3$_{10}$ and $\pi$-Helices: Poisson events on sequence space, possible implications.

Param Priya Singh, Anirban Banerji*
Bioinformatics Centre, University of Pune
Pune-411007, India.
Contact address : anirbanab@gmail.com

November 25, 2009

Abstract

Although $\alpha$−helices and $\beta$-sheets dominate the composition of proteins, other secondary structures find their places therein too. 3$_{10}$ and $\pi$-helices are two such rare secondary structures. There is very little objective insight about various statistical aspects regarding the nature of their occurrences. Comprehensive set of reasons behind the existence of 3$_{10}$ and $\pi$-helices can only be obtained if the occurrence profile of these on the primary structure is unambiguously described from the perspectives of sequence, structure and evolution. Although studies about the compositional and energetic profile of 3$_{10}$ and $\pi$-helices aren’t uncommon, merely that doesn’t tell us why these (rather unstable) structures are found in the proteins at the first place. Considering all the non-redundant protein structures across all the major structural classes, the present study attempts to find the probabilistic distributions that describe several facets of the occurrence of these rare secondary structures in proteins. Structural causes for observing these statistical patterns are explained too. Probabilistic profiling of the occurrence of 3$_{10}$ and $\pi$-helices reveal their presence to follow Poisson flow on the sequence. Thorough statistical analysis of sequence intervals between consecutive occurrences of 3$_{10}$ and $\pi$-helices, support this finding. With extensive analysis from varying standpoints we prove here that, such Poisson flows suggest the 3$_{10}$-helices and especially $\pi$-helices to be evidences of nature’s mistakes on folding pathways. This hypothesis is further supported by results of critical evolutionary analysis on 20 major protein domain families, which reveal the definitive trend that proteins try to dispose these structures off during evolution, in favor of $\alpha$-helices. Alongside these unexpected and significant results, a new algorithm to differentiate between related sequences is proposed here that reliably studies evolutionary distance with respect to protein secondary structures. Building upon a firm foundation of structural information, this study attempted to address protein evolution from the perspective of secondary structures, with rigorous statistical and algorithmic framework.

Keywords : 3$_{10}$-helix, $\pi$-helix, Poisson flow, protein evolution, evolutionary distance calculation, secondary structures.

Introduction :

3$_{10}$-helices are enigmatic characters. Possibilities of their existence were discussed a good eight years before Pauling proposed the structure of $\alpha$−helix[1,2]. But the number of studies on them are minuscule, when compared to the same on $\alpha$−helices. However, even though the amassed information isn’t prolific, a systematic survey of them might take any student of structural biology for a bumpy ride. That is because works on them often contradict each other, leaving us with a (kind of) confusing picture.
To begin from the beginning, we jot down the things that we know about 3\textsubscript{10}-helices. They have 3 residues per turn (that explains the name) with a translation of 2Å along the helix axis with a hydrogen bonding pattern \([i → (i + 3)]\). \(~3\%\) of the residues are known to exist in 3\textsubscript{10}-helices (compared to\(~32\%\) of residues in \(\alpha\)-helices)[3-5]. 3\textsubscript{10}-helices are irregular in shape with an average length of 3-4 residues; finding one of them with more than 6 residues is extremely rare[4]; in contrast, \(\alpha\)-helices adopt long, regular helical structures in proteins. This marks, to quite an extent, the end of range \(\sim\) 3 residues per turn (that explains the name) with a translation of 2Å along the helix axis with a hydrogen bonding pattern.

To begin from the beginning, we jot down the things that we know about 3\textsubscript{10}-helices in the torsion angle space, the ideal assessment was \((-60^0, -30^0)\)[3]. In reality, one study [4] asserted the preferred mean value of \(\phi - \psi\) for them is \((-71^0, -18^0)\), the other[6] claimed it to be \((-50^0, -30^0)\) (apparently, a more precise variant of it [7] read \((-49^0, -26^0)\)), according to another [8] the same is \((-63^0, -17^0)\), while according to still another [9] it is \((-68^0, -17^0)\); and for the 3\textsubscript{10}-helices containing non-native amino acids the preferred \(\phi - \psi\) value was found to be \((-54^0, -35^0)\)[10].

