Phylogenetic, sequence and structural analysis of Insulin superfamily proteins reveals an indelible link between evolution and structure-function relationship

Shrilakshmi S.[1,2#], Shankar V. Kundapura[1,2#], Debayan Dey[1, 3], Ananda Kulal[1*], Udupi A. Ramagopal[1*]

[1]Division of biological sciences, Poornaprajna Institute of Scientific Research, Poornaprajnapura, Bidalur (Post), Bengaluru, India.

[2]Manipal Academy of Higher Education, Manipal, Karnataka, India-576104.

[3]Present address: Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA.

# Equal contribution

***To whom correspondence may be addressed**:

Udupi A. Ramagopal, Division of Biological Sciences, Poornaprajna Institute of Scientific Research, #4, 16th Cross, Sadashivanagar, Bangalore 560080, India
Ph.: +91 9900810182; 918023611836; E-mail: ramagopal.udupi@gmail.com
Ananda, Kulal , Division of Biological Sciences, Poornaprajna Institute of Scientific Research, #4, 16th Cross, Sadashivanagar, Bangalore 560080, India
Ph.: +91 9945900336 E-mail:annuk9@yahoo.co.in.

Keywords: Insulin, IGFs, relaxin, bombyxin, insulin superfamily proteins, phylogenetic tree, partners imposed evolutionary restraints, IGFBPs.

Author contributions: S.S, S.V.K, D.D, A.K, U.A.R. together conceptualized the problem. S.S. and S.V.K performed the analysis. S.S, S.V.K, A.K, U.A.R together evaluated the analysis and wrote the manuscript.

Conflict of interest statement: The authors declare no conflict of interest.

Abstract:

The insulin superfamily proteins (ISPs), in particular, insulin, IGFs and relaxins are key modulators of animal physiology. They are known to have evolved from the same ancestral gene and have diverged into proteins with varied sequences and distinct functions, but maintain a similar structural architecture stabilized by highly conserved disulphide bridges. A recent surge of sequence data and the structures of these proteins prompted a need for a comprehensive analysis which connects the evolution of these sequences in the light of available functional and structural information and their interaction with cognate receptors. This study reveals a) unusually high sequence conservation of IGFs (>90%), which has never been reported before. In fact, it was interesting to observe that the functional domains (excluding signal peptide) of human, horse, pig and Ord's kangaroo rat are 100% identical. (b) an updated definition of the signature motif of the relaxin family (c) a non-canonical C-peptide cleavage site in a few killifish insulin sequences and so on. We also provide a structure-based rationale for such conservation by introducing a concept called binding partners imposed evolutionary constraints. Furthermore, the high conservation of IGFs appears to represent a classic case of resistance to sequence diversity exerted by physiologically important interactions with multiple partners. Furthermore, we propose a probable mechanism for C-peptide cleavage in killifish insulin sequences.

Introduction

Insulin superfamily proteins (ISPs) are involved in critical functions and few of which are extensively studied considering their therapeutic applications. The insulin superfamily includes many homologous proteins such as Insulin (1), IGF-1 (insulin-like growth factor-1) (2), IGF-2 (3), Relaxin (Relaxin 1 2 &3) (4), Insulin-like peptides (INSL3, INSL4, INSL5, INSL6, ILP-1 and ILP-2)(5,6), Bombyxin (7), Locust insulin related peptides (8), Molluscan insulin- related peptides (9) and Caenorhabditis elegans insulin-like peptides (10), which are found in most of the animal phyla. ISPs share similar structural architecture with a distinct insulin fold formed by three conserved disulphide linkages. All the insulin superfamily proteins and peptides are proposed to be evolved from a single ancestral gene after undergoing conducive multiple independent gene duplications (11).

Insulin is the most studied member of this family. Considering its importance in glucose metabolism and diabetes, it is probably most explored or a protein experimentally "beaten to death". Since, ISPs have evolved from a single ancestral gene, we asked the question whether the collective analysis of these members with respect to their sequence divergence/similarity, structure, function and their known interaction with cognate partner(s) reveal any novel insights? With the above question in mind, we have analyzed insulin, insulin-like growth factors (IGFs), bombyxin and relaxin. Among the four members of ISPs considered here, three of these proteins are found in humans (insulin, IGFs and Relaxins) whereas, Bombyxin is an insect protein. Since this study was aimed at understanding the evolution of ISP members in relation to their structure and function, we chose the above mentioned four families of ISPs as these members are relatively well explored, either biochemically and/or structural information on these molecules individually or in complex with their receptors are available. Although, in the last few decades many structure, evolution and functional studies (12–15) have been performed, to our knowledge, there is no study that involves sequence, structure, function and phylogenetic perspective encompassing all the four major families of insulin superfamily. With the surge of sequence information in the post genomics era and recently available structures of several of these ISP members (16–20) in few cases their complexes with receptors, we were curious to know, whether any novel insights could be derived from the cumulative analysis of these evolutionarily related but functionally diverse molecules having similar structural architecture.

As mentioned above, insulin is the well-known member of this family. In brief, this miniature protein is secreted by the pancreatic β cells. The main function of insulin is to maintain optimal blood glucose levels by facilitating glucose uptake by cells, which is mediated by interaction of insulin with its receptor (IR). It also plays a vital role in gene regulation and metabolism of proteins, carbohydrates and lipids (21).

Insulin-like growth factors (IGF-1 and IGF-2) have important roles in regulating the growth and metabolism of all cells and are primarily produced by the liver as a endocrine hormone (2). Unlike insulin, IGFs retain their C-peptide (also called C-domain) and are single polypeptide

chains which maintains the insulin fold. IGF-1 and IGF-2 have similar functional roles except that they are expressed at different stages of physiological development. IGF-1 is expressed mainly after birth whereas IGF-2 is expressed during early embryonic and fetal development in most of the somatic cells (22).

Relaxins are a family of proteins in insulin superfamily having many paralogs among them like relaxin-1, relaxin-2, relaxin-3 and so on (23). Relaxin-2 is known to mediate the hemodynamic changes that occur during pregnancy and is pivotal in mammals which relaxes pelvic ligaments, hence the name relaxin (24). Besides its role in pregnancy, Relaxin-2 also plays a key role in vasodilation, antifibrinolysis, wound healing and is recently found to be involved in cardiovascular regulation against acute heart failure (25–27). Relaxin-1 is considered to be a homolog of relaxin-2 in primates (5). In humans, relaxin-2 is a major circulating hormone (28) and is expressed in variety of tissues whereas, expression of relaxin-1 is restricted to decidua, placenta and prostrate (5). Relaxin-3 is a neuropeptide hormone found in the brain which is said to have a role in learning, stress, alertness, memory, locomotion and appetite regulation (16, 26). It is also claimed to act as a neurotransmitter (29). Relaxin-3 is widely accepted to be the ancestral gene in relaxin family (5).

Bombyxin is largely found in the lepidopteron class of insects. Bombyxin, previously known as prothoracicotropic hormone (PTTH) (7) stimulates the prothoracic glands to synthesize and release ecdysone, which is a key hormone that regulates insect moulting and metamorphosis (30). It also has a role in trehalose sugar metabolism (31). Like relaxins, bombyxins are known to have many paralogs, although the functions of most paralogs have not been determined experimentally.

