3|gi:328791294).[29] The genome of the *D. melanogaster* fly contains multiple ORFs, which encode small proteins with a CN doublet (one of the signatures of TFPDs) at the C-terminal part followed by a hydrophobic stretch that can link the protein to the extracellular membrane via

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

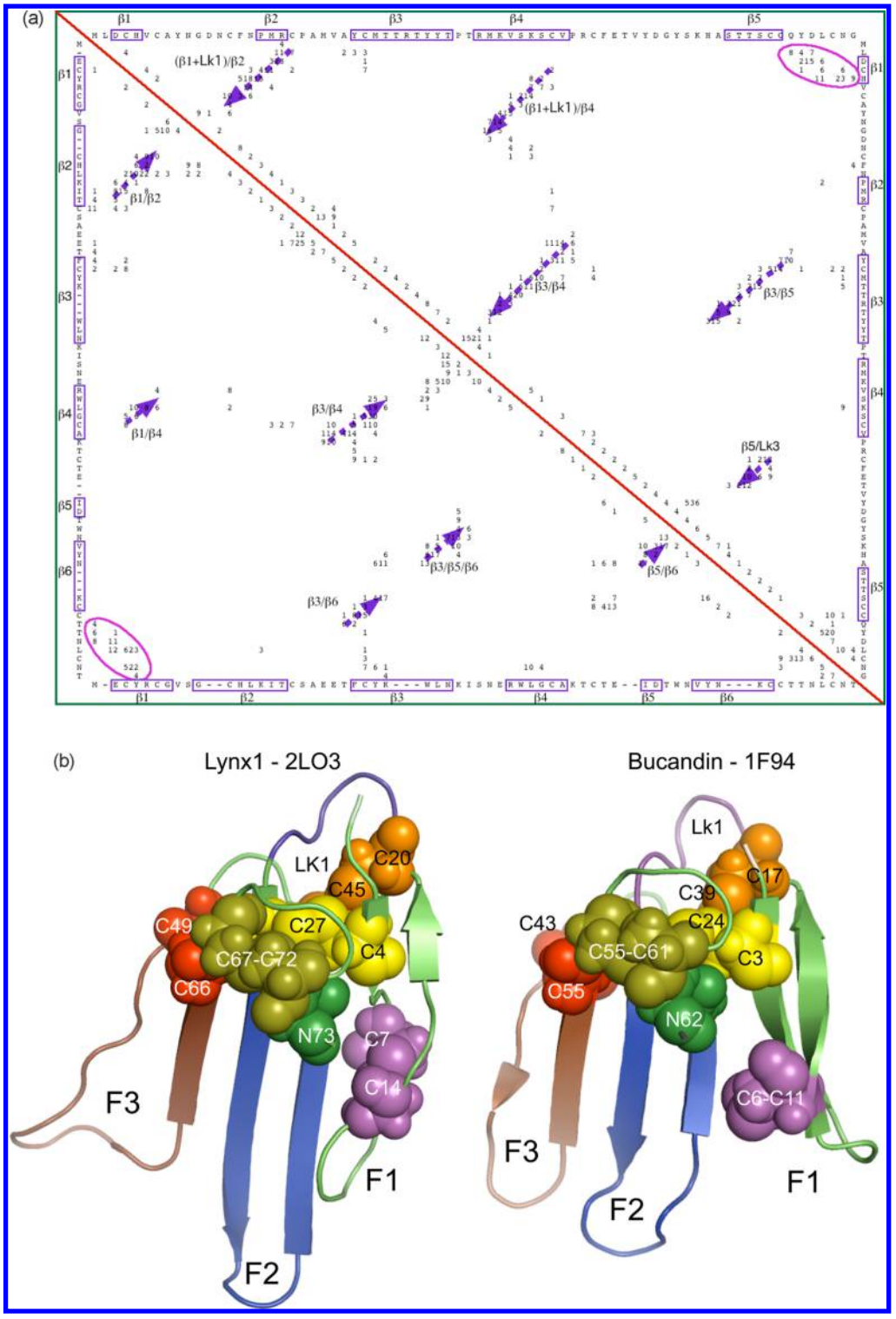


Figure 4. (A) 2D distance map calculated from the NMR-structure of soluble human Lynx1 protein (2L03, upper panel)[59] and the X-ray structure of bucandin (1F94, lower panel).[60] Antiparallel β-structures (violet arrows) are perpendicular to the diagonal (light green), whereas the IMICs between the N- and C terminus are indicated in red ovals. (B) The structure of soluble form of hLynx1[59] (left panel) and bucandin[60] (right panel); F1, F2, and F3 designate consecutive fingers; Lk1 is the loop linking F1 with F2. The four conserved disulfides are shown: (internal S−S, yellow spheres), which is flanked by orange and C-terminal (deep olive, C-terminal S−S bond); the S−S bond which is at the base of the palm but is positioned beyond the three stacked disulfide bridges is shown as red spheres in both structures; the disulfide at the bottom of F1 is shown as a violet sphere.

glycophosphatidylinositol (GPI) modification.[50] For example, it has been shown that retroactive (NP_572693|gi:18859727) is crucial for cuticle organization in larva,[51] whereas the proteins called crooked (NP_609557|gi:20129467), coiled (NP_608536|gi:19920474), and crimpled (NP_648500|gi:21357259) are crucial for proper localization of septate junction components.[52] Indeed, "coiled" has the sequence arrangement of the C residues similar to that in the Ly-6 signature, whereas the sequences of retroactive, crooked, and crimpled only resemble it to some extent (Figure S9 and Table S2 in the

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

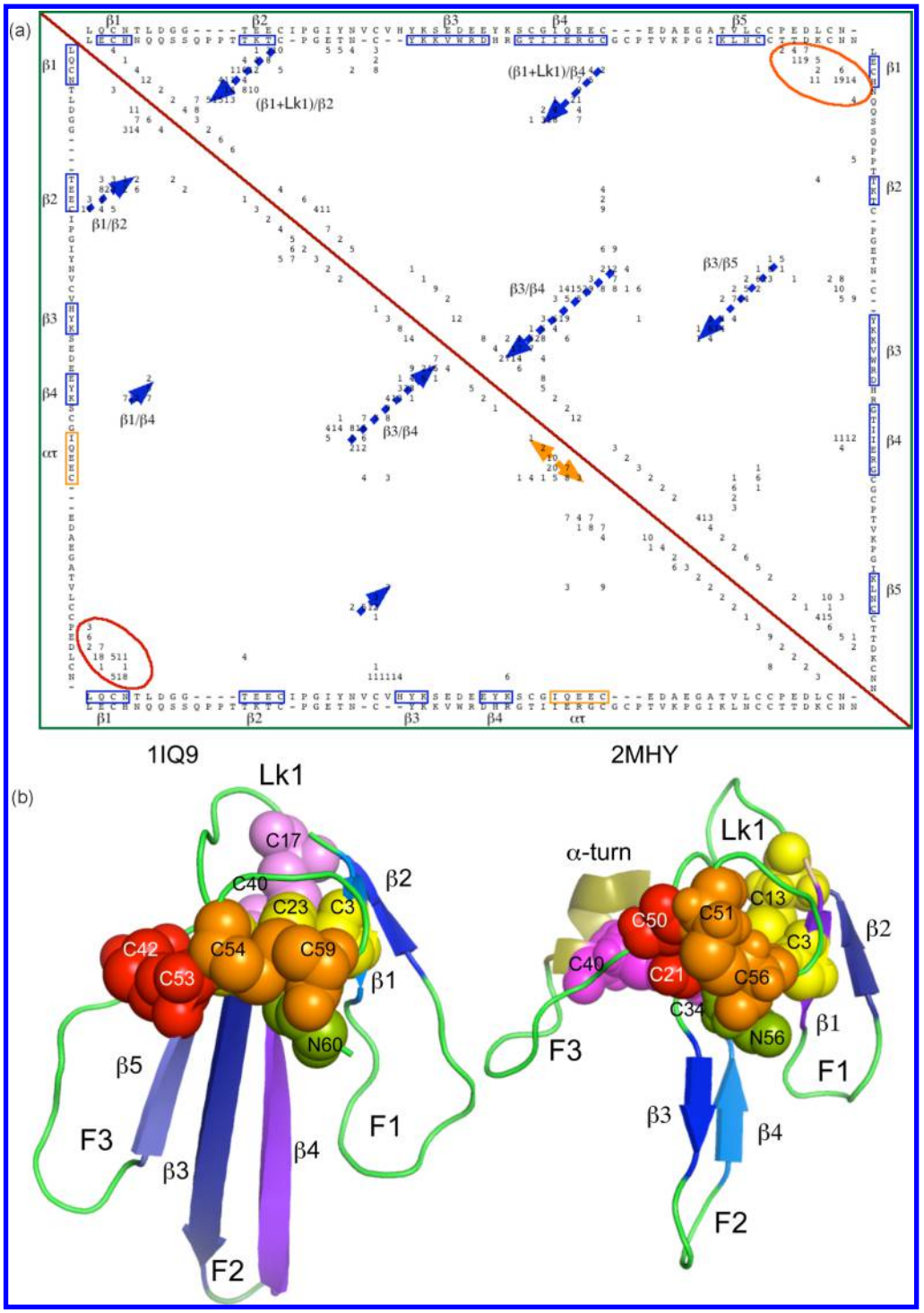


Figure 5. (A) Upper panel: 2D tandem distance map of the IMICs calculated from the structure of short neurotoxin (right triangle) from *Naja nigricolis* venom (1IQ9)[63] and the courtship pheromone plethodontid modulating factor (2MHY)[62] from *Plethodon shermani* (left triangle); sequence alignment (ID = 21%) is shown at the horizontal axes. Blue arrows indicate antiparallel β-structure, whereas the yellow arrow represents short α-helical segment; (B) At the lower panel are shown the X-ray structure of the short neurotoxin and the PMF from Sherman salamander, respectively.