Moving on to energetic perspective, while a number of theoretical and computational studies asserted that \(\alpha\)-helices are energetically more stable than the 3\textsubscript{10}-helices[11-15], analysis of ESR (Electron Spin Resonance) spectral data [16,17] pointed at a coexistence of 3\textsubscript{10} and \(\alpha\)-helices. Although it is easy to understand that initiation for helix formation will be easier in case of 3\textsubscript{10}-helix than for \(\alpha\)-helix (one fewer unit to consider before the formation of first hydrogen bond), it is found from literature [18] that it is not 3\textsubscript{10}-helices but "only" \(\alpha\)-helices, \(\beta\)-sheets, or short covalently bridged cycles (as in conotoxins or in metallothioneins), which can serve as nucleations for initiating protein folding. Easy melting of 3\textsubscript{10}-helices at lower temperature[19], serves as another proof of the unstable nature of their structures.

Variation in the \(\phi - \psi\) magnitude in \(\alpha\)-helices can also be detected easily; however, such variations can always be observed to be limited within a small area of 'Ramachandran map', implying clearly a deterministic well-cut-out plan of nature to assign categorical significance to \([i → (i + 4)]\) hydrogen bonding pattern. Remarkably, as shown later in this work, despite sharing the same energy minima with \(\alpha\)-helices in \(\phi - \psi\) space [20] and possessing an \(\alpha\)-helix-like CD spectrum[21], the 3\textsubscript{10}-helices’ \(\phi - \psi\) values can be observed to populate all possible nooks and crannies of 'Ramachandran map'. The considerable extent of \(\phi - \psi\) variation in these helices hinted us to explore the probability of looking at these helices as some kind of ad-hoc structures that had been used while making the primary structure fold, but with which much of well-cut-out significance scheme had not been attached, as it was done with \(\alpha\)-helices. A previous work [3] had studied particular causalities behind unexpected torsion angles in 3\textsubscript{10}-helices, but it was from a bottom-up perspective (treating amino-acids one-by-one) and therefore could not provide a general answer to a simple question “why at the first place, are they there?”. Many a studies (helix-coil transition [22], helix nucleation [3], transitions between 3\textsubscript{10} and \(\alpha\)-helical conformers in domain motions ([23],[24]), molecular dynamics based work to observe \([i → (i + 3)]\) hydrogen bonding pattern in helical peptides[25],[26]) - have described the role of 3\textsubscript{10}-helices. But a careful observation into them reveals that all of them merely tried to justify the existence of them in proteins. But none of them, could explain what the proteins would have missed if 3\textsubscript{10}-helices were not there (analogous question on \(\alpha\)-helices, would have been an easy assignment for freshman structural biology student; but not this one).

Rigorous literature survey could dig up works reporting the presence of a 3\textsubscript{10}-helix in a groove formed between two \(\alpha\)-helices within a protein[27], their (possible) role in protein-protein interaction [28, the English abstract of a Japanese article] and in motif formation[29]. Some other studies revealed the presence of single 3\textsubscript{10}-helices in active sites [30,31] and possible role of a single 3\textsubscript{10}-helix in RNA-binding[32] and receptor-binding[33]. However, an immensely interesting trend could be observed in the way the “role’s of 3\textsubscript{10}-helices were reported in these studies. Expressions like - "the larger flexible loop includes one turn of a 3\textsubscript{10}-helix that comprises the binding site .."[30], or “a short 3\textsubscript{10}-helix, found immediately N-terminal to the first \(\beta\)-strand in RRM1, may interact with RNA directly"[32], or “the active site cysteine lies in a cleft formed by a coil region that includes the 3\textsubscript{10}-helix and a
loop...[31], or “a possible physiological role of the 310-helix present in G-CSF for its receptor binding activity”[32] strongly indicate the predominant trend of 310-helix research. For a student of structural biology, it becomes a little difficult to assign great importance to “one turn of a 310-helix” or “a short 310-helix” (when we know that their average length is 3-4 residues[4]); in rare and isolated occurrences that do not follow a general pattern across universe of proteins. On top of it, use of the words like “may” and “possible”, and an observation of occurrence of a 310-helix alongside a loop within a coil etc. - suggests the student that the aura of importance attached to the 310-helices, might be a little too amplified.