Considering the fact, that ISPs have evolved from a single ancestral gene, it is possible that the evolutionary conservation and variations of residues among ISPs, could provide critical clues for rational modification of insulin. The study also attempts to correlate the mode of interaction of ISPs to their respective receptors, the determinants responsible for oligomerization observed in few members of ISPs and the consequent evolutionary restraints. During the course of this analysis we came across few interesting aspects such as (a) unusually high sequence conservation in IGFs for which we attempt to provide a rationale. (b) evolutionarily diverged insulin killifish sequences, with a distinct non-canonical C peptide cleavage site (c) an unnoticed conservation in the receptor binding motif of relaxins, particularly in relaxin-3 and so on. Finally, we have borrowed information from the ISPs to design few plausible insulin analogues, computationally analyzed their structural stability and their affinity to insulin receptor.

Results and discussion

Insulin is synthesized as a single polypeptide consisting of (a) a N-terminal stretch called B-chain consisting of around 30 residues (b) a post-translationally cleavable stretch consisting around 30 residues called C-peptide and (c) a C-terminal stretch consisting of around 21 residues called A chain (Figure S1) (32). This naming convention is common to all the members of this family, even for the members that do not require post-translational cleavage of C-peptide. We maintain similar convention all through the manuscript. Most of our analysis considers variation and similarities in the B and A chain unless otherwise mentioned.

Insulin superfamily proteins (ISPs) have varying sequence identity amongst themselves (Table S1), yet have a similar structural architecture consisting of three uniquely arranged helices forming the "insulin fold" (12), stabilized by completely conserved three disulfide bridges (Figure S2). For example, in humans, insulin shares a sequence identity of ~30% with IGF-1 and its sequence identity with relaxin-2 is as low as 14%. All the three members belong to functionally divergent family (Table S1), suggesting, in ISPs a sequence identity of below 30% probably represent two sequences of divergent function. Similarly, when the two IGF paralogs having similar function (both are growth hormones) were compared they had a sequence identity 63% between them. Further, within the relaxin family, Relaxin-1 shares a sequence identity of ~80% with its functionally equivalent member Relaxin-2 (both play a role in pregnancy and vasodilation), whereas, its sequence identity with functionally distinct paralog, a neuropeptide, relaxin-3 is as low as 23%. The above observations suggest that the sequence similarities/differences, although indicative of their function, are not useful to distinguish ISP members into different families.

We have used minimum evolution as the phylogenetic method to understand the relationship between the proteins of the insulin superfamily. When the insulin superfamily phylogenetic tree (Figure 1) was generated, it clearly diverged into four families (Insulin, IGFs, bombyxin and relaxin). The sequences obtained from the clades of these families were considered to recognize position specific conservation and variation within a given family and amongst the families.

The phylogenetic tree (Figure 1) suggests that there was an early divergence of relaxin family from other ISPs. The next family to diverge was bombyxin which consists of sequences largely from *Bombyx mori*. A small clade diverges out before insulin and IGFs which will be discussed later in section on C-peptide cleavage site of killifish insulins. A common ancestor of insulin and IGFs is observed after which insulin and IGFs diverge out together. Further, two clades diverge from IGFs in to IGF-1 and IGF-2. The phylogenetic analysis indicates that the insulin superfamily proteins are closely related to each other with an exception of relaxin which is distant and diverges out early. It is also noteworthy to indicate that, each family has undergone gene duplication events amongst themselves and thereby have different paralogs in the same family.

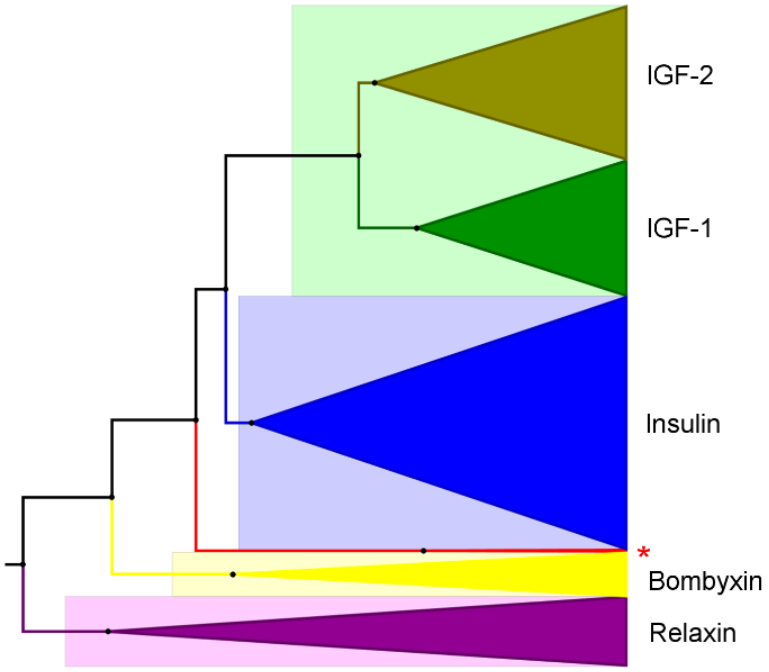

Figure 1: Phylogenetic tree of ISPs constructed using minimum evolution and visualized by graphical tool FigTree. Green clade represents IGF-1, dull green clade represents IGF-2, blue clade represents insulin family, yellow clade represents bombyxin family and purple clade represents relaxin family. The star represents a separate clade of divergent killifish insulin sequences.

The sequences clustered based on phylogenetic analysis were analyzed for conservation within each family and among all the ISP members. The sequence logo generated for each family and the common sequence logo for all the ISP members called "master sequence" is represented in the figure 2. It is clear that B8Gly and the cysteines involved in the formation of canonical disulphide linkages are absolutely conserved throughout the family (Figure 2, black asterisks). In insulin, the B chain has a higher conservation than A chain, which is consistent with the previous report (14). Insulin is the only protein of this family which is stored as a hexamer in its stable form. Hexamerization is stabilized by few hydrophobic interactions between the insulin dimers together with zinc mediated trimerization of insulin dimers, where B10His of every monomer coordinates with the zinc ion to form a hexamer (32). Analysis of the structures of insulin and its complex with its receptor reveal that the highly conserved residues, other than cysteines (see Figure 2) participate in either oligomerization and/or are involved in receptor binding.