Supporting Information). Moreover, the ID_{ave} for these 15 proteins is about 15%, which indicates for an extreme differentiation of their sequences. The equally low ID_{ave} = 16.5% was calculated from the MSA of 40 five TFPDs belonging to 40 two proteins expressed in humans, i.e., 40 monodomain and two multidomain TFPs (Table S3 in the Supporting Information).

It has been shown that the sleepless protein (*D. melanogaster;* ACF58241.1|gi:194369501) and its sequence variant called quiver (gi|194369501|gb|ACF58241.1;NP_001137646.1|gi:22-1330178) bind to the potassium channel Shaker and control the sleeping period.[53−55] Although these two proteins are being considered as the members of the Ly-6 family, their sequences however do not have all the elements that are typical for the

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

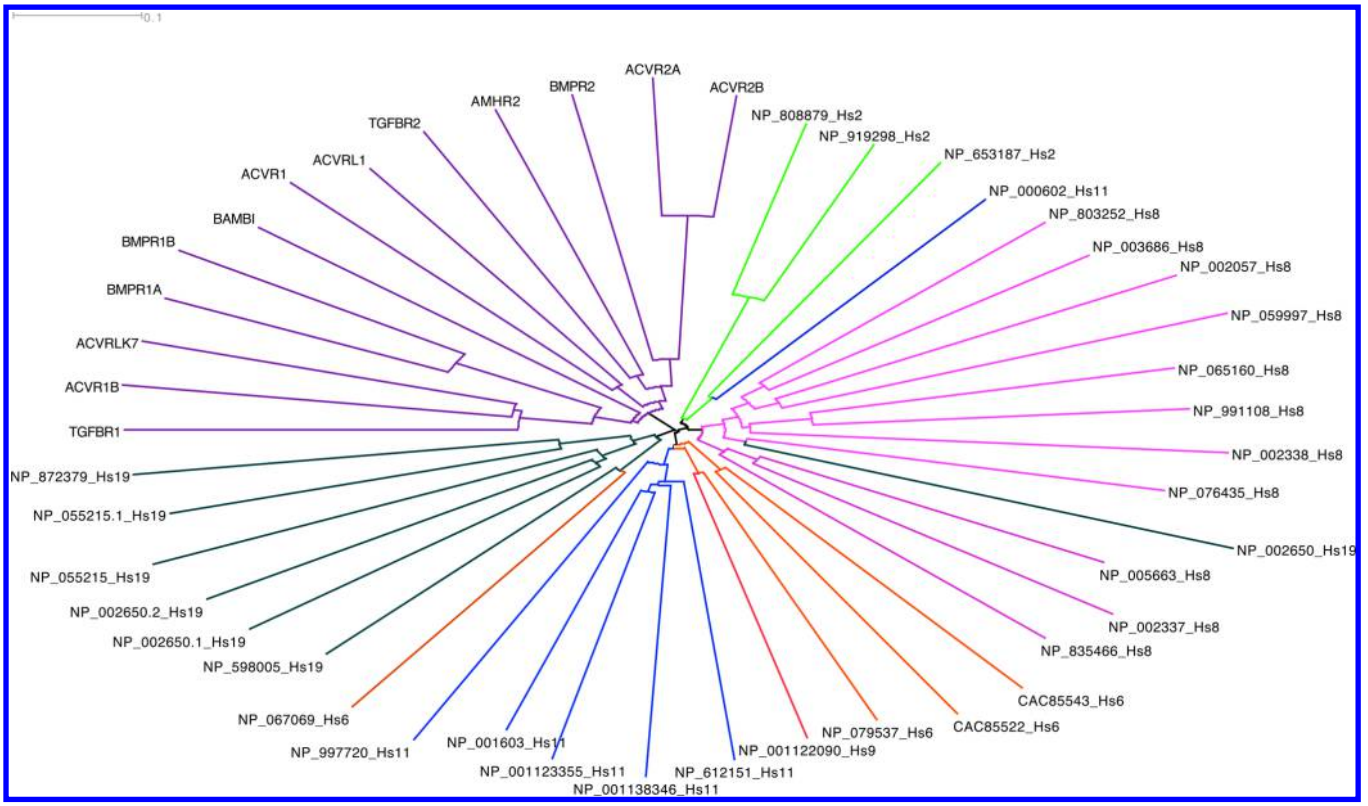


Figure 6. Unrooted tree calculated from the MSA of 45 human TFPDs (Figure S17, Supporting Information). TGFβ-series of receptors bearing their genes' names are in deep violet, whereas the branches bearing database codes for the other TFPDs that are encoded on different chromosomes have the following colors: chromosome II (light green), VI (orange), VIII (rose), IX (red), XI (blue), and XIX (deep green).

canonical TFPD signature.[9,10] For example, they have an odd number of C residues (Figure S9 in the Supporting Information) with two pairs of CCs, one of which could form an intermolecular SS bond. In fact, several structural studies coupled with functional assays have shown that some TFPs form homo- or heterodimers that are linked via an intermolecular disulfide bond such as α-cobratoxin from *Naja kaouthia* (4AEA)[48] and irditoxin (2H7Z),[49] respectively. As to whether the proteins with apparent Ly6-like sequence signature[50−55] that are expressed in the fly could adopt the canonical TFPD fold or would they rather adopt other SS-rich folds remain to be solved. For example, the Argos protein that sequesters Spitz (EGF-family of proteins in *D. melanogaster*) has several fingers (3C9A),[56] but its fold does not have any typical attribute of the TFPD. Likewise, the X-ray and NMR-structures of Dickkopf (3S2K and 2JTK; Figure S10 in the Supporting Information), which binds to low-density lipoprotein receptors (Wnt coreceptor LRP5/6) display "three-fingers", but their spatial arrangements and disulfide interaction networks are different than it is in the TFPD fold.

Organisms belonging to different phyla express a number of TFPDs containing proteins whose functions remain conserved throughout the evolution, but some species had adapted this fold to proteins with particular functions. For example, small TF-like proteins form a large spectrum of toxins (7−9 kDa) that are produced in venomous glands of some snakes.[9,57,58] In Figure 4A is shown a tandem 2D map of intramolecular interaction clusters calculated from the NMR-established structure of soluble form of human Lynx1[59] (upper triangle) and the snake neurotoxin bucandin[60] (lower triangle); structures of these two proteins are shown in Figure 4B. Despite low sequence similarity (ID = 26%) the distribution of the IMICs are similar for both structures. Various types of salamanders produce plethodontid-modulating factors (PMFs)[19,61] that have some sequence similarity to the canonical signature of snake short neurotoxins. Despite the different topology of disulfides[62] in the PMF from Sherman salamander as compared to the disulfide topology in the TFPD of a snake short neurotoxin from the cobra *Naja nigricollis* (1IQ9)[63] our analyses of the IMICs show that these patterns share a significant similarity to each other (Figure 5, upper panel). Moreover, the IMICs between the N-terminus and the C-terminal CN doublet remain conserved in both structures as well as the overall interaction patterns between β-structures, which are two typical features of the TFPD fold.[9,10] It is noticeable however that the stacking pattern of the disulfides is altered in the PMF as compared to that in the TFPD fold[9,10] (Figure 5, lower panel).

2.4. Multidimensional Drift of Sequence Attributes (MDSA) Causing Fine Diversification of Functional Profiles in the LU-Domain (Ly-6/uPAR and Diverse TFP-like Domains). In the human genome are encoded at least 40 two genes of TFPs, which contain from one to three TFPDs.[9] Archetypal forms of some members of genes on human chromosomes II, III, and XII coding for the TGFβ-receptors (Table S3 in the Supporting Information) had to be already coined at an early stage of evolution of multicellular organisms such as *A. queenslandica*. In Figure 6 is shown a simple tree calculated from the MSA of all 45 TFPDs present in human proteins (Figure S17 in the Supporting Information). The ECDs of TGFβ-series of receptors are clustered together as well as the majority of TPFDs of proteins encoded on