Picture of 310-helix studies from the realm of helix-coil transition and (separately) from helix formation studies is as eminently confusing as it is from any other sphere; with two notable differences. First, many of these studies involve peptides (especially, the α-Aminoisobutyric acid (AIB)). Second, the minute aspects \(i \rightarrow (i + 3)\) and \(i \rightarrow (i + 4)\) hydrogen bond studies depend on exact definition of a hydrogen bond used in the calculation. It is interesting to note that AIB is not a proteinogenic amino acid (in fact, it is rare to be found in nature), but in oligomeric form it readily forms 310-helices. To what extent a hypothesis constructed out of AIB studies be relevant in the complex domain of proteins, is a rather philosophical question and we refrain ourselves from commenting on that. However the very fact that proteins are complex and enormously diverse machines than the peptides, can never be questioned. Studying the transition phenomenon with various peptides and (some) proteins, some works [22,16,34] claim the formation of 310-helices as necessary step in α-helix formation; while another[20,35], drawing upon the results of double-label ESR and NMR studies on Ala-rich peptide sequences, claim a coexistence of α and 310-helices. The \(i \rightarrow (i + 3)\) and \(i \rightarrow (i + 4)\) hydrogen-bonding studies faithfully reproduces this conundrum. In a study [36] of exceptional relevance to the present one, it was confirmed that \(i \rightarrow (i + 3)\) bond formation does take place while helix formation and denaturation were studied with synthetic Ala-based peptides; but the same work failed to observe the formation complete 310-helices !! Another work [37] reported the “curious” case of \(i \rightarrow (i + 3)\) bond formation amidst \(i \rightarrow (i + 4)\) bond breaking while exploring alanine-based peptides with MD. In an expression, reminiscent of the ones presented in the last paragraph, “transient turn structures with \(i \rightarrow (i + 3)\) hydrogen bond” was reported [38] while studying unfolding on an 18-residue peptide with 1 ns MD. But tackling the problem the other way round, another study asserted that formation of 310-helices is not a necessary step in the transition from coil to helix[39]. - Taken as a whole, perfect democratic coexistence of all kinds of contradicting findings with no clear picture.

Thus, to summarize it all, first, we couldn’t to find a single work where a 310-helix has been identified to perform either a structural or functional role which a (small) α-helix at the same coordinate would have failed to perform. In fact, a whole slew of studies point out how unstable they are. Second, population of 310-helices in all possible corners of Ramachandran map (a cursory glance at 310 \(\phi-\psi\) space (fig-1)a) reveals the presence of points like : \((-60^0, +140^0)\), \((+70^0, +30^0)\), \((+100^0, -150^0)\), \((-40^0, -40^0)\) with two remarkable cases at about \((+170^0, +170^0)\) and \((+90^0, -90^0)\) - suggested categorically that, nature doesn’t always prioritize the act of ensuring local or locally-global or global energy minima for the 310-helices. (The same for \(\pi\)-helices is presented in (fig-1)b). In fact, as a recent study [40] pointed out, since there are position-specific shifts in \(\phi-\psi\) of the 310-helices in proteins, attempting to describe structural invariants in them with \(\phi-\psi\) construct, isn’t tenable at all. Third, although the role of 310-helices on the formation and uncoiling of α-helices has been pointed out previously ([20],[23]), no study have ever proved conclusively that presence of 310-helices is obligatory in either the formation or uncoiling of α-helices. Fourth, “one turn of a 310-helix” or “a short 310-helix” etc., might at times be associated with some role in either protein structure stabilization or in protein function; but an overwhelming absence of that trend across protein universe implies clearly that such singular occurrences are mere accidents and not a part of general (efficient) mechanism that characterizes functioning of nature. Hence, for an observer of protein reality who relies purely on the data, it might not be unreasonable to consider 310-helices as pure liabilities, from a structural per-
spective. Although one study [41] had reported the occurrence of $3_{10}$-helices on the protein surfaces, merely that does not ensure their possible role in protein function. We could not find a study that shows a statistically significant (or at least five or ten cases) where a $3_{10}$-helix performs an extremely important protein function and where this particular function couldn’t have been performed by a small $\alpha$-helix.