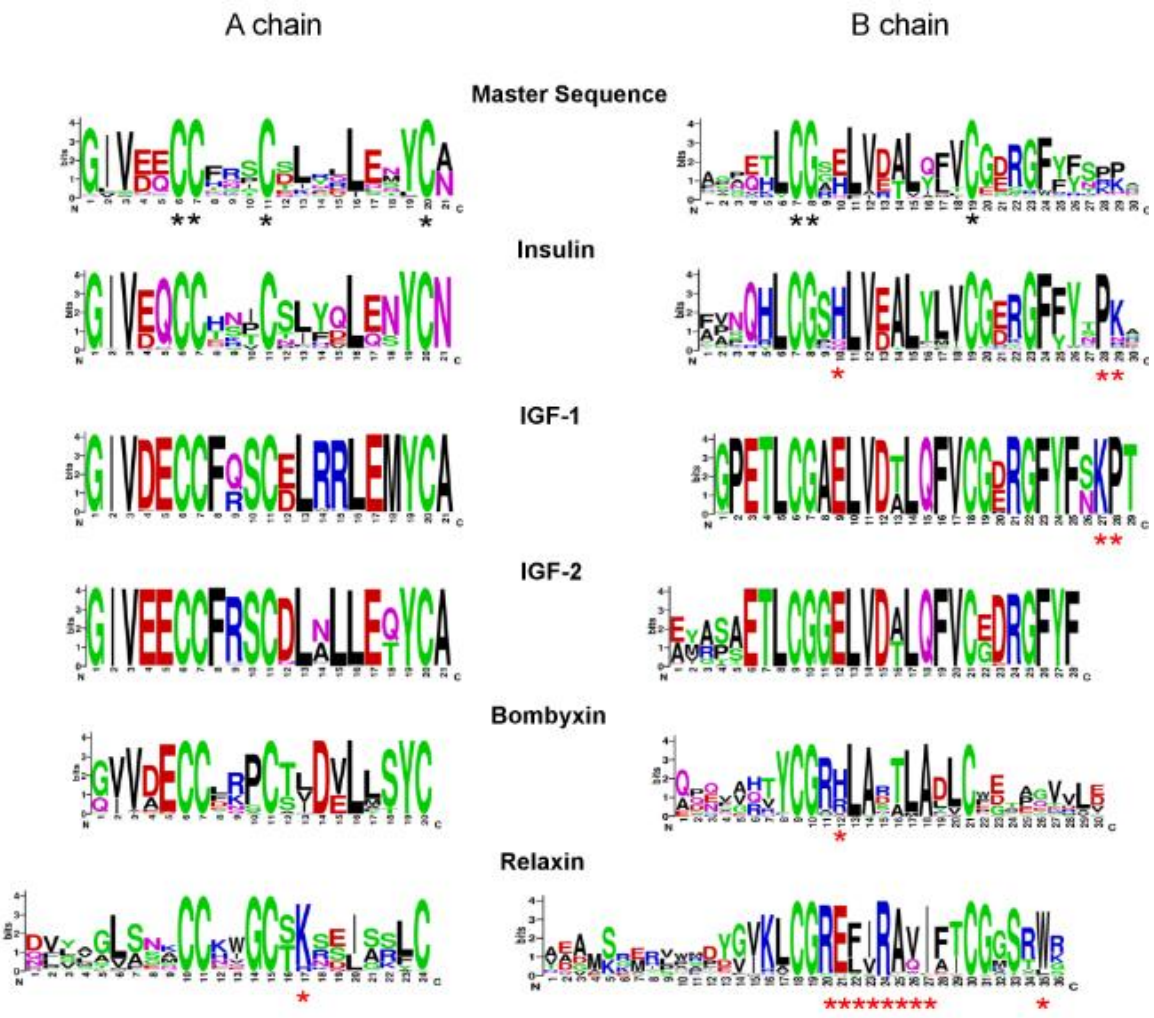

Figure 2: Sequence logos of A and B chains of insulin superfamily. Master alignment sequence logo represents the whole data set whereas; the individual sequence logo of proteins represents their respective cluster (family). The conservation of the residues corresponds to the height of the letter and the sequence logo was generated using the program WebLogo. For convenience, sequence logos of individual protein families are stacked upon each other in such a manner that the conserved canonical cysteines are in the same position. The asterisks in black below the master sequence represents the completely conserved residues in the ISPs. The asterisks in red represent the residues will be discussed in the following sections.

IGFs (IGF-1 and IGF-2) are monomers and they bind to their respective cognate receptors (IGF-1R and IGF-2R) and have low affinity cross reactivity with insulin receptors (IR). From figure 2, it is clear that IGF-1 and IGF-2, despite being monomers, have the most number of completely conserved residues in both A and B chains among ISP members (except for the N-terminal part of IGF-2). This unusual conservation of IGFs, to our knowledge has not been noticed before. The rationale for such high sequence conservation in IGFs will be discussed later.

Bombyxin shares around 30% sequence identity with human insulin (Table S1). While the insulin B chain is relatively more conserved than the A chain, our analysis suggests that the

bombyxin A chain is more conserved compared to its B chain (Figure 2). Concurrent with this observation, it has been shown that the modification of the A chain reduces the affinity to the bombyxin receptor (33). Unfortunately, the experimental evidence on bombyxin receptor and its interactions are elusive. A report on a probable receptor suggests that there is a 300kDa protein present on the ovarian cells of lepidopteran species (34). It was proposed that bombyxin doesn't form a hexamer (35, 36) though it contains the B10His (64%), which in insulin plays a vital role in zinc mediated hexamerization.

Relaxins are the first family to diverge according to the phylogenetic tree and have the least sequence conservation when compared to the other families (Figure 2). A plausible reason could be the presence of functionally variant paralogs like relaxin-1, 2 and 3 as mentioned earlier. Although, functionally divergent, relaxin sequences have known to possess a distinct receptor binding motif (RXXXRXXI/V), which is proposed to be important for its interaction with G-protein coupled receptors (GPCRs) (37). Our analysis indicates that this binding motif needs to be redefined with regards to sequence conservation.

Distinct insulin sequence of Killifish reveals a non-canonical C-peptide cleavage site

As observed in the phylogenetic tree (Figure 1), a small clade diverges out from the ancestral node of insulin and IGFs. This clade consisting of only two insulin sequences belong to the killifish species and share a sequence identity of 39% with human insulin. These sequences belong to Turquoise killifish (Uniprot ID: A0A1A8ACH4) and Beira killifish (Uniprot ID: A0A1A8ISB4). Further, most of the residues involved in oligomerization and receptor binding interactions are maintained when compared to human insulin (Table S2 and S3). Moreover, when the insulin receptor sequences of these killifishes were compared with human insulin receptor sequence, it was observed that they had a sequence identity of around 60%. All the above observations indicate that the insulin from these killifish species might oligomerize and interact with their cognate receptor in a similar manner to human insulin.

The detailed observation of these diverged killifish sequences revealed an interesting anomaly at one of their post-translational C peptide cleavage sites. This cleavage is essential for processing of pro-insulin to its active form by the action of proconvertases. Canonically, the C-peptide cleavage site at the C-terminal end of B chain (cleaved by proconvertase 1/3) and N-terminal end of A chain (cleaved by proconvertase 2), is occupied by two consecutive dibasic residues, consisting of arginine or lysine (figure 3, black asterisk- B chain C-terminal, red asterisk- A chain N terminal). This is classified as type 2 precursor cleavage motif identified by proconvertases (38), which is marked by ((Arg/Lys)-(Arg/Lys)) motif. Interestingly, this dibasic cleavage site at C-terminal end of B chain is replaced by di-acidic aspartate residues (-Asp-Asp-) in the above killifish sequences (Figure 3).

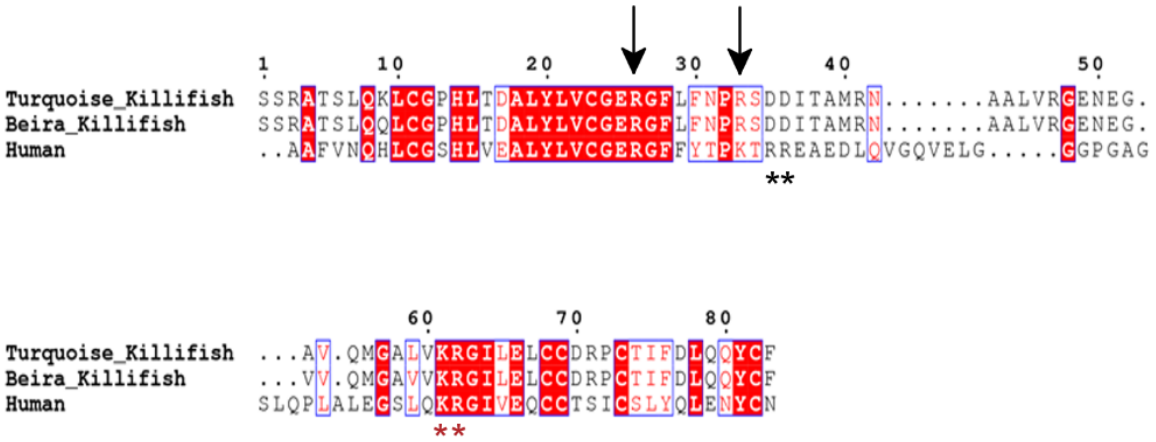

Figure 3: Multiple sequence alignment of insulin sequences, Human insulin (Uniprot ID: P01308), Turquoise killifish insulin (Uniprot ID: A0A1A8ACH4) and Beira killifish insulin (Uniprot ID: A0A1A8ISB4). The two black asterisks represent the change in cleavage site of C peptide at the C-terminal end of B chain. The two red asterisks represent the N-terminal cleavage site of the A chain which are conserved. The black arrows indicate the plausible type 3 cleavage motif found in killifish insulin sequence.