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

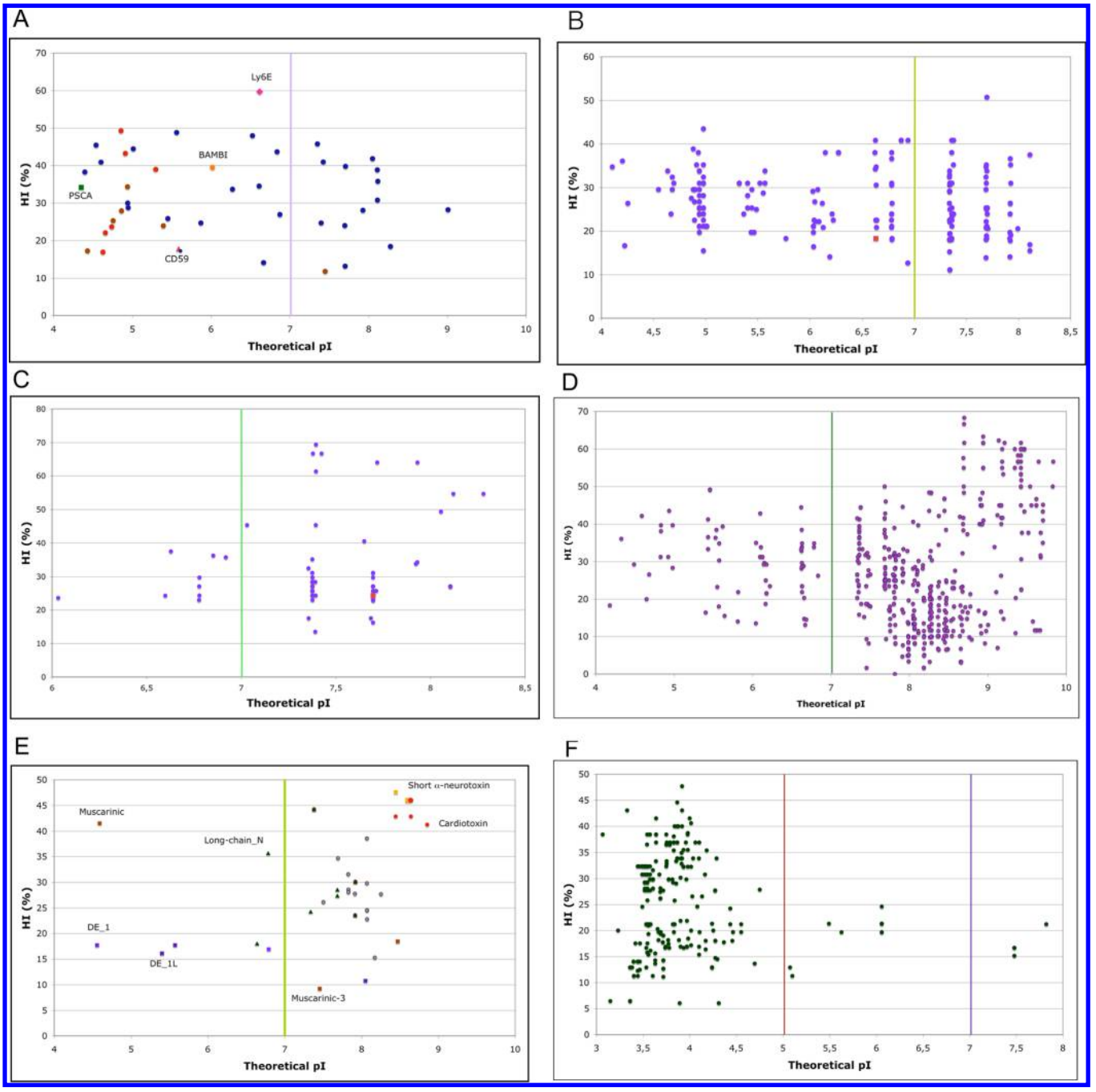


Figure 7. Distributions of the HIs versus the theoretical pIs: (A) 45 TFPDs from human proteins; CD59 is red triangle, Ly6E is red diamond, and the series of TGFβRI and TGFβRII are in red and brown circles, respectively (average ID = 16.5%; pI_{ave} = 6.2 ± 1.4, skewness = 0.3, kurtosis = 1.8, and HI_{ave} = 32.0 ± 11.2; skewness = 0.1, kurtosis = 2.3 (Table S3 in the Supporting Information); (B) 186 CD59s from various sources (pI_{ave} = 6.4 ± 1.2, skewness = −0.3, kurtosis = 1.6, and HI_{ave} = 26.3 ± 7.2; skewness = 0.4, kurtosis = 2.8 (Table S5 in the Supporting Information); (C) 76 Lynx1 from different sources (HI_{ave} = 7.4 ± 0.8, skewness = 0.3, kurtosis = 1.8 (Table S4 in the Supporting Information); (D) TFPDs in diverse TFPs present in snake (Table S6 in the Supporting Information); (E) 40 toxins from king cobra's venom; muscarinic toxins are as brown squares, weak neurotoxins as violet squares, cardiotoxins are red circles, short hydrophobic neurotoxins are as yellow squares, some of the long neurotoxins are green triangles, and the rest is red-blue circles (ID_{ave} = 40%; pI_{ave} = 7.4 ± 1.4, skewness = −2.8, kurtosis = 11.7; HI_{ave} = 25.3 ± 9.2, skewness = 0.3, kurtosis = 2.1 (Table S8 in the Supporting Information); (F) 231 TFPD-like sequences in the PMFs from miscellaneous salamander species (ID_{ave} = 51%; pI_{ave} = 3.9 ± 0.7, skewness = 3.3, kurtosis = 15.8; HI_{ave} = 25.3 ± 9.2, skewness = 0.0, kurtosis = 2.0 (Table S9 in the Supporting Information).

chromosomes II, VIII, XI, and XIX. Only CD59 (chromosome XI) is clustered with the TFPDs from chromosome II. The former is located on arm p (11p13), whereas the genes coding for the remaining three TFPs are on arm q (11q24.2). Domain I from uPAR clustered with the TFP locus on chromosome VIII. Domain I of uPAR however does not have one of the SS bonds at the C-terminus, which is an unusual feature in the LU-type fold.[9] This dendrogram may suggest that extensive duplication events of genes coding for various TFPs had taken place at different stages of evolution of species.

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

In Figure 7A-F are shown distributions of the HIs versus the theoretical pIs for various human TFPDs (7A), the members of the Lynx1 and CD59s expressed in various organisms (7B-C), several hundreds of TFPs expressed in venoms of miscellaneous snakes (7D), 40 toxins from King Cobra (*Ophiophagus hannah*) venom (7E),[64] and the PMFs from diverse salamanders (7F). The HI versus pI diagram for human TFPDs shows that these two attributes span a relatively large range of values, namely the HIs change from the highly hydrophobic values such as that for the domain in Ly6E to the highly hydrophilic range as it is in the domain of CD59, whereas their pIs vary from the basic to acidic values. There is however a lesser dispersion of these attributes for the LU-domain of the Lynx1 proteins in comparison to the CD59s expressed in miscellaneous species. These two sequence attributes of the Lynx1 proteins remained well conserved throughout evolution of various organisms, and they are similar to those of snake short neurotoxins. In contrast, these two sequence attributes have relatively large dispersion in the LU-domain of the CD59s, which could indicate that the genes coding for Lynx1 and CD59 had been subjected to different levels of evolutionary pressure; changes of AAs in the sequences ordered according to descending IDs between human protein and its orthologous forms expressed in various organisms are shown in Figures S11−S12 and Tables S4−S5 in the Supporting Information. A distribution of the HIs versus pIs of the TFPs present in the venoms of some snakes illustrate that a great part of them are basic proteins (Table S6 in the Supporting Information). The diagram suggests that diversification of their functional profiles was achieved via modifications of the overall hydrophobicity levels (HIs). For example, the cardiotoxins are basic (pI_{ave} = 9.2 ± 0.4) for 95 sequences (Figure S14 and Table S7 in the Supporting Information), and most of them are both hydrophobic and positively charged proteins (average HI = 48.0% ± 11.1%). In fact, the cytotoxins bind to cellular membranes and may target some receptors, which could be dependent on their hydrophobicity levels.[65−67] The group of short neurotoxins from snake venoms is composed of basic and hydrophilic proteins. Weak, long, and muscarinic toxins have the pIs from 4.5 to 7.5, and their HIs vary from 15 to 30%. Some of the long neurotoxins however have their pIs below 7. A similar distribution of the HIs versus the pIs is shown for the 40 toxins from the venom of king cobra (Figure S15 and Table S8 in the Supporting Information). A different distribution of the HIs versus pIs was obtained for the PMFs (Table S9 in the Supporting Information), which indicates that they share a coherent function in some amphibians.

2.5. A Hypothesis on Multidimensional Drift of Sequence Attributes in TFPs. The analyzed genomic databases contain from 15,000 to nearly 70,000 sequences including various isoforms. Having this in mind one could ask the following question: how was sequence/structure/function paradigm sampled in various groups of the TFPs. We propose that the following steps may explain a considerable genetic drift of genes coding for those proteins: 1° The TFPDs were divided into two groups of proteins, namely those that have the acidic pIs (the majority of Ly-6s, CD59s, and the ECDs of the TGFβ receptors) and those with the basic pIs such as a part of the Ly-6 domain-containing proteins expressed in various animals and toxic components secreted in venomous glands of some snakes.[68,69] For example, the majority of the neurotoxins and cardiotoxins from snake venoms are positively charged (their pIs > 7), since they must have a sizable tropism toward the negatively charged cellular membranes[65−67] and some receptors.[11,12,68−70] In contrast, the PMFs whose sequences have TFPD-like signatures exhibit the low pI values, since their function is to interact with pheromone sensing vomeronasal receptors (Jacobson's organ). It implies that they cannot be "too sticky" to a negatively charged cell surface. 2° The structural framework that is stabilized by the closely interacting three SS bonds and fully conserved N residue at the C-terminus (the CN doublet) was varied according to the hydrophobicity level, namely from an extremely hydrophobic range that is occupied by some cardiotoxins, some short neurotoxins and Ly6E proteins to a very hydrophilic range that is populated with several groups of proteins such as the CD59s or the putative TFPDs in plants' cation transporters II. Variations of these two sequence attributes are different for each group of functionally related proteins. For example, the CD59s underwent a more extensive species-specific adjustment of their inhibitory properties that was necessary to counter the lysis capacity of the complement complex, whereas the Lynx1s had to maintain their physical-chemical sequence attributes conserved in order to specifically target the nicotinic receptors. 3° Functional epitopes were varied via extensive AA-substitutions and indels within the three fingers of the TFPD fold. In consequence, only several sequence positions remain highly conserved (Shannon's entropy I_e = 0.0) in diverse groups of the TFPs. For example, in Figure 8 are shown three diagrams with Shannon's entropy calculated from the MSAs of the series of 95 sequences of cardiotoxins (A), 76 sequences of Lynx1s (B), and 40 sequences of toxic polypeptides from the venom king cobra (C) (see also Figure S16 in the Supporting Information).