To find an answer to our simple yet fundamental question “why are they there?” we chose to define the problem from a statistical standpoint. To be precise, we wanted to characterize the statistical pattern of occurrence of $3_{10}$-helices on primary structures of all the protein structures of non-redundant PDB[42] ranging across all the structural domains of SCOP[43]. The objective and unambiguous nature of statistical distribution in their occurrence profile will provide an ideal framework to study other outstanding issues related to $3_{10}$-helices. To explore the possibility of a latent periodicity or hidden pattern in the (rare) occurrences of $3_{10}$-helices, a rigorous mathematical survey of the separation between individual occurrences of $3_{10}$-helices on the sequence axis, across all the SCOP classes, were carried out.

As one scans through the primary structure of a protein, certain characteristics in the mode of the appearance of $3_{10}$-helices stand out prominently. In objective description of the situation, suppose at the coordinate $S_1$ of the primary structure the observer notices a $3_{10}$-helix. He notices the next $3_{10}$-helix at another point $S_2$ and then at $S_3$ and so forth. Although the energetic details of co-existence of $3_{10}$-helices and $\alpha$-helices have been reported in some previous studies([20],[23]); it is not biophysically viable to assume that every time there’s a $3_{10}$-helix, it would have to be in the vicinity of some $\alpha$-helix in the primary structure. In our dataset we observed that these rare secondary structures can occur anywhere in the structure: in the vicinity of $\alpha$-helices and sheets, as well as independent of any other regular secondary structure. Hence, while considering the occurrences of $3_{10}$-helices on the primary structure, we described them as they are (i.e., without resorting to $\alpha$-helices to describe the coordinate of a $3_{10}$-helix). With such a construct to scan the sequence, the observer might notice some extremely interesting facts, namely:

**The biophysical details:**

**B.1)** Very few residues fall into the class of $3_{10}$-helices (it was observed in a previous study [5] that only 4% of the residues could be classified into $3_{10}$-helices, from their select data-base).

**B.2)** Even within the helix population, $\sim 20\%$ of all protein helices adopt the $3_{10}$-helix conformation [20]. The number of $\pi$-helices is even less. (On a related note, although a previous work [4] had reported the presence of $3_{10}$-helices in all-$\beta$ proteins, we failed to find a single instance of such occurrences).

**The mathematical details:**

**M.1)** No periodic (or other global) pattern could be detected in the occurrence profile of $3_{10}$-helices.

**M.2)** Probability of two or more $3_{10}$-helices occurring at the same coordinate of primary sequence is zero (hence the occurrence profile of the $3_{10}$-helices constitutes an ordinary flow of events).

**M.3)** (Related to M.1) The chance of not finding a $3_{10}$-helix for $i$ units of sequence during a search is the same as that of a fresh search that fails to find a $3_{10}$-helix in the next $i$ unit of sequence. In other words, past history (implying search conducted along the sequence till the point search has reached) has no effect on finding or not-finding a $3_{10}$-helix. Mathematically:

$$P(S > s_i + \delta | S > s_i) = P(S > s_i),$$

where $s_i$ denotes the coordinate of sequence where the search operation is halting at present and $\delta$ is the distance measured on the sequence units, on which the search for $3_{10}$-helix is being carried out.

This property suggests that occurrence of $3_{10}$-helices on the primary structure possess the 'memoryless property'. We note further that a distribution that attempts to describe the occurrence profile of the $3_{10}$-helices, should be able to describe the amount of sequence length one needs to scan before the
event of detecting another $3_{10}$-helix occurs. At the same time, this distribution should be continuous in nature.

**M.4** (Related to M.1 and M.3) Differences in the occurrence profile of $3_{10}$-helices on the sequence axis ($S_2 - S_1$, $S_3 - S_2$, ..., $S_{i+1} - S_i$, ...) are stochastically independent for any natural $i$.