Replacement of two dibasic residues by consecutive residues of identical nature, -Asp-Asp-, requires a change of bases in adjacent codons that lead to two identical residues. Further, the change of codon from Arg to Asp requires a mutation of at least two bases in each codon. All these observations lead to prudent questions about possible C-peptide cleavage or lack thereof. However, upon closer observation, a plausible non-canonical cleavage motif for C-peptide in these killifish insulin was identified. This C-peptide cleavage site seems to adopt a type 3 precursor cleavage motif marked by R/K-X_n-R motif, where X is any other amino acid and "n" takes integer values of 0, 2, 4, 6 (in killifish sequences n=6, Figure 3) (38). The type 3 precursor cleavage motif is seen in pro-Dynorphin, pro-ANF, and E-peptide cleavage site of pro IGFs, but was never reported in insulin sequences. This suggests that proconvertase 1/3 which probably cleaves the C-peptide at the C-terminal of B chain, does so, by recognizing a different cleavage motif. It is interesting to note that these two killifishes inhabit landlocked fresh water ephemeral pools present mainly in Zimbabwe and Mozambique, and hence the observed non-canonical cleavage site might be the result of location specific evolutionary pressure. This difference in the C-peptide cleavage site of these killifishes might be attributed to their ability to evolve rapidly in response to environmental changes (39–41). In contrast, insulin sequence from a geographically less restricted killifish, Mummichog, found in marine sources of the coast of USA and Canada contains the canonical type 2 cleavage site. It should be noted that the rare type 3 precursor cleavage motif is observed in few other chordates (Figure S3).

Binding partner(s) imposed evolutionary constraint in IGFs

Another interesting observation to emerge out of this study is the unusually high conservation of IGF family members (IGF-1 and IGF-2), previously unreported. In fact, the functional domain of

IGF-1 sequence (excluding signal peptide) of human, horse, pig and Ord's kangaroo rat are 100% identical.

Unlike IGFs, insulin is stored in its stable hexameric form. Thus, the residues which are important for the stabilization of insulin fold, involved in oligomerization (Table S2), and those that interact with the cognate receptor (Table S3) are expected to be conserved during the evolution. However, IGFs are known to be monomers and only obvious evolutionary constraint is on those residues that are involved in interaction with their cognate receptors (IGF-1R and IGF-2R), suggesting IGFs should have less restraints on residue conservation. Surprisingly, among the ISPs, IGFs have the highest number of conserved residues in the A and B chains. In IGF-1, among the 50 residues of A and B chains, around 45 residues have conservation above 98% (Figure 2). Similarly, in the case of IGF-2 among the 49 residues, 37 of them have conservation above 98% (Figure 2).

It is known that the action of IGFs is regulated by IGF-binding proteins (IGFBPs) (42). IGFBPs bind to IGFs and increase the half-life of IGFs in the circulation (43) and regulate receptor binding, thus regulate their signaling (42). There are six IGFBPs (IGFBP-1 to IGFBP-6), they bind to IGFs with equal or greater affinity than the IGF-1R (44). Further, among ISPs, IGFs are the only two members whose interaction with their receptors is known to be controlled by regulatory proteins. The interaction of IGF-1 with its receptors and IGFBPs is depicted with a simple interactome (Figure S4). Based on these observations, we speculate that the observed high conservation of IGFs is probably due to the constraints imposed by its cognate receptors and several of these IGFBPs.

Through biophysical and mutational studies, it has been shown that the N and C-terminal domains of the IGFBPs are involved in binding to IGFs (45). Later, it was confirmed by the crystal structure of the complex of IGF-1 together with N and C-terminal domains of IGFBP-4 (46). Similarly, crystal structure of the complex of IGF-1 in combination with C-terminal domain of IGFBP-1 and N-terminal domain of IGFBP-4 also confirmed the similar mode of binding of IGF-1 with the IGFBPs (PDB entry: 2DSR, 2DSQ) (17). These structures revealed that the IGF-1 binds in a deep cleft formed between the N and C-terminal domains of the IGFBPs (Figure 4A and 4B). Further, buried surface area of IGF-1 upon complex formation (only A and B chain) with IGFBPs, was found to be around 60%. Assuming that the mode of binding of all the IGFBPs (IGFBP1-6) with IGF-1 is similar, even a slight variation in interaction of these IGFBPs with IGF-1 requires burial of additional surface area on IGF-1. Similarly, with additional solvent exposed area of IGF-1 involved in IGF-1R interaction, in particular α-CT region (Figure 4C), there is little room for surface exposed residues to change during the evolution. The above analysis suggests that the N and C-terminal domains of IGFBPs, almost engulf this miniature protein upon complex formation and this mode of interaction with several of its partners appear to exert constraints on its evolution. Although, in protein-protein interaction the interacting surfaces co-evolve (47), any change on IGF-1 requires concomitant change on all the six binding proteins and on IGF-1R, which is difficult to expect as these changes are random in nature. In

fact, almost all the residues of IGF-1 are highly conserved throughout chordates. In agreement with the above analysis, it was interesting to note that residues of IGF-1 (from the A and B chains) that are exposed in the complex are those that are not conserved (Figure 2 and Figure 4A).

Superposition of IGF-1 from the IGF-1:IGF-1R complex on to IGF-1 of IGF-1:IGFBP4 complex or IGF-1:IGFBP1/4 complex revealed that both IGFBPs and IGF-1R compete for largely a common surface on IGF-1. Further, the extent of interaction surface involved in the IGF-1:IGFBP interaction is more than that observed in the case of IGF-1:IGF-IR complex (Figure 4C). From the above analysis of the co-complex structures, it was clear that IGFBPs inhibit the interaction of IGF-1 with its receptor and hence regulate their function. Further it was observed that the α-CT region occupies some additional area other than that occupied by IGFBPs.