The sequences of cardiotoxins from snakes' venoms have several fully conserved positions (ID_{ave} = 69.5%). Likewise, 76 sequences of Lynx1s Figure 8B) and 40 sequences of polypeptides from the venom of king cobra (Figure 8C) have nine and ten sequence positions fully conserved. However, there is a general trend for an increase of the I_e values from A to C which implies that the sequences in king cobra venom have a higher functional diversification than in the other two cases. It is also noticeable that the ID_{ave} = 40% for the 40 toxins present in king cobra venom is much better than the ID_{ave} for the TFPDs encoded in the human genome, which implies for an extreme diversification of functional properties of the Ly-6/uPAR domains expressed in humans. It could be concluded that hypervariability of some sequence positions within the TFPD fold created a multitude of recognition profiles of their targets that became also compatible with the physical-chemical condition of the matrix in which they function. 4° Due to low overall sequence similarity in the superfamily of proteins with the Ly-6/uPAR/TFP signature (LU domain), which contain from 6 to 12 Cys residues, the pairing of some of them may differ from the S−S pairings in the canonical TFPDs. For example, even if the series of the PMFs have low sequence similarity to snake short neurotoxins and their disulfide bridge connectivity is different, they possess a TFP-like fold.

■ CONCLUSIONS

Our study has revealed that the genomes of various species contain proportionally similar extent of the hydrophobic sequence space (HSS), which is one of the main factors in protein folding.[71,72] About 40% of all the proteins in a given genomic database contain high hydrophobicity peaks (HHPs). Several HHPs were detected in the sequences of the ALK domains and in some of the sequences of diverse TFPDs. We

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

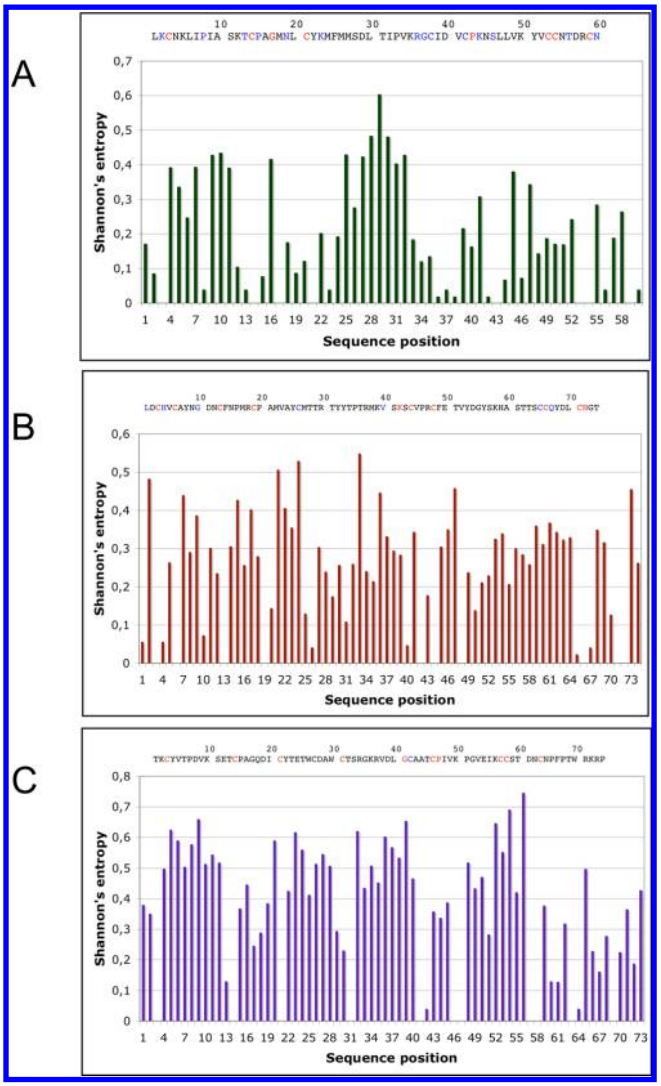


Figure 8. Shannon's entropy (I_e) for three sets of sequences calculated as the logarithm at base 20. (A) Ninety-five sequences of cardiotoxins from the venoms of different snakes; (B) Seventy-six sequences of Lynx1s from disparate organisms; (C) Forty sequences of toxic polypeptides from the venom of king cobra. At the top of each diagram is given the first sequence from the MSA as a reference with the AAs in red if fully conserved, namely in A is the sequence of cardiotoxin I-like protein [CAA63978, *Naja naja*]; (B) Ly-6/neurotoxin-like protein 1 isoform c (NP_803252; *Homo sapiens*), and (C) cardiotoxin precursor [ABB83631; *Ophiophagus hannah*]. Lists of the analyzed sequences are given in the Supporting Information.

have described a hypothesis on multidimensional drift of sequence attributes, which involves local and total hydrophobicity variations, charge distribution, polarizability, and the overall pI and which cause significant diversification of specific functions within multigene family of proteins.[73] As an example we used different genetic constructions containing from one to three TFPDs. For example, the analyzed structures of the TFPDs and ALKs have different levels of conservation of their sequences and hydrophobic patterns. For example, the ALK domain has a higher sequence conservation level across the metazoan kingdom than the ECDs of the TGFβ series of receptors. It is due to the fact that there are few constraints on sequence modifications within the three fingers of the ECDs, which could have acquired fine targeting epitopes for their activation ligands and inhibitors. In contrast, many sequence positions in the ALKs must remain conserved in order to sustain the intrinsic kinase activity of the domain.

In each of the analyzed genomic databases we have detected at least several TGFβ-like receptors. Thus, it is probable that the ECD of a TGFβ-like receptor was the primordial functional genomic construct with the TFPD fold. The exons of such a primordial ORF could have been acquired via reshuffling of exons coding for their ligands, which have the cystine-knot fold.[6] It is conceivable that from such a primordial genetic construction could have emerged other forms of TFPDs, namely BAMBI-like pseudoreceptor and other small proteins with TFPD-like fold. It is worth noting that the sequences of the ECDs of the TGFβ receptors are longer (85−120 AAs) than the nominal sequences of snake short neurotoxins (60−65 AAs). The former contain additional spatial motifs such as α-helices or disulfide bonds in the first or third fingers, but their spatial forms still retain common structural traits with each other.[9]

Our analyses have shown that relatively few single domain TFPDs are encoded in the genomes of miscellaneous marine organisms, whereas they contain from two to six homologues of the TGFβ-like receptors. Some organisms of a relatively ancient origin such as salamanders secrete a specific group of small proteins called courtship pheromones (PMFs), which have TFPD-related sequence signature and fold. It is tempting to speculate as shown in Figure 8 that there could have been a transformation of an archetypal gene coding for a small TFP to an ORF coding for a primordial PMF-like factor, from which would have later radiated a plethora of diverse TFPs that are produced in the venoms of some snakes.[57,58,68−70]

Such transformation could have been achieved via multidimensional drift of sequence attributes (MDSA). The sequences of the PMFs and some toxic components in snakes' venoms were shortened as compared to their longer forms adapted for the ECDs of the TGFβ-series of receptors (see Figure 9). For example, it is conceivable that an archetypal gene coding for an isoform of Lynx1C, which does not have the C-terminal hydrophobic segment (GPI modification), could have been adequate as an origin of primordial genes coding for the amphibian-specific PMFs. This type of gene could have been later transformed into a series of genes coding for the TFPs that are expressed in the glands of venomous snakes (class: reptilians). It is worth noting however that primordial snake-like organisms probably originated about 140,000 millions years ago,[57] which is nearly half a billion years after the formation of primordial multicellular organisms whose genomes had acquired multiple ORFs coding for the ECDs of the TGFβ-series of receptors. Moreover, AA mutations in various TFPDs do not have similar functional constraints. For example, all AA mutations in the TFPDs that are a part of given organism must be functionally acceptable in order to accommodate the overall genetic drift and differentiation of a given genome, whereas AA mutations in the TFPs which function on the outside of the system such as neurotoxins and other components expressed in snakes' venoms could have mutated more freely their AA-profiles in order to effectively target crucial receptors in their prey.

Diversified functional profiles of various groups of proteins having from one to three consecutive TFPDs encoded in the genomes of disparate species were summarized in Table S10 (Supporting Information). This list shows that the gain-of-function in those disparate TFPDs was probably fashioned via

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

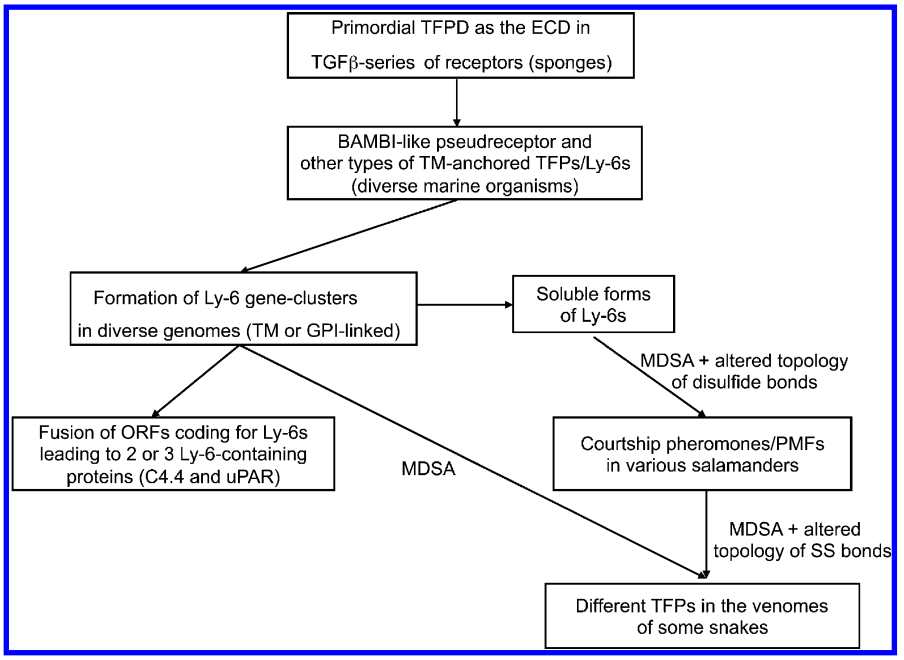


Figure 9. Schema of hypothetical events that could have been involved in transformation of the ECD in the TGFβ series of receptors to a membrane-anchored TFPD (Ly6/uPAR/TFP fold, aka LU-domain). The latter could have been transformed into soluble TFPs, which via extensive MDSAs and alteration in disulfide bond topology acquired PMF-like activity in some primordial salamanders, and which could have been one of the possible sources for a series of genes coding for diverse TFP-like toxins secreted to glands of some venomous snakes.