Denoting the event of detection of a $3_{10}$-helix by $D_{S_i}$, we can generalize M.4 to describe the stochastic independence of the differences in the occurrence profile of $3_{10}$-helices on the sequence axis by asserting:

$$[ (D_{S_1} - D_{S_0}), (D_{S_2} - D_{S_1}), \ldots, (D_{S_i} - D_{S_{i-1}}) ]$$

are independent for any natural $i$ and further, for every $0 < S_0 < S_1 < \ldots < S_i$.

Together these properties (M.2 and M.1, M.3, M.4, along with B.1 and B.2) imply that there exists a real parameter $\lambda$, such that for every $i$ on the distribution of $S_i$, it holds true; and as a whole, describes a Poisson distribution with parameter $\lambda s$, where $s$ denotes the particular length on sequence axis between two consecutive $3_{10}$-helices.

In fact, a distribution endowed with characteristics and requirements as stated above, can be considered as a special class of continuous probability distributions, which describe the sequence intervals between detection of two consecutive $3_{10}$-helices. The process of detections occur independently yet continuously at a constant average rate on any sequence under consideration, a hallmark of Poisson processes. The exponential distribution occurs naturally when describing the lengths of the inter-arrival times in a homogeneous Poisson process [44-46].

Hence to define our problem categorically, we attempt to construct a counting algorithm that scans through the primary structure and tries to detect the $3_{10}$-helix, before verifying whether the occurrence profile of these follow a mathematical pattern, or not. Search for any particular $3_{10}$-helix takes place on a random length of sequence of extent $S$ (that is, the length of entire primary structure under consideration). Searching operation has an exponential distribution with parameter $\mu(\mu = 1/\beta)$, where $\beta$ denotes the actual number of occurrence of the $3_{10}$-helix in $S$. On the basis of observation (discussed above), we accept that the appearance of $3_{10}$-helices follow a Poisson flow with intensity $\lambda$ ($\lambda$ being the average number of detection of $3_{10}$-helix per unit traversal of sequence length; thus $\lambda = \lambda(s)$. The constancy or variability of $\lambda$ will be determined by the sequence under consideration).

Designing a full-proof detection scheme to identify $3_{10}$-helices is extremely difficult. Hence, we introduce a parameter $\rho$ that describes the probability with which the counting scheme can detect the $3_{10}$-helices. Finally, we designate a random variable $X$ to describe the number of recorded $3_{10}$-helices from a given primary structure. In this work, we propose to find its distribution and corresponding characteristics of the mean ($m_x$) and variance (Var$_x$), before verifying these theoretical predictions with actual profile of occurrence of rare secondary structures. Once functional, on utilitarian front, the present algorithm can help the researchers to estimate the number of $3_{10}$-helices, when the sequence information is provided. On the theoretical aspect, the ramification is deep. Success of this algorithm will imply that our interpretation of $\phi - \psi$ distribution (fig.-1) of $3_{10}$-helices was correct. Perhaps even more importantly, it will suggest that occurrence profile of $3_{10}$-helices on sequence can indeed be described reliably with Poisson distribution; which in its turn will imply that occurrence of $3_{10}$-helices on sequence is a randomly occurring event that take place purely 'by chance' (the hallmark of Poisson process).

Spectrum of results from various types of statistical surveys had provided us with reasonable indications that our hypothesis regarding Poisson nature of occurrence profile of $3_{10}$-helices (and its implications), to a large extent, is true. If it is so, one would expect that during evolution, proteins must somehow try to dispose the $3_{10}$-helices off. To put this hypothesis to strict and critical test, a thorough study of PDB structures classified by Pfam (release 24.0) [47] was undertaken. Pfam is a database based on hidden Markov model profiles (HMMs), which combines high quality and complete
protein domain families with high quality alignments. We selected sequences from top 20 Pfam families for which the structures are known. We wanted to observe the changes in the parts of sequence having $3_{10}$ and $\pi$ structural state in the pairs of proteins with different evolutionary distances within a family. Since such changes should be more clearly revealed in the proteins with different evolutionary distances within a family; Pfam provided an ideal framework for such an analysis.