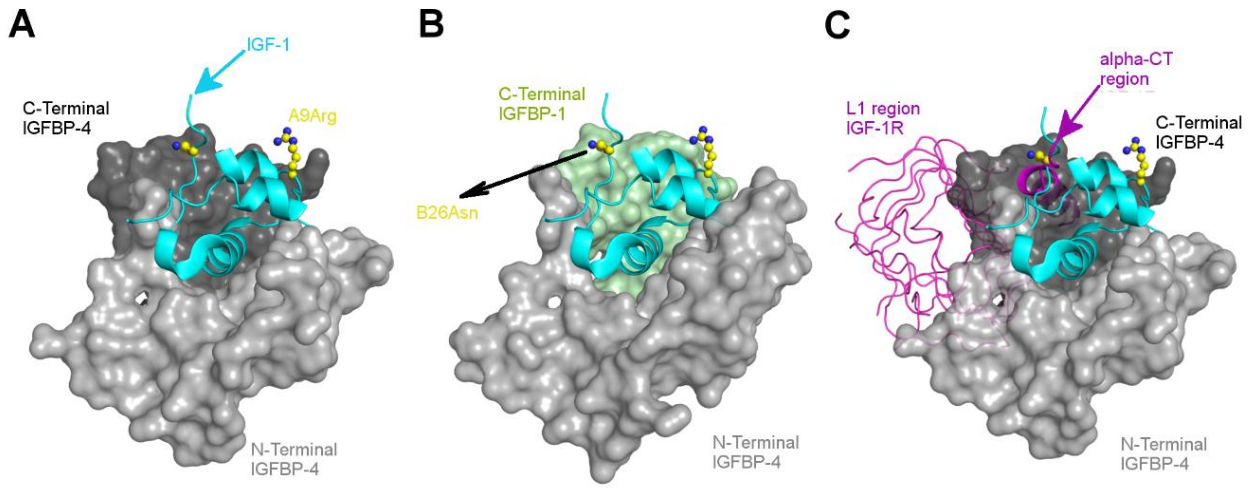

Figure 4: Interaction of IGF-1 with IGFBPs and IGF-1R: A Co-complex structure of IGF-1 with IGFBP-4 (PDB entry: 2DSR, N-terminal domain: light grey, surface representation and C-terminal domain: dark grey, surface representation). **B**: Co-complex structure of IGF-1 with IGFBP-4 and IGFBP-1 (PDB entry: 2DSQ, N-terminal domain IGFBP-4: light grey, surface and C-terminal domain of IGFBP-1: light green , surface). **C**: Superimposed structure of co-complex of IGF-1 and its receptor (5U8Q; L1 region-magenta, ribbon; αCT region- magenta cartoon) with co-complex structure of IGF-1 with IGFBP-4 (N-terminal; light grey, surface). IGF-1 is displayed in cartoon representation (cyan) across all the three panels. Note that the mode of interaction of IGF-1 with IGFBPs is very similar, where IGF-1 is bound between N and C-terminal domains of IGFBPs burying itself in a deep cleft.

All IGFBPs have the capability to bind to both IGF-1 and IGF-2 and IGFBP-1 to 5 are said to have no preference between IGF-1 and IGF-2, whereas, IGFBP-6 is supposed to have higher affinity to IGF-2 than IGF-1 (48). IGF-2 is known to bind to IGF-2R which is also known as cation-independent mannose-6-phosphate receptor, with distinct sequence and domain organizations when compared to IGF-1R and IR (49). Unfortunately, there are no experimentally determined structures of IGF-2 with any of the IGFBPs to draw inferences from. However, considering the similarity of IGF-1 and IGF-2 structures, it can be envisaged that the mode of interaction between IGFBPs and IGF-2 might be comparable to that of IGF-1, and hence similar

structural constraints as observed in the case of IGF-1 might be applicable to IGF-2. Thus, it appears that IGF's resistance to sequence diversity appears to be driven by binding partner(s) imposed constraints.

The other interesting aspect is that the IGF-1 has the shortest C-peptide (12 residue) among ISP members, which is known to interact with its receptor (50). In line with this observation, C1Gly, C2Tyr, C3Gly, C7Arg and C8Arg are highly conserved (Figure S4). From the analysis of co-complex structures (PDB entry: 5U8Q) it was clearly evident that the basic residues (C7Arg and C8Arg) are important in IGF-1-IGF-1R interaction where they interact with acidic loop in the CR region, which is supported by experimental evidences (50). Further discussions can be found in the supplementary information.

Low sequence conservation in relaxins and redefining their receptor binding motif (Relaxin 1, 2 and 3)

As observed in the phylogenetic tree (Figure 1), relaxins are distantly related to insulin and IGFs, and have very low conservation of sequence among themselves (Figure 5). Relaxins are known to perform different functions and interact with distinct GPCRs, which probably explain their sequence divergence. Further, only relaxin-2 is known to form dimers and there are no reports on dimerization of relaxin-1 or relaxin-3 (24, 27). Moreover, their function is not regulated by carrier proteins as observed in IGFs and hence their evolution is not probably regulated by partner imposed evolutionary restraints as observed in IGFs. This places less demand on the conservation of these surface exposed residues. Probably, all the reasons mentioned above contribute to the observed variation in relaxin sequences. Conversely, it is tempting to propose that the observed variation in the sequences of relaxin family hints at the possibility that they might not interact with many binding partners, unlike seen in the case of IGFs.

With all these sequence variations among relaxins, a conserved motif (RXXXRXXI/V) located on the B chain is shown to be important for receptor binding (51). When the receptor binding motif is analyzed separately for relaxin-1, 2 and 3, we noticed complete conservation in the receptor binding motif of relaxin-3 (**REFIRAVI**) which is not observed in relaxin-1 and 2 (Figure 5). As relaxin-3 is functionally divergent to relaxin-1 and 2, the high conservation of its receptor binding motif indicates a strict requirement for its receptor specific binding.

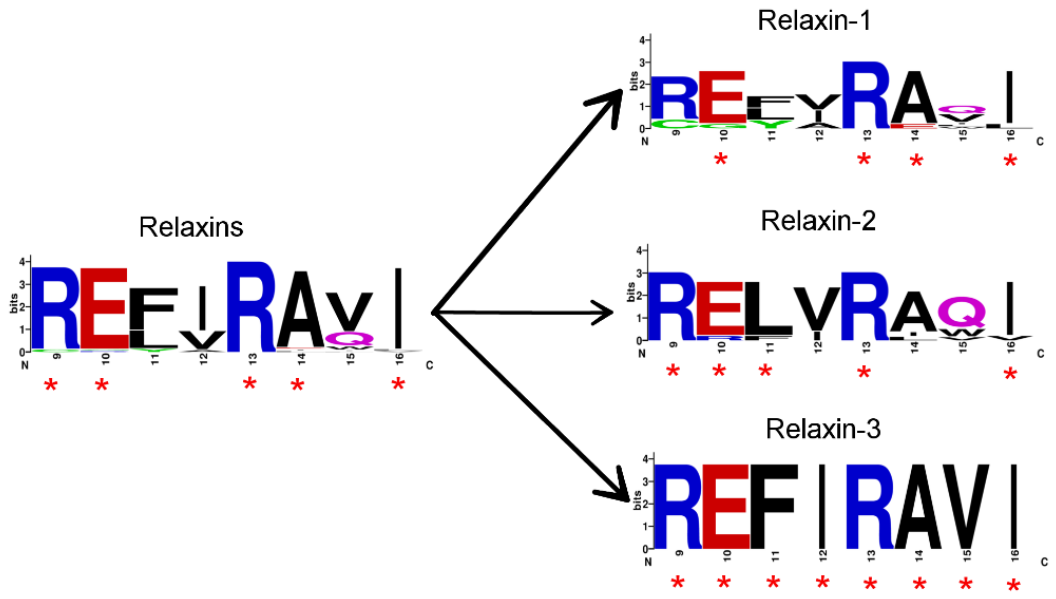

Figure 5: Receptor binding motif of relaxins. On the left WebLogo of receptor binding motif of relaxins (1, 2 and 3) and on the right the receptor binding motif of relaxin-1, 2 and 3 separately depicting the conservation of residues. The red asterisks represent the highly conserved residues (above 90%).