acquisition of adequacy, finesse, and strength of interactions between their surface epitopes and those on their targets. In fact, analyses of intermolecular interaction clusters of various complexes containing TFPD bound to different targets have indicated that the fold is a multifaceted entity.[6,9] Our analyses presented here could suggest that the gain-of-function within this multigene family of proteins had been achieved via genetic mutations involving diverse indel events which had caused multidimensional drift of their sequence attributes and allowed them to attain the physical-chemical and unique functional properties that are also compatible with the conditions in which a given receptor/ligand became operational. It is probable that the canonical TFPD gave rise to a multitude of TFPD-like signatures, which are present in the sequences of many proteins that are expressed in various species belonging to miscellaneous phyla. To what extent however the folds of the sleepless, quiver, crumbled, and the other Ly6-like proteins in *D. melanogaster* and other species[74] as well as the CD177[75] or TEX101[76] in the vertebrates resemble the fold of the canonic TFPD would be revealed once X-ray structures of these proteins become available. More studies are needed for the assessment of functional inputs of the diversified groups of proteins with TFP/Ly-6/uPAR-like signature to miscellaneous processes such as cell−cell recognition,[77,78] inflammation,[79,80] tissue regeneration,[81,82] and regulation of various signaling cascades.[4−8,83]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00322.

Input and output files from the software used in this work, MSAs of various groups of proteins, tabulated sequence attributes of the aligned sequences, and several figures with X-ray structures and Shannon's entropy (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Fax: 33-1-69089071. E-mail: galat@dsvidf.cea.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported financially by the SIMOPRO/IBITECS/DSV/CEA.

■ REFERENCES

(1) Low, B. W.; Preston, H. S.; Sato, A.; Rosen, L. S.; Searl, J. E.; Rudko, A. D.; Richardson, J. S. Three Dimensional Structure of Erabutoxin-B Neurotoxic Protein: Inhibitor of Acetylcholine Receptor. *Proc. Natl. Acad. Sci. U. S. A.* **1976**, *73*, 2991−2994.

(2) Tsernoglou, D.; Petsko, G. A. The Crystal Structure of a Post-Synaptic Neurotoxin From Sea Snake at Å Resolution. *FEBS Lett.* **1976**, *68*, 1−4.

(3) Greenwald, J.; Fischer, W. H.; Vale, W. W.; Choe, S. Three-finger Toxin Fold for the Extracellular Ligand-Binding Domain of the Type II Activin Receptor Serine Kinase. *Nat. Struct. Biol.* **1999**, *6*, 18−22.

(4) Vitt, U. A.; Hsu, S. Y.; Hsueh, A. J. Evolution and Classification of Cystine Knot-Containing Hormones and Related Extracellular Signaling Molecules. *Mol. Endocrinol.* **2001**, *15*, 681−694.

(5) Chang, H.; Brown, C. W.; Matzuk, M. M. Genetic Analysis of the Mammalian Transforming Growth Factor-β Superfamily. *Endocr. Rev.* **2002**, *23*, 787−823.

(6) Galat, A. Common Structural Traits for Cystine Knot Domain of the TGFβ Superfamily of Proteins and Three-Fingered Ectodomain of Their Cellular Receptors. *Cell. Mol. Life Sci.* **2011**, *68*, 3437−3451.

(7) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912−1934.