As a part of phylogenetic analysis, an extensive (computational) analysis was conducted to count substitutions of a RSS on the sequence space. Although literature on contemporary Biology abounds with sequence-based evolution studies, a rigorous work that unambiguously lays down the criteria set to characterize substitutions with the help of evolutionary distance and alignment information, was difficult to find. In most of the studies on protein evolution, either the complete sequence, or the complete 3D structures or domains have been considered. This leaves us with little insight as to how secondary structures evolve and on the other hand, how the changing profile of structural parameters of secondary structures contribute in evolution of proteins as a whole. We propose a methodology to achieve the same in the present work. It assumes significance to mention here that our approach differs notably from any of previous studies on similar paradigm [48-53], not only in its basic motivation but also in the implementational procedures.

Evolutionary analysis were carried out on all the protein sequences drawn from twenty most prominent Pfam families. An elaborate set of comparisons were undertaken to observe the evolutionary trends in them, from the perspective of secondary structural elements. Although we could note the previous attempts to investigate this (rather difficult) paradigm [50, 54-56]; our approach was different both in its motivation and implementational details. In a two part scheme, we started by describing evolution of aforementioned primary structures with respect to: a) position-specific retainment of $3_{10}$-helices, b) position-specific replacement of $3_{10}$-helices, and c) position-specific introduction of new $3_{10}$-helices. In the next stage, the cases of position-specific replacement of $3_{10}$-helices were surveyed across the twenty most prominent Pfam classes, to obtain an idea about the proportion of transformation of $3_{10}$-helices to $\alpha$-helices. Since such analysis involved seven major structural classes of protein domains (SCOP), the present study could capture the entire paradigm from primary structures to protein domains, through the standpoint of secondary structures along evolution.

$\pi$-helices (characterized by the hydrogen-bonding [$i \rightarrow (i + 5)$] between residues) are even more rare in proteins than the $3_{10}$-helices[7, and references therein]; the number of studies on them is less too. Reasons for their energetically unfavorable structure is described masterly in a previous work[7]. All our aforementioned assertions about $3_{10}$-helices can as well be extended to the realm of $\pi$-helices on the primary structures too. Hence exactly the same methodology is applied to investigate the occurrence profile of $\pi$-helices too. In fact, if the mathematical and biophysical prerequisites are satisfied under certain biological contexts, one can even generalize the present set of reasons and algorithms to the entire set of other rare structural patterns too (poly-Proline helix, $\beta$-bulge, $\alpha$-turn, $\beta$-turn, $\gamma$-turn, $\pi$-turn, $\omega$-loop etc.). In the present work though, the results and implications of present algorithm is discussed with respect to it’s application to only $3_{10}$ and $\pi$-helices on the primary structure.

**Materials (Dataset collection and classification):**

We obtained our dataset of protein structures from Protein Data Bank. All the structures were obtained using advanced query feature of PDB and the domain sequences were classified in different classes according to structural classification proposed by SCOP (Structural Classification of Proteins). SCOP is a database of protein domains, accordingly we limited our analysis to the domains only, entire protein chains were not considered. Only true classes in SCOP were taken into consideration (all-$\alpha$, all-$\beta$, $\alpha + \beta$, $\alpha/\beta$, membrane, multidomain and small proteins). Analysis was performed in a SCOP
class-specific manner, to identify variabilities due to differences in the distribution of $3_{10}$ and $\pi$-helices in different protein folds. However, since the number of $\pi$-helices is small and statistically insignificant in any one SCOP class, class-specific analysis could not be performed for $\pi$-helices. We used a relaxed criteria (70% sequence identity, crystal structure resolution better than 3.5 Å) while collecting the sequences in any domain class to avoid missing out on any piece of information.

In each SCOP class, $3_{10}$ and $\pi$-helices were identified using DSSP algorithm since PDB assignments are known to be subjective and incomplete [57].

**Methodology**:

**Section -1) Mathematical backbone**

For the purpose of detection of $3_{10}$ and $\pi$-helices, we describe a random continuous part of the entire length of primary structure ($S$) by $s$. We proceed to find the conditional probability that $X = \gamma (0, 1, 2, ...)$. 