Surprisingly, we observed that the residues succeeding the conserved arginines (B9Arg and B13Arg) of this motif are also highly conserved among all the three relaxin paralogues. This additional conservation of B10Glu (95.9%) and B14Ala (93.9 %) in the relaxin family (relaxin-1, 2 and 3) may redefine the receptor binding motif as **RE**XX**RA**X**I/V**, instead of **R**XXX**R**XX**I/V**. Two other residues, A13Lys and B24Trp were also found to be conserved. While, B24Trp (89.1%) is known to be involved in receptor binding interaction (52), the role of highly conserved A13Lys (91.8% and 8.2% of Arg) was not mentioned before. Structural analysis reveals that A13Lys and B10Glu are in close proximity to the other two positively charged residues (B9Arg and B13Arg) of the receptor binding motif. Hence, A13Lys might also have a role in receptor binding interactions as it forms a largely positively charged patch along with the conserved residues B9Arg and B13Arg. Thus, we propose that A13Lys may also be a key residue involved in receptor binding interactions in relaxins along with B24Trp and the receptor binding motif.

Conclusion

Our analysis of sequence conservation/variation, which stems from the phylogenetic analysis and the information from functional and structural studies, provide unprecedented insights. For example, rationale for unexpectedly high evolutionary conservation of IGFs presented here, takes into account, the function of IGFs, their regulation by IGFBPs and structural information that decipher the mode of interaction of IGFs with IGFBPs and IGF-1R. Similarly, novel findings such as (a) non-canonical C-peptide cleavage site in killifish insulin sequences (b) redefined signature motif of relaxin family, all of which are a result of collective analysis of

phylogeny, sequence together with their structural and functional information. Further, in the absence of critical information regarding bombyxin receptor or its interaction with bombyxin, the sequence conservation alone suggests, unlike other insulin family members, bombyxin probably interacts with its receptor majorly utilizing its A chain. We were surprised to learn, how such an interlinked analysis that exploits information from multiple disciplines provide many profound insights on this well studied superfamily.

Insulin family proteins are known to evolve from a single ancestral gene and these Lilliputians of the protein world are the regulators of many physiological processes in higher organisms, including sugar metabolism, growth and reproduction. A recent discovery (53) reveals that the fairly new member of this family, relaxin-3, is a neuropeptide and a potential target for the treatment of various neuropsychiatric disorders (54). It is intriguing to note that nature has optimized these sequences to perform optimal and intended function over millions of years, encompassing species of diverse nature and habitat. It appears that it would be apt to borrow this wisdom from nature rather than engaging in high throughput experiments for the modification of these molecules. For example, Insulin analogue “insulin lispro” where the natural swapping of residues at the end of B-chain of IGF-1(-B28-B29-) with respect to insulin, is borrowed and incorporated on to insulin to create therapeutically successful fast acting analogue of insulin. Similar mutations can be strategized by learning from other members of the insulin superfamily for production of better insulin analogues. Intuitive designing of proteins by introducing characteristic mutations inspired from other members of this protein family might lead to potential therapeutics like dual-hormone Insulaxin (28). Conversely, this comprehensive study also points to how one can learn from nature to slightly manipulate the sequence of the protein without loss of primary function probably devoid of any cross reactivity.

Materials and methods

Sequence identification and analysis

Insulin family protein sequences were obtained from UniProt database (55) by using InterPro ID: IPR022352 and the clusters of each ISP members were obtained by giving 50% sequence identity cut off using UniRef database. The clusters thus obtained which consisted of at least 10 sequences were considered for the study. Some of the sequences were manually removed as they either contained incomplete sequence or were annotated as uncharacterized protein. CD-HIT server (56) was used to remove redundant dataset having 100% sequence identity by giving a higher identity cut-off of 99%, which gave 463 sequences (consisting B, C, D and E domains; for conventions of naming these domains see Figure S1) were obtained and utilized to perform multiple sequence alignment (MSA) using Clustal Omega (57). Out of the 463 sequences, insulin clade had 175 sequences, IGF-1- 98, IGF-2- 109, Bombyxin- 32 and relaxin- 49 sequences. Signal peptide was manually removed from all the sequences. MSA visualization, generation of consensus sequence and extraction of A, B and C chains were performed using Geneious (58). The images for multiple sequence alignment was generated using ENDscript server (59). The average sequence identity and most distant pair was determined using ALISTAT server (http://caps.ncbs.res.in/iws/alistat_ali.html). Residue numbering is in accordance with the numbering of master MSA sequence of processed A chain and B chains. Residues are considered to be conserved if they exhibit a conservation of >85% in the data set.

Phylogenetic analysis

The evolutionary history was inferred using the Minimum Evolution (ME) method (60). The optimal tree with the sum of branch length = 20.99532940 is shown. Rigorous bootstrapping (500 replicates) was performed. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (61) and are in the units of the number of amino acid substitutions per site. The ME tree was searched using the Close-Neighbor-Interchange (CNI) algorithm (62) at a search level of 1. The Neighbor-joining algorithm (63) was used to generate the initial tree. All positions with less than 95% site coverage were eliminated. Evolutionary analyses were conducted in MEGA6 (64). Phylogenetic tree was visualized by FigTree (http://tree.bio.ed.ac.uk/software/figtree/). Further, MSA for individual clades representing the families and few major paralogs as in the case of IGFs were obtained.

Structural analysis:

The protein structures of interest obtained from the world wide Protein Data Bank (wwPDB) were visualized using Pymol (65) and COOT (66). The PDB entry of proteins which were used for the study are as follows; insulin: 1ZNJ, 1EVR, 4F8F and 5BTS, IGF-1: 1GZR, IGF-2: 3KR3, relaxin-3: 2FHW, relaxin-2: 2MV1 and 6RLX and Bombyxin: 1BOM, IGF-1 binding proteins in

complex with IGF-1: 2DSR, 2DSQ, Insulin receptor in complex with insulin: 4OGA, IGF-1 receptor in complex with IGF-1: 5U8Q.

PDBe PISA server was used for the analysis to decipher the interacting residues in protein complexes and in the insulin oligomer (67). This was accomplished by analyzing the buried surface area between the protein complexes and oligomers. The buried surface area of IGF-1 upon interaction with the IGFBPs was calculated using CCP4's AreaImol (68).

Acknowledgments
SS would like to thank Admar mutt education foundation (AMEF) and Kongovi foundations for fellowship.