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

(8) Edson, M. A.; Nagaraja, A. K.; Matzuk, M. M. The Mammalian Ovary From Genesis to Revelation. *Endocr. Rev.* **2009**, *30*, 624−712.
(9) Galat, A.; Gross, G.; Drevet, P.; Sato, A.; Ménez, A. Conserved Structural Determinants in Three-Fingered Protein Domains. *FEBS J.* **2008**, *275*, 3207−3225.
(10) Galat, A. The Three-Fingered Protein Domain of the Human Genome. *Cell. Mol. Life Sci.* **2008**, *65*, 3481−3493.
(11) Levitin, F.; Weiss, M.; Hahn, Y.; Stern, O.; Papke, R. L.; Matusik, R.; Nandana, S. R.; Ziv, R.; Pichinuk, E.; Salame, S.; Bera, T.; Vincent, J.; Lee, B.; Pastan, I.; Wreschner, D. H. PATE Gene Clusters Code for Multiple, Secreted TFP/Ly-6/uPAR Proteins that are Expressed in Reproductive and Neuron-Rich Tissues and Possess Neuromodulatory Activity. *J. Biol. Chem.* **2008**, *283*, 16928−16939.
(12) Tsetlin, V. I. Three-finger Snake Neurotoxins and Ly6 Proteins Targeting Nicotinic Acetylcholine Receptors: Pharmacological Tools and Endogenous Modulators. *Trends Pharmacol. Sci.* **2015**, *36*, 109−123.
(13) Kieffer, B.; Driscoll, P. C.; Campbell, I. D.; Willis, A. C.; van der Merwe, P. A.; Davis, S. J. Three-dimensional Solution Structure of the Extracellular Region of the Complement Regulatory Protein CD59, a New Cell-surface Protein Domain Related to Snake Venom Neurotoxins. *Biochemistry* **1994**, *33*, 4471−4482.
(14) Kong, H. K.; Park, J. H. Characterization and Function of Human Ly-6/uPAR Molecules. *BMB Rep.* **2012**, *45*, 595−603.
(15) Ploug, M.; Ellis, V. Structure-Function Relationships in the Receptor for Urokinase-Type Plasminogen Activator. Comparison to Other Members of the Ly-6 Family and Snake Venom α-Neurotoxins. *FEBS Lett.* **1994**, *349*, 163−168.
(16) Technau, U.; Rudd, S.; Maxwell, P.; Gordon, P. M.; Saina, M.; Grasso, L. C.; Hayward, D. C.; Sensen, C. W.; Saint, R.; Holstein, T. W.; Ball, E. E.; Miller, D. J. Maintenance of Ancestral Complexity and Non-Metazoan Genes in Two Basal Cnidarians. *Trends Genet.* **2005**, *21*, 633−639.
(17) Richards, G. S.; Degnan, B. M. The Dawn of Developmental Signaling in the Metazoa. *Cold Spring Harbor Symp. Quant. Biol.* **2009**, *74*, 81−90.
(18) Watanabe, H.; Schmidt, H. A.; Kuhn, A.; Hoger, S. K.; Kocagoz, Y.; Laumann-Lipp, N.; Ozbek, S.; Holstein, T. W. Nodal Signalling Determines Biradial Asymmetry in. *Nature* **2014**, *515*, 112−115.
(19) Palmer, C. A.; Hollis, D. M.; Watts, R. A.; Houck, L. D.; McCall, M. A.; Gregg, R. G.; Feldhoff, P. W.; Feldhoff, R. C.; Arnold, S. J. Plethodontid Modulating Factor, a Hypervariable Salamander Courtship Pheromone in the Three-Finger Protein Superfamily. *FEBS J.* **2007**, *274*, 2300−2310.
(20) Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Edgar, R.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Miller, V.; Ostell, J.; Pruitt, K. D.; Schuler, G. D.; Shumway, M.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2008**, *36* (Database issue), D13−D21.
(21) Srivastava, M.; Begovic, E.; Chapman, J.; Putnam, N. H.; Hellsten, U.; Kawashima, T.; Kuo, A.; Mitros, T.; Salamov, A.; Carpenter, M. L.; Signorovitch, A. Y.; Moreno, M. A.; Kamm, K.; Grimwood, J.; Schmutz, J.; Shapiro, H.; Grigoriev, I. V.; Buss, L. W.; Schierwater, B.; Dellaporta, S. L.; Rokhsar, D. S. The *Trichoplax* Genome and the Nature of Placozoans. *Nature* **2008**, *454*, 955−960.
(22) Srivastava, M.; Simakov, O.; Chapman, J.; Fahey, B.; Gauthier, M. E.; Mitros, T.; Richards, G. S.; Conaco, C.; Dacre, M.; Hellsten, U.; Larroux, C.; Putnam, N. H.; Stanke, M.; Adamska, M.; Darling, A.; Degnan, S. M.; Oakley, T. H.; Plachetzki, D. C.; Zhai, Y.; Adamski, M.; Calcino, A.; Cummins, S. F.; Goodstein, D. M.; Harris, C.; Jackson, D. J.; Leys, S. P.; Shu, S.; Woodcroft, B. J.; Vervoort, M.; Kosik, K. S.; Manning, G.; Degnan, B. M.; Rokhsar, D. S. The *Amphimedon queenslandica* Genome and the Evolution of Animal Complexity. *Nature* **2010**, *466*, 720−726.
(23) Chapman, J. A.; Kirkness, E. F.; Simakov, O.; Hampson, S. E.; Mitros, T.; Weinmaier, T.; Rattei, T.; Balasubramanian, P. G.; Borman, J.; Busam, D.; Disbennett, K.; Pfannkoch, C.; Sumin, N.; Sutton, G. G.; Viswanathan, L. D.; Walenz, B.; Goodstein, D. M.; Hellsten, U.; Kawashima, T.; Prochnik, S. E.; Putnam, N. H.; Shu, S.; Blumberg, B.; Dana, C. E.; Gee, L.; Kibler, D. F.; Law, L.; Lindgens, D.; Martinez, D. E.; Peng, J.; Wigge, P. A.; Bertulat, B.; Guder, C.; Nakamura, Y.; Ozbek, S.; Watanabe, H.; Khalturin, K.; Hemmrich, G.; Franke, A.; Augustin, R.; Fraune, S.; Hayakawa, E.; Hayakawa, S.; Hirose, M.; Hwang, J. S.; Ikeo, K.; Nishimiya-Fujisawa, C.; Ogura, A.; Takahashi, T.; Steinmetz, P. R.; Zhang, X.; Aufschnaiter, R.; Eder, M. K.; Gorny, A. K.; Salvenmoser, W.; Heimberg, A. M.; Wheeler, B. M.; Peterson, K. J.; Bottger, A.; Tischler, P.; Wolf, A.; Gojobori, T.; Remington, K. A.; Strausberg, R. L.; Venter, J. C.; Technau, U.; Hobmayer, B.; Bosch, T. C.; Holstein, T. W.; Fujisawa, T.; Bode, H. R.; David, C. N.; Rokhsar, D. S.; Steele, R. E. The Dynamic Genome of *Hydra*. *Nature* **2010**, *464*, 592−596.
(24) Putnam, N. H.; Srivastava, M.; Hellsten, U.; Dirks, B.; Chapman, J.; Salamov, A.; Terry, A.; Shapiro, H.; Lindquist, E.; Kapitonov, V. V.; Jurka, J.; Genikhovich, G.; Grigoriev, I. V.; Lucas, S. M.; Steele, R. E.; Finnerty, J. R.; Technau, U.; Martindale, M. Q.; Rokhsar, D. S. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science* **2007**, *317*, 86−94.
(25) Dehal, P.; Satou, Y.; Campbell, R. K.; Chapman, J.; Degnan, B.; De Tomaso, A.; Davidson, B.; Di Gregorio, A.; Gelpke, M.; Goodstein, D. M.; Harafuji, N.; Hastings, K. E.; Ho, I.; Hotta, K.; Huang, W.; Kawashima, T.; Lemaire, P.; Martinez, D.; Meinertzhagen, I. A.; Necula, S.; Nonaka, M.; Putnam, N.; Rash, S.; Saiga, H.; Satake, M.; Terry, A.; Yamada, L.; Wang, H. G.; Awazu, S.; Azumi, K.; Boore, J.; Branno, M.; Chin-Bow, S.; DeSantis, R.; Doyle, S.; Francino, P.; Keys, D. N.; Haga, S.; Hayashi, H.; Hino, K.; Imai, K. S.; Inaba, K.; Kano, S.; Kobayashi, K.; Kobayashi, M.; Lee, B. I.; Makabe, K. W.; Manohar, C.; Matassi, G.; Medina, M.; Mochizuki, Y.; Mount, S.; Morishita, T.; Miura, S.; Nakayama, A.; Nishizaka, S.; Nomoto, H.; Ohta, F.; Oishi, K.; Rigoutsos, I.; Sano, M.; Sasaki, A.; Sasakura, Y.; Shoguchi, E.; Shin-I, T.; Spagnuolo, A.; Stainier, D.; Suzuki, M. M.; Tassy, O.; Takatori, N.; Tokuoka, M.; Yagi, K.; Yoshizaki, F.; Wada, S.; Zhang, C.; Hyatt, P. D.; Larimer, F.; Detter, C.; Doggett, N.; Glavina, T.; Hawkins, T.; Richardson, P.; Lucas, S.; Kohara, Y.; Levine, M.; Satoh, N.; Rokhsar, D. S. The Draft Genome of *Ciona intestinalis*: Insights into Chordate and Vertebrate Origins. *Science* **2002**, *298*, 2157−2167.
(26) Putnam, N. H.; Butts, T.; Ferrier, D. E.; Furlong, R. F.; Hellsten, U.; Kawashima, T.; Robinson-Rechavi, M.; Shoguchi, E.; Terry, A.; Yu, J. K.; Benito-Gutiérrez, E. L.; Dubchak, I.; Garcia-Fernandez, J.; Gibson-Brown, J. J.; Grigoriev, I. V.; Horton, A. C.; de Jong, P. J.; Jurka, J.; Kapitonov, V. V.; Kohara, Y.; Kuroki, Y.; Lindquist, E.; Lucas, S.; Osoegawa, K.; Pennacchio, L. A.; Salamov, A. A.; Satou, Y.; Sauka-Spengler, T.; Schmutz, J.; Shin-I, T.; Toyoda, A.; Bronner-Fraser, M.; Fujiyama, A.; Holland, L. Z.; Holland, P. W.; Satoh, N.; Rokhsar, D. S. The *Amphioxus* Genome and the Evolution of the Chordate Karyotype. *Nature* **2008**, *453*, 1064−1071.
(27) Sodergren, E.; Weinstock, G. M.; Davidson, E. H.; Cameron, R. A.; Gibbs, R. A.; Angerer, R. C.; Angerer, L. M.; Arnone, M. I.; Burgess, D. R.; Burke, R. D.; Coffman, J. A.; Dean, M.; Elphick, M. R.; Ettensohn, C. A.; Foltz, K. R.; Hamdoun, A.; Hynes, R. O.; Klein, W. H.; Marzluff, W.; McClay, D. R.; Morris, R. L.; Mushegian, A.; Rast, J. P.; Smith, L. C.; Thorndyke, M. C.; Vacquier, V. D.; Wessel, G. M.; Wray, G.; Zhang, L.; Elsik, C. G.; Ermolaeva, O.; Hlavina, W.; Hofmann, G.; Kitts, P.; Landrum, M. J.; Mackey, A. J.; Maglott, D.; Panopoulou, G.; Poustka, A. J.; Pruitt, K.; Sapojnikov, V.; Song, X.; Souvorov, A.; Solovyev, V.; Wei, Z.; Whittaker, C. A.; Worley, K.; Durbin, K. J.; Shen, Y.; Fedrigo, O.; Garfield, D.; Haygood, R.; Primus, A.; Satija, R.; Severson, T.; Gonzalez-Garay, M. L.; Jackson, A. R.; Milosavljevic, A.; Tong, M.; Killian, C. E.; Livingston, B. T.; Wilt, F. H.; Adams, N.; Bellé, R.; Carbonneau, S.; Cheung, R.; Cormier, P.; Cosson, B.; Croce, J.; Fernandez-Guerra, A.; Geneviève, A. M.; Goel, M.; Kelkar, H.; Morales, J.; Mulner-Lorillon, O.; Robertson, A. J.; Goldstone, J. V.; Cole, B.; Epel, D.; Gold, B.; Hahn, M. E.; Howard-Ashby, M.; Scally, M.; Stegeman, J. J.; Allgood, E. L.; Cool, J.; Judkins,

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

K. M.; McCafferty, S. S.; Musante, A. M.; Obar, R. A.; Rawson, A. P.; Rossetti, B. J.; Gibbons, I. R.; Hoffman, M. P.; Leone, A.; Istrail, S.; Materna, S. C.; Samanta, M. P.; Stolc, V.; Tongprasit, W.; Tu, Q.; Bergeron, K. F.; Brandhorst, B. P.; Whittle, J.; Berney, K.; Bottjer, D. J.; Calestani, C.; Peterson, K.; Chow, E.; Yuan, Q. A.; Elhaik, E.; Graur, D.; Reese, J. T.; Bosdet, I.; Heesun, S.; Marra, M. A.; Schein, J.; Anderson, M. K.; Brockton, V.; Buckley, K. M.; Cohen, A. H.; Fugmann, S. D.; Hibino, T.; Loza-Coll, M.; Majeske, A. J.; Messier, C.; Nair, S. V.; Pancer, Z.; Terwilliger, D. P.; Agca, C.; Arboleda, E.; Chen, N.; Churcher, A. M.; Hallbook, F.; Humphrey, G. W.; Idris, M. M.; Kiyama, T.; Liang, S.; Mellott, D.; Mu, X.; Murray, G.; Olinski, R. P.; Raible, F.; Rowe, M.; Taylor, J. S.; Tessmar-Raible, K.; Wang, D.; Wilson, K. H.; Yaguchi, S.; Gaasterland, T.; Galindo, B. E.; Gunaratne, H. J.; Juliano, C.; Kinukawa, M.; Moy, G. W.; Neill, A. T.; Nomura, M.; Raisch, M.; Reade, A.; Roux, M. M.; Song, J. L.; Su, Y. H.; Townley, I. K.; Voronina, E.; Wong, J. L.; Amore, G.; Branno, M.; Brown, E. R.; Cavalieri, V.; Duboc, V.; Duloquin, L.; Flytzanis, C.; Gache, C.; Lapraz, F.; Lepage, T.; Locascio, A.; Martinez, P.; Matassi, G.; Matranga, V.; Range, R.; Rizzo, F.; Rottinger, E.; Beane, W.; Bradham, C.; Byrum, C.; Glenn, T.; Hussain, S.; Manning, G.; Miranda, E.; Thomason, R.; Walton, K.; Wikramanayke, A.; Wu, S. Y.; Xu, R.; Brown, C. T.; Chen, L.; Gray, R. F.; Lee, P. Y.; Nam, J.; Oliveri, P.; Smith, J.; Muzny, D.; Bell, S.; Chacko, J.; Cree, A.; Curry, S.; Davis, C.; Dinh, H.; Dugan-Rocha, S.; Fowler, J.; Gill, R.; Hamilton, C.; Hernandez, J.; Hines, S.; Hume, J.; Jackson, L.; Jolivet, A.; Kovar, C.; Lee, S.; Lewis, L.; Miner, G.; Morgan, M.; Nazareth, L. V.; Okwuonu, G.; Parker, D.; Pu, L. L.; Thorn, R.; Wright, R. The Genome of the Sea Urchin *Strongylocentrotus purpuratus*. *Science* **2006**, *314*, 941−952.