Thus, banking on the aforementioned background, we have:

$$P \{ X = \gamma \mid s \} = \frac{(\lambda ps)^\gamma}{\gamma!} e^{-\lambda ps}$$

Hence, the total probability of the event $\{X = \gamma\}$ is given by:

$$P \{ X = \gamma \mid s \} = \int_0^\infty \left( \frac{(\lambda ps)^\gamma}{\gamma!} e^{-\lambda ps} \mu e^{-\mu s} ds \right)$$

$$= \frac{\mu}{\gamma!} (\lambda p)\gamma \int_0^\infty \left( s^\gamma e^{-(\lambda p + \mu)s} \right) ds$$

$$= \frac{\mu}{\gamma!} (\lambda p)^\gamma (\lambda p + \mu)^{\gamma+1}$$

$$= \frac{\mu}{(\lambda p + \mu)} \left( \frac{\lambda p}{\lambda p + \mu} \right)^\gamma (\gamma = 0, 1, 2, ...).$$

This is a geometric distribution with parameter $\frac{\mu}{(\lambda p + \mu)}$, and therefore, the mean and variance of this distribution is given by:

$$\gamma_x = \left( \frac{\lambda p}{\lambda p + \mu} \right) \lambda p \mu$$

(1)

and,

$$Var_x = \left( \frac{\lambda p}{(\lambda p + \mu)} \right)^2 \frac{\lambda p(\lambda p + \mu)}{\mu^2} = \left( \frac{\lambda p}{\mu} \right)^2 + \frac{\lambda p}{\mu} = \gamma_x (\gamma_x + 1)$$

(2)
Section 2.1) Algorithmic Implementation:

Section 2.1.1) Algorithm to (statistically) characterize the occurrence profile of 3_{10} and \pi helices:

Crux of the algorithmic implementation of eq^n − 1 and eq^n − 2 rested with reliable detection of 3_{10}-helices. For this purpose, an inclusive criterion was chosen as the first step; only to be revised by a restrictive criterion in the second. A torsion angle range that suitably covers most of the 3_{10}-helices of the data-set was considered. Due to immense diversity of torsion angle range for 3_{10}-helices, a fairly large range \( \phi : (-40^\circ \text{ to } -90^\circ) \) and \( \psi : (-25^\circ \text{ to } -55^\circ) \) were considered. A run of at least 3 consecutive residues in the aforementioned torsion angle range, served as the filtering criterion (since most 3_{10}-helices are of length 3, or more residues).

Denoting the actual number of occurrence any of 3_{10}-helices in arbitrarily long \( s \) in \( S \), as \( \gamma \); and defining \( \mu = 1/\gamma \); \( \gamma \) and \( \mu \) were calculated. Since occurrences of 3_{10}-helices are rare and highly nonuniform, considering the entire sequence \( S \) in one go to detect the pattern in occurrence of 3_{10}-helices, may have introduced various coarse-graining type errors. Hence \( S \) was sub-divided into overlapping shorter sequences \( s \), where each \( s \) is statistically significant (if \( |S| = n \) and statistically significant length of \( s \) is \( r \) \((r \geq 32)\), then \((i = 0) \rightarrow r \), \((i = i + 1) \rightarrow r + 1 \), \ldots \((i = n - r + 1) \rightarrow n \)). In this manner, the entire sequence could be scanned, taking into consideration the (possible) local bias that may be associated with 3_{10}-helix occurrence. Calculations were repeated with various magnitudes of \( r \), to identify the (possible) latent bias. To not miss out on the overall picture, \( \lambda = (|\gamma|/|S|) \) is calculated, where \( \lambda \) denoted the average number of detection of 3_{10}-helices per unit traversal of sequence length using torsion angle assignment of 3_{10}-helices.