References

1. F.G.Banting, M.B, and C.H.Best (1922) The Internal Secretion Of The Pancreas. *J. Lab. Clin. Med.* **VII**, 251–266
2. Rinderknecht, E. (1977) The Amino Acid Sequence of Human Insulin-like Growth Factor I and Its Structural Homology with Proinsulin *. *J. Biol. Chem.* **253**, 2769–2776
3. Rinderknecht, E., and Humbel, R. E. (1978) Primary structure of human insulin-like growth factor II. *FEBS Lett.* **89**, 283–286
4. H. I,. Fevold, F. L. H. A. R. K. M. (1930) The Relaxative Hormone Of The Corpus Luteum. Its Purification And Concentration. *J. Am. Chem. Soc.* **52**, 3340–3348
5. Wilkinson, T. N., Speed, T. P., Tregear, G. W., and Bathgate, R. A. (2005) Evolution of the relaxin-like peptide family. *BMC Evol. Biol.* **5**, 14
6. Antonova, Y., Arik, A. J., Moore, W., Riehle, M. A., and Brown, M. R. (2012) *Insulin-like Peptides: Structure, Signaling, and Function*, Elsevier, 10.1016/B978-0-12-384749-2.10002-0
7. Nagasawa, H., Kataoka, H., Isogai, A., Tamura, S., Suzuki, A., Mizoguchi, A., Fujiwara, Y., Suzuki, A., Takahashi, S. Y., and Ishizaki, H. (1986) Amino acid sequence of a prothoracicotropic hormone of the silkworm *Bombyx mori*. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 5840–5843
8. Lagueux, M., Lwoff, L., Meister, M., Goltzenfi, F., and Hoffmann, J. A. (1990) cDNAs from neurosecretory cells of brains of Locusta migratoria (Insecta , Orthoptera) encoding a novel member of the superfamily of insulins. **254**, 249–254
9. Smit, A. B., E.VREUGDENHIL, R.H.M.EBBERINK, W.P.M.GERAERTS, J.KLOOTWIJIK, and J.JOOSSE (1988) Growth-controlling molluscan neurons produce the precursor of an insulin-related peptide. *Lett. to Nat.* **331**, 535–538
10. Duret, L., Guex, N., Peitsch, M. C., Bairoch, A., Lyon, B., Cedex, F.-V., and Biomedical, G. (1998) New Insulin-Like Proteins with Atypical Disulfide Bond Pattern Characterized in Caenorhabditis elegans by Comparative Sequence Analysis and Homology Modeling. *Genome Res.* **8**, 348–353
11. Chan, S. J., and Steiner, D. F. (2000) Insulin Through the Ages: Phylogeny of a Growth Promoting and Metabolic Regulatory Hormone. *Am. Zool.* **40**, 213–222
12. J.Murray-rust, A.N.McLeod, T.L.Blundell, and S.P.Wood (1992) Structure and Evolution of Insulins: Implications for receptor binding. *BioEssays*. **14**, 325–331
13. Bajaj, M., Blundell, T., and Wood, S. (1984) Evolution in the Insulin Family - Molecular Clocks That Tell the Wrong Time. *Biochem. Soc. Symp.*
14. Conlon, J. M. (2001) Evolution of the insulin molecule : insights into structure-activity and phylogenetic relationships. *Peptides*. **22**, 1183–1193
15. Ebberink, R. H. M., Smit, A. B., and Minnen, J. Van (1989) The Insulin Family: Evolution of Structure and Function in Vertebrates and Invertebrates. *Biol. Bull.* **177**, 176–182

16. Rosengren, K. J., Lin, F., Bathgate, R. A. D., Tregear, G. W., Daly, N. L., Wade, J. D., and Craik, D. J. (2006) Solution structure and novel insights into the determinants of the receptor specificity of human relaxin-3. *J. Biol. Chem.* **281**, 5845–5851
17. Sitar, T., Popowicz, G. M., Siwanowicz, I., Huber, R., and Holak, T. A. (2006) Structural basis for the inhibition of insulin-like growth factors by insulin-like growth factor-binding proteins. *PNAS.* **103**, 13028–13033
18. Dransfield, D. T., Cohen, E. H., Chang, Q., Sparrow, L. G., Bentley, J. D., Dolezal, O., Xiao, X., Peat, T. S., Newman, J., Pilling, P. A., Phan, T., Priebe, I., Brierley, G. V., Kastrapeli, N., Kopacz, K., Martik, D., Wassaf, D., Rank, D., Conley, G., Huang, Y., Adams, T. E., and Cosgrove, L. (2010) A Human Monoclonal Antibody against Insulin-Like Growth Factor-II Blocks the Growth of Human Hepatocellular Carcinoma Cell Lines In vitro and In vivo. *Mol. Cancer Ther.* **9**, 1809–1819
19. Xu, Y., Kong, G. K. W., Menting, J. G., Margetts, M. B., Delaine, C. A., Jenkin, L. M., Kiselyov, V. V., De Meyts, P., Forbes, B. E., and Lawrence, M. C. (2018) How ligand binds to the type 1 insulin-like growth factor receptor. *Nat. Commun.* **9**, 1–13
20. Hjorth, C. F., Norrman, M., Wahlund, P. O., Benie, A. J., Petersen, B. O., Jessen, C. M., Pedersen, T., Vestergaard, K., Steensgaard, D. B., Pedersen, J. S., Naver, H., Hubálek, F., Poulsen, C., and Otzen, D. (2016) Structure, Aggregation, and Activity of a Covalent Insulin Dimer Formed during Storage of Neutral Formulation of Human Insulin. *J. Pharm. Sci.* **105**, 1376–1386
21. Skyler, J. S. (2012) *Atlas of Diabetes*
22. Bergman, D., Halje, M., Nordin, M., and Engström, W. (2013) Insulin-Like Growth Factor 2 in Development and Disease : A Mini-Review. *Gerontology*. **59**, 240–249
23. Patil, N. A., Rosengren, K. J., Separovic, F., Wade, J. D., Bathgate, R. A. D., and Hossain, M. A. (2017) Relaxin family peptides: structure–activity relationship studies. *Br. J. Pharmacol.* **174**, 950–961
24. Eigenbrot, C., Randal, M., Quan, C., Burnier, J., Lori, O., Rinderknecht, E., and Kossiakoff, A. A. (1991) X-ray Structure of Human Relaxin at 1.5 A : Comparison to insulin and implications for Receptor Binding Determinants. *J. Mol. Biol.* **221**, 15–21
25. Grossman, J., and Frishman, W. H. (2010) A New Approach for the Treatment of Acute Congestive Heart Failure. *Cardiol. Rev.* **18**, 305–312
26. Sherwood, O. D. (2004) Relaxin's physiological roles and other diverse actions. *Endocr. Rev.* **25**, 205–234
27. Haugaard-Kedstrom., L. M., Hossain, M. A., Daly, N. L., Bathgate, R. A. D., Rinderknecht, E., Wade, J. D., Craik, D. J., and Rosengren, K. J. (2015) Solution structure, aggregation behavior, and flexibility of human relaxin-2. *ACS Chem. Biol.* **10**, 891–900
28. Schwabe, C., and Büllesbach, E. (1998) *Relaxin and the Fine Structure of Proteins*, Springer
29. Tanaka, M., Iijima, N., Miyamoto, Y., Fukusumi, S., Itoh, Y., Ozawa, H., and Ibata, Y. (2005) Neurons expressing relaxin 3/INSL 7 in the nucleus incertus respond to stress. *Eur. J. Neurosci.* **21**, 1659–1670
30. Ishizaki, H., and Suzuki, A. (1994) The brain secretory peptides that control moulting and metamorphosis of the silkmoth, Bombyx mori. *Int. J. Dev. Biol.* **38**, 301–310
31. Mizoguchi, A., and Okamoto, N. (2013) Insulin-like and IGF-like peptides in the silkmoth Bombyx mori: Discovery, structure, secretion, and function. *Front. Physiol.* **4 AUG**, 1–11
32. Blundell, T., Dodson, G., Hodgkin, D., and Mercola, D. (1972) Insulin: The Structure in the Crystal and its Reflection in Chemistry and Biology. *Adv. Protein Chem.* **26**, 279–402
33. Kawakami, A., Iwami, M., and Ishizaki, H. (1989) Structure and organization of four clustered genes that encode bombyxin , an insulin-related brain secretory peptide of the silkmoth Bombyx mon. *Biochemistry*. **86**, 6843–6847
34. Fullbright, G., Lacy, E. R., and Büllesbach, E. E. (1997) The prothoracicotropic hormone bombyxin has specific receptors on insect ovarian cells. *Eur. J. Biochem.* **245**, 774–80
35. Blundell, T. L., Ishizaki, H., Jhoti, H., and Mcleod, A. N. (1987) Prothoracicotropic hormone has