(28) Freeman, R. M., Jr.; Wu, M.; Cordonnier-Pratt, M. M.; Pratt, L. H.; Gruber, C. E.; Smith, M.; Lander, E. S.; Stange-Thomann, N.; Lowe, C. J.; Gerhart, J.; Kirschner, M. cDNA Sequences for Transcription Factors and Signaling Proteins of the Hemichordate *Saccoglossus kowalevskii*: Efficacy of the Expressed Sequence Tag (EST) Approach for Evolutionary and Developmental Studies of a New Organism. *Biol. Bull.* **2008**, *214*, 284−302.

(29) Elsik, C. G.; Worley, K. C.; Bennett, A. K.; Beye, M.; Camara, F.; Childers, C. P.; de Graaf, D. C.; Debyser, G.; Deng, J.; Devreese, B.; Elhaik, E.; Evans, J. D.; Foster, L. J.; Graur, D.; Guigo, R.; HGSC production teams; Hoff, K. J.; Holder, M. E.; Hudson, M. E.; Hunt, G. J.; Jiang, H.; Joshi, V.; Khetani, R. S.; Kosarev, P.; Kovar, C. L.; Ma, J.; Maleszka, R.; Moritz, R. F.; Munoz-Torres, M. C.; Murphy, T. D.; Muzny, D. M.; Newsham, I. F.; Reese, J. T.; Robertson, H. M.; Robinson, G. E.; Rueppell, O.; Solovyev, V.; Stanke, M.; Stolle, E.; Tsuruda, J. M.; Vaerenbergh, M. V.; Waterhouse, R. M.; Weaver, D. B.; Whitfield, C. W.; Wu, Y.; Zdobnov, E. M.; Zhang, L.; Zhu, D.; Gibbs, R. A. Honey Bee Genome Sequencing Consortium. Finding the Missing Honey Bee Genes: Lessons Learned From a Genome Upgrade. *BMC Genomics* **2014**, *15*, 86 http://www.biomedcentral.com/1471-2164/15/86,.

(30) Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y. J. International Human Genome Sequencing Consortium. Initial Sequencing and Analysis of the Human Genome. *Nature* **2001**, *409*, 860−921.

(31) Altschul, S. F.; Wootton, J. C.; Gertz, E. M.; Agarwala, R.; Morgulis, A.; Schäffer, A. A.; Yu, Y. K. Protein Database Searches Using Compositionally Adjusted Substitution Matrices. *FEBS J.* **2005**, *272*, 5101−5109.

(32) Tweedie, S.; Ashburner, M.; Falls, K.; Leyland, P.; McQuilton, P.; Marygold, S.; Millburn, G.; Osumi-Sutherland, D.; Schroeder, A.; Seal, R.; Zhang, H. The FlyBase Consortium. FlyBase: Enhancing *Drosophila* Gene Ontology Annotations. *Nucleic Acids Res.* **2009**, *37*, D555−D559.

(33) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947−2948.

(34) Huson, D. H.; Richter, D. C.; Rausch, C.; Dezulian, T.; Franz, M.; Rupp, R. Dendroscope: An Interactive Viewer for Large Phylogenetic Trees. *BMC Bioinf.* **2007**, *8*, 460.

(35) Galat, A. Function-Dependent Clustering of Orthologues and Paralogues of Cyclophilins. *Proteins: Struct., Funct., Genet.* **2004**, *56*, 808−820.

(36) Galat, A. Involvement of Some Large Immunophilins and Their Ligands in the Protection and Regeneration of Neurons: a Hypothetical Mode of Action. *Comput. Biol. Chem.* **2006**, *30*, 348−359.

(37) Phillips, J. C. Hydropathic Self-Organized Criticality: a Magic Wand for Protein Physics. *Protein Pept. Lett.* **2012**, *19*, 1089−1093.

(38) Phillips, J. C. Thermodynamic Description of β Amyloid Formation Using Physicochemical Scales and Fractal Bioinformatic Scales. *ACS Chem. Neurosci.* **2015**, *6*, 745−750.

(39) Galat, A. On Transversal Hydrophobicity of Some Proteins and Their Modules. *J. Chem. Inf. Model.* **2009**, *49*, 1821−1830.

(40) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157*, 105−132.

(41) Arndt, C. *Information Measures: Information and Its Description in Science and Engineering*; Springer: Berlin, 2004.

(42) Galat, A. Functional Drift of Sequence Attributes in the FK506-Binding Proteins (FKBPs). *J. Chem. Inf. Model.* **2008**, *48*, 1118−1130.

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

(43) Berman, H. M.; Henrick, K.; Nakamura, H.; Markley, J. L. The Worldwide Protein Data Bank (wwPDB): Ensuring a Single, Uniform Archive of PDB Data. *Nucleic Acids Res.* **2007**, *35*, D301−D303.
(44) Roskoski, R., Jr. ERK1/2 MAP Kinases: Structure, Function, and Regulation. *Pharmacol. Res.* **2012**, *66*, 105−143.
(45) Sawyer, J. S.; Beight, D. W.; Britt, K. S.; Anderson, B. D.; Campbell, R. M.; Goodson, T.; Jr; Herron, D. K.; Li, H. Y.; McMillen, W. T.; Mort, N.; Parsons, S.; Smith, E. C.; Wagner, J. R.; Yan, L.; Zhang, F.; Yingling, J. M. Synthesis and Activity of New Aryl- and Heteroaryl-Substituted 5,6-Dihydro-4H-pyrrolo[1,2-b]pyrazole Inhibitors of the Transforming Growth Factor-β Type I Receptor Kinase Domain. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3581−3584.
(46) Martin, M. P.; Zhu, J. Y.; Lawrence, H. R.; Pireddu, R.; Luo, Y.; Alam, R.; Ozcan, S.; Sebti, S. M.; Lawrence, N. J.; Schonbrunn, E. A Novel Mechanism by which Small Molecule Inhibitors Induce the DFG Flip in Aurora A. *ACS Chem. Biol.* **2012**, *7*, 698−706.
(47) Affolter, M.; Marty, T.; Vigano, M. A.; Jazwinska, A. Nuclear Interpretation of Dpp Signaling in *Drosophila*. *EMBO J.* **2001**, *20*, 3298−3305.
(48) Osipov, A. V.; Rucktooa, P.; Kasheverov, I. E.; Filkin, S. Y.; Starkov, V. G.; Andreeva, T. V.; Sixma, T. K.; Bertrand, D.; Utkin, Y. N.; Tsetlin, V. I. Dimeric α−Cobratoxin X-ray Structure: Localization of Intermolecular Disulfides and Possible Mode of Binding to Nicotinic Acetylcholine Receptors. *J. Biol. Chem.* **2012**, *287*, 6725−6734.
(49) Pawlak, J.; Mackessy, S. P.; Sixberry, N. M.; Stura, E. A.; Le Du, M. H.; Menez, R.; Foo, C. S.; Menez, A.; Nirthanan, S.; Kini, R. M. Irditoxin, a Novel Covalently Linked Heterodimeric Three-Finger Toxin with High Taxon-Specific Neurotoxicity. *FASEB J.* **2009**, *23*, 534−545.
(50) Hijazi, A.; Masson, W.; Augé, B.; Waltzer, L.; Haenlin, M.; Roch, F. Boudin is Required for Septate Junction Organisation in *Drosophila* and Codes for a Diffusible Protein of the Ly-6 Superfamily. *Development* **2009**, *136*, 2199−2209.
(51) Moussian, B.; Soding, J.; Schwarz, H.; Nusslein-Volhard, C. Retroactive, a Membrane-Anchored Extracellular Protein Related to Vertebrate Snake Neurotoxin-Like Proteins, is Required for Cuticle Organization in the Lrva of *Drosophila melanogaster*. *Dev. Dyn.* **2005**, *233*, 1056−1063.
(52) Nilton, A.; Oshima, K.; Zare, F.; Byri, S.; Nannmark, U.; Nyberg, K. G.; Fehon, R. G.; Uv, A. E. Crooked, Coiled and Crimpled are Three Ly6-Like Proteins Required for Proper Localization of Septate Junction Components. *Development* **2010**, *137*, 2427−2437.
(53) Wu, M.; Robinson, J. E.; Joiner, W. J. Sleepless is a Bifunctional Regulator of Excitability and Cholinergic Synaptic Transmission. *Curr. Biol.* **2014**, *24*, 621−629.
(54) Wu, M. N.; Joiner, W. J.; Dean, T.; Yue, Z.; Smith, C. J.; Chen, D.; Hoshi, T.; Sehgal, A.; Koh, K. Sleepless, a Ly-6/Neurotoxin Family Member, Regulates the Levels, Localization and Activity of Shaker. *Nat. Neurosci.* **2010**, *13*, 69−75.
(55) Koh, K.; Joiner, W. J.; Wu, M. N.; Yue, Z.; Smith, C. J.; Sehgal, A. Identification of Sleepless, a Sleep-Promoting Factor. *Science* **2008**, *321*, 372−376.
(56) Klein, D. E.; Stayrook, S. E.; Shi, F.; Narayan, K.; Lemmon, M. A. Structural Basis for EGFR Ligand Sequestration by Argos. *Nature* **2008**, *453*, 1271−1275.
(57) Fry, B. G.; Scheib, H.; van der Weerd, L.; Young, B.; McNaughtan, J.; Ramjan, S. F.; Vidal, N.; Poelmann, R. E.; Norman, J. A. Evolution of an Arsenal: Structural and Functional Diversification of the Venom System in the Advanced Snakes (*Caenophidia*). *Mol. Cell. Proteomics* **2008**, *7*, 215−246.
(58) Fry, B. G.; Roelants, K.; Champagne, D. E.; Scheib, H.; Tyndall, J. D.; King, G. F.; Nevalainen, T. J.; Norman, J. A.; Lewis, R. J.; Norton, R. S.; Renjifo, C.; de la Vega, R. C. The Toxicogenomic Multiverse: Convergent Recruitment of Proteins into Animal Venoms. *Annu. Rev. Genomics Hum. Genet.* **2009**, *10*, 483−511.
(59) Lyukmanova, E. N.; Shenkarev, Z. O.; Shulepko, M. A.; Mineev, K. S.; D'Hoedt, D.; Kasheverov, I. E.; Filkin, S. Y.; Krivolapova, A. P.; Janickova, H.; Dolezal, V.; Dolgikh, D. A.; Arseniev, A. S.; Bertrand, D.; Tsetlin, V. I.; Kirpichnikov, M. P. NMR Structure and Action on Nicotinic Acetylcholine Receptors of Water-Soluble Domain of Human LYNX1. *J. Biol. Chem.* **2011**, *286*, 10618−10627.
(60) Kuhn, P.; Deacon, A. M.; Comoso, S.; Rajaseger, G.; Kini, R. M.; Uson, I.; Kolatkar, P. R. The Atomic Resolution Structure of Bucandin, a Novel Toxin Isolated from the Malayan Krait, Determined by Direct Methods. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2000**, *56*, 1401−1407.
(61) Janssenswillen, S.; Vandebergh, W.; Treer, D.; Willaert, B.; Maex, M.; Van Bocxlaer, I.; Bossuyt, F. Origin and Diversification of a Salamander Sex Pheromone System. *Mol. Biol. Evol.* **2015**, *32*, 472−480.
(62) Wilburn, D. B.; Bowen, K. E.; Doty, K. A.; Arumugam, S.; Lane, A. N.; Feldhoff, P. W.; Feldhoff, R. C. Structural Insights into the Evolution of a Sexy Protein: Novel Topology and Restricted Backbone Flexibility in a Hypervariable Pheromone from the Red-Legged Salamander, *Plethodon shermani*. *PLoS One* **2014**, *9*, e96975.
(63) Gilquin, B.; Bourgoin, M.; Ménez, R.; Le Du, M. H.; Servent, D.; Zinn-Justin, S.; Ménez, A. Motions and Structural Variability Within Toxins: Implication for Their Use as Scaffolds for Protein Engineering. *Protein Sci.* **2003**, *12*, 266−277.
(64) Vonk, F. J.; Casewell, N. R.; Henkel, C. V.; Heimberg, A. M.; Jansen, H. J.; McCleary, R. J.; Kerkkamp, H. M.; Vos, R. A.; Guerreiro, I.; Calvete, J. J.; Wuster, W.; Woods, A. E.; Logan, J. M.; Harrison, R. A.; Castoe, T. A.; de Koning, A. P.; Pollock, D. D.; Yandell, M.; Calderon, D.; Renjifo, C.; Currier, R. B.; Salgado, D.; Pla, D.; Sanz, L.; Hyder, A. S.; Ribeiro, J. M.; Arntzen, J. W.; van den Thillart, G. E.; Boetzer, M.; Pirovano, W.; Dirks, R. P.; Spaink, H. P.; Duboule, D.; McGlinn, E.; Kini, R. M.; Richardson, M. K. The King Cobra Genome Reveals Dynamic Gene Evolution and Adaptation in the Snake Venom System. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 20651−20656.
(65) Lazdunski, M.; Renaud, J. F. The Action of Cardiotoxins on Cardiac Plasma Membranes. *Annu. Rev. Physiol.* **1982**, *44*, 463−473.
(66) Sue, S. C.; Brisson, J. R.; Chang, S. C.; Huang, W. N.; Lee, S. C.; Jarrell, H. C.; Wu, W. Structures of Heparin-Derived Disaccharide Bound to Cobra Cardiotoxins: Context-Dependent Conformational Change of Heparin Upon Binding to the Rigid Core of the Three-Fingered Toxin. *Biochemistry* **2001**, *40*, 10436−10446.
(67) Dubovskii, P. V.; Utkin, Y. N. Cobra Cytotoxins: Structural Organization and Antibacterial Activity. *Acta Naturae* **2014**, *6*, 11−18.
(68) Nirthanan, S.; Gwee, M. C. Three-Finger α-Neurotoxins and the Nicotinic Acetylcholine Receptor, Forty Years On. *J. Pharmacol. Sci.* **2004**, *94*, 1−17.
(69) Kini, R. M.; Doley, R. Structure, Function and Evolution of Three-Finger Toxins: Mini Proteins with Multiple Targets. *Toxicon* **2010**, *56*, 855−867.
(70) Vonk, F. J.; Jackson, K.; Doley, R.; Madaras, F.; Mirtschin, P. J.; Vidal, N. Snake Venom: From Fieldwork to the Clinic: Recent Insights into Snake Biology, Together with New Technology Allowing High-Throughput Screening of Venom, Bring New Hope for Drug Discovery. *BioEssays* **2011**, *33*, 269−279.
(71) Levitt, M.; Gerstein, M.; Huang, E.; Subbiah, S.; Tsai, J. Protein Folding: the Endgame. *Annu. Rev. Biochem.* **1997**, *66*, 549−579.
(72) Ferreiro, D. U.; Komives, E. A.; Wolynes, P. G. Frustration in Biomolecules. *Q. Rev. Biophys.* **2014**, *47*, 285−363.
(73) Soria, P. S.; McGary, K. L.; Rokas, A. Functional Divergence for Every Paralog. *Mol. Biol. Evol.* **2014**, *31*, 984−992.
(74) Tanaka, K.; Diekmann, Y.; Hazbun, A.; Hijazi, A.; Vreede, B.; Roch, F.; Sucena, E. Multi-Species Analysis of Expression Pattern Diversification in the Recently Expanded Insect Ly6 Gene Family. *Mol. Biol. Evol.* **2015**, *32*, 1730−1747.
(75) Klippel, S.; Strunck, E.; Busse, C. E.; Behringer, D.; Pahl, H. L. Biochemical Characterization of PRV-1, a Novel Hematopoietic Cell Surface Receptor, which is Overexpressed in Polycythemia Rubra Vera. *Blood* **2002**, *100*, 2441−2448.
(76) Fujihara, Y.; Tokuhiro, K.; Muro, Y.; Kondoh, G.; Araki, Y.; Ikawa, M.; Okabe, M. Expression of TEX101, Regulated by ACE, is Essential for the Production of Fertile Mouse Spermatozoa. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 8111−8116.

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041

(77) Ramachandran, P.; Pellicoro, A.; Vernon, M. A.; Boulter, L.; Aucott, R. L.; Ali, A.; Hartland, S. N.; Snowdon, V. K.; Cappon, A.; Gordon-Walker, T. T.; Williams, M. J.; Dunbar, D. R.; Manning, J. R.; van Rooijen, N.; Fallowfield, J. A.; Forbes, S. J.; Iredale, J. P. Differential Ly-6C Expression Identifies the Recruited Macrophage Phenotype, which Orchestrates the Regression of Murine Liver Fibrosis. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E3186−E3195.
(78) Ben-Shlomo, I.; Hsueh, A. J. Three's Company: Two or More Unrelated Receptors Pair with the Same Ligand. *Mol. Endocrinol.* **2005**, *19*, 1097−1109.
(79) Rosas, M.; Thomas, B.; Stacey, M.; Gordon, S.; Taylor, P. R. The Myeloid 7/4-Antigen Defines Recently Generated Inflammatory Macrophages and is Synonymous with Ly-6B. *J. Leukocyte Biol.* **2010**, *88*, 169−180.
(80) Lee, P. Y.; Wang, J. X.; Parisini, E.; Dascher, C. C.; Nigrovic, P. A. Ly6 Family Proteins in Neutrophil Biology. *J. Leukocyte Biol.* **2013**, *94*, 585−594.
(81) Brockes, J. P.; Kumar, A. Comparative Aspects of Animal Regeneration. *Annu. Rev. Cell Dev. Biol.* **2008**, *24*, 525−549.
(82) Lévesque, M.; Gatien, S.; Finnson, K.; Desmeules, S.; Villiard, E.; Pilote, M.; Philip, A.; Roy, S. Transforming Growth Factor: β Signaling is Essential for Limb Regeneration in Axolotls. *PLoS One* **2007**, *2*, e1227.
(83) Walsh, D. W.; Godson, C.; Brazil, D. P.; Martin, F. Extracellular BMP-Antagonist Regulation in Development and Disease: Tied up in Knots. *Trends Cell Biol.* **2010**, *20*, 244−256.

DOI: 10.1021/acs.jcim.5b00322
J. Chem. Inf. Model. 2015, 55, 2026−2041