an insulin-like structure. *FEBS Lett.* **219**, 419–425
36. Mizoguchi, A., Okamoto, N., Riddiford, L. M., Hughes, H., and Brown, M. R. (2013) Insulin-like and IGF-like peptides in the silkmoth Bombyx mori : discovery , structure , secretion , and function. **4**, 1–11
37. Bathgate, R. A. D., Halls, M. L., Westhuizen, E. T. Van Der, Callander, G. E., Kocan, M., and Summers, R. J. (2013) Relaxin Family and Their Receptors. *Physiol. Rev.* **93**, 405–480
38. Seidah, N. G., Day, R., and Mieczyslaw Marcinikiewicz, A. Mi. C. (1998) Precursor Convertases: An Evolutionary Ancient, Cell-Specific, Combinatorial Mechanism Yielding Diverse Bioactive Peptides and Proteins. *Ann. NEW YORK Acad. Sci.* **May**, 15–24
39. Valenzano, D. R., Singh, P. P., Beyer, A., Johnson, E. A., and Brunet, A. (2015) The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Resource The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan. *Cell*. **163**, 1539–1554
40. Whitehead, A., Clark, B. W., Reid, N. M., Hahn, M. E., and Nacci, D. (2017) When evolution is the solution to pollution: Key principles, and lessons from rapid repeated adaptation of killifish (*Fundulus heteroclitus*) populations. *Evol. Appl.* **10**, 762–783
41. Reid, N. M., Proestou, D. A., Clark, B. W., Warren, W. C., Colbourne, J. K., Shaw, J. R., Karchner, S. I., Hahn, M. E., Nacci, D., Oleksiak, M. F., Crawford, D. L., and Whitehead, A. (2016) The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*. **354**, 1305–1308
42. Kim, H., Rosenfeld, R. G., and Oh, Y. (1997) Biological roles of insulin-like growth factor binding proteins (IGFBPs). *Exp. Mol. Med.* **29**, 85–96
43. Guler, H., Schmid, C., and Froesch, E. R. (1989) Insulin-like growh factors I and II in healthy man. Estimations of half-lives and production rates. *Acta Endocrinol. (Copenh).* **121**, 753–758
44. Allard, J. B., and Duan, C. (2018) IGF-Binding Proteins: Why Do They Exist and Why Are There So Many? *Front. Endocrinol. (Lausanne).* **9**, 113–129
45. Carrick, F. E., Forbes, B. E., and Wallace, J. C. (2001) BIAcore Analysis of Bovine Insulin-like Growth Factor (IGF) -binding Protein-2 Identifies Major IGF Binding Site Determinants in Both the Amino- and Carboxyl-terminal Domains *. *J. Biol. Chem.* **276**, 27120–27128
46. Siwanowicz, I., Popowicz, G. M., Wisniewska, M., Huber, R., Kuenkele, K., Lang, K., Engh, R. A., Holak, T. A., and Martinsried, D.- (2005) Structural Basis for the Regulation of Insulin-like Growth Factors by IGF Binding Proteins. *Structure*. **13**, 155–167
47. Teppa, E., Zea, D. J., and Marino-Buslje, C. (2017) Protein–protein interactions leave evolutionary footprints: High molecular coevolution at the core of interfaces. *Protein Sci.* **26**, 2438–2444
48. Headey, S. J., Keizer, D. W., Yao, S., Brasier, G., Kantharidis, P., Bach, L. A., and Norton, R. S. (2004) C-Terminal Domain of Insulin-Like Growth Factor (IGF) Binding Protein-6: Structure and Interaction with IGF-II. *Mol. Endocrinol.* **18**, 2740–2750
49. Brown, J., Delaine, C., Zaccheo, O. J., Siebold, C., Gilbert, R. J., Van Boxel, G., Denley, A., Wallace, J. C., Hassan, A. B., Forbes, B. E., and Jones, E. Y. (2008) Structure and functional analysis of the IGF-II/IGF2R interaction. *EMBO J.* **27**, 265–276
50. Bayne, M. L., Applebaum, J., Underwood, D., Chicchi, G. G., Green, B. G., Hayes, N. S., and Cascieri, M. a (1989) The C region of human insulin-like growth factor (IGF) I is required for high affinity binding to the type 1 IGF receptor. *J. Biol. Chem.* **264**, 11004–8
51. Bathgate, R. A. D., Halls, M. L., van der Westhuizen, E. T., Callander, G. E., Kocan, M., and Summers, R. J. (2013) Relaxin Family Peptides and Their Receptors. *Physiol. Rev.* **93**, 405–480
52. Kuei, C., Sutton, S., Bonaventure, P., Pudiak, C., Shelton, J., Zhu, J., Nepomuceno, D., Wu, J., Chen, J., Kamme, F., Seierstad, M., Hack, M. D., Bathgate, R. A. D., Hossain, M. A., Wade, J. D., Atack, J., and Lovenberg, T. W. (2007) R3 (B ⌬ 23 – 27) R / I5 Chimeric Peptide , a Selective Antagonist for GPCR135 and GPCR142 over Relaxin Receptor LGR7. *J. Biol. Chem.* **282**, 25425–25435
53. Bathgate, R. A. D., Samuel, C. S., Burazin, T. C. D., Layfield, S., Claasz, A. A., Reytomas, I. G.

T., Dawson, N. F., Zhao, C., Bond, C., Summers, R. J., Parry, L. J., Wade, J. D., and Tregear, G. W. (2002) Human Relaxin Gene 3 (H3) and the Equivalent Mouse Relaxin. *J. Biol. Chem.* **277**, 1148–1157

54. Kumar, J. R., Rajkumar, R., Jayakody, T., Marwari, S., Hong, J. M., Ma, S., Gundlach, A. L., Lai, M. K. P., and Dawe, G. S. (2017) Relaxin' the brain: a case for targeting the nucleus incertus network and relaxin-3/RXFP3 system in neuropsychiatric disorders. *Br. J. Pharmacol.* **174**, 1061–1076
55. Bateman, A. (2019) UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515
56. Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**, 1658–1659
57. Sievers, F., and Higgins, D. G. (2014) Clustal Omega. *Curr. Protoc. Bioinforma.* **2014**, 3.13.1-3.13.16
58. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. **28**, 1647–1649
59. Robert, X., and Gouet, P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, 320–324
60. Rzhetsky, A., and Nei, M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.* **10**, 1073–1095
61. Zuckerkandl, E., and Pauling, L. (1965) Evolutionary Divergence and Convergence in Proteins. *Evol. Genes Proteins*. 10.1016/b978-1-4832-2734-4.50017-6
62. Nei, M., and Kumar, S. (2000) *Molecular evolution and phylogenetics*, Oxford University Press, [online] https://global.oup.com/academic/product/molecular-evolution-and-phylogenetics-9780195135855?cc=jp&lang=en& (Accessed August 6, 2019)
63. Saitou, N., and Nei, M. (1998) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* **4**, 1–20
64. Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729
65. The PyMOL Molecular Graphics System, Version 1.8 (2015)
66. Emsley, P., and Cowtan, K. (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 2126–2132
67. Krissinel, E., and Henrick, K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*
68. B.Lee and F.M.Richards (1971) The Interpretation of Protein Structures : Estimation of Static Accessibility. *J.Mol.Bio*