The actual number of 3_{10}-helices for every protein was calculated using DSSP; this is denoted by \( \gamma_s \) (DSSP). The \( \gamma_s \) (DSSP) is subsequently equated to \( p \times \left( \frac{\lambda}{\mu} \right) \), to find the correction parameter \( p \) for every protein. (Correction parameter is necessary to address the fact that certain 3_{10}-helices might be missed even with the best of the efforts to identify them with torsion angle range and some of the 3_{10}-helices may be incorrectly assigned using torsion angle range). Using the magnitude of \( p \) obtained in the last step, \( \gamma_s \) (torsion) is calculated for the test cases applying the formula \( p \times \left( \frac{\lambda}{\mu} \right) \). Finally, the magnitudes of \( \gamma_s \) (torsion) and \( \gamma_s \) (DSSP) are compared (for the entire data-set) to test the hypothesis.

Section 2.1.2) Algorithm to observe (statistical) trends in intervals between 3_{10} and \pi helices:

Further analysis were carried out to (statistically) model the patterns of sequence intervals between consecutive occurrences of 3_{10}-helices and \pi-helices. This study was essential because of the inconclusive nature of findings obtained from the torsion angle based study on the occurrence profile of the 3_{10}-helices. The inter-arrival sequence-distances between the observed occurrences of consecutive 3_{10}-helices (from DSSP assignment), for every protein containing 3_{10}-helix (resolution < 3Å) in non-redundant PDB, across all the SCOP classes, were studied. For each SCOP domain, the statistical distribution with best score of \(' \chi^2 \)-goodness of fit\' that models the inter-arrival sequence lengths between 3_{10}-helix occurrences was identified.

Analysis of the property set of statistical distributions often reveals latent trends in a dataset that are not detectable otherwise. Such analysis can be of enormous benefit when the causality behind the occurrence of certain events are difficult to ascertain. Our anticipation regarding the necessity of rigorous and exhaustive statistical analysis on occurrence profile properties of 3_{10}-helices, as carried out by the implementation of aforementioned algorithm, proved to be correct as the findings (kept in 'Results and Discussions' section) demonstrated.
Section 2.2) Phylogenetic Analysis:

2.2.1) Framework of the work:

Sequences were aligned and evolutionary distances between each pair of the families were computed using MEGA [58]. Distance calculation requires at least one common aligned site in the multiple sequence alignment. Hence, only the sequences with comparable lengths were considered. Results from a previous empirical study suggested that the substitution rate usually varies among amino acid sites during protein evolution; furthermore, this rate variation approximately follows the gamma distribution[59]. Since substitution rate also varies with amino acid pair, we computed JTT (Jones-Taylor-Thornton) distances[60] for evolutionary distance comparison. To ensure the best possible result, substitution rate was assumed to follow gamma distribution with shape parameter 2.4 [59].

Distance comparison analysis yielded a symmetric matrix of pairwise evolutionary distances for each family. For each pair of proteins in a family in the evolutionary distance matrix (it is an upper-triangular matrix excluding self), the number of substitutions of any 3_10-helix with any other secondary structure in the multiple sequence alignment of the proteins were counted. Suitable substitution criterion was decided after an extensive survey of the evolutionary distances and alignments (detailed in the next section).

2.2.2) Determination of substitution criterion:

Any one of the sequences from the pair under consideration was chosen as the reference sequence. Locations of the 3_10-helices (assigned by DSSP) were mapped on the sequence alignment. If any of the 3_10-helices in the reference sequence occurs within a 3 residue range in the other sequence, it was considered to be on the same sequence coordinates. This 3 residue buffer was considered merely to account for (possible) insertion/deletion in another region that might affect the position of the 3_10-helix under consideration. The event of absence of a 3_10-helix within this 3 residue range in the reference sequence, can be either due to a replacement of 3_10-helix or incorporation of a new 3_10-helix in one of the sequences. The decision to select between these two possibilities was taken on the basis of thorough study of evolutionary distance. Trends from the obtained results suggested that with the increase of evolutionary distance, the tendency to lose 3_10-helix in a pair of aligned proteins also increases. From a comprehensive survey of various distances, we found that the mean evolutionary distance can be used as a consistent and efficient parameter to choose between replacement and new RSS formation. As a logical continuation, when the evolutionary distance of the pair under consideration was found to be greater than the mean distance for that family, it was considered to be a case of replacement; whereas when it is less, a case of insertion of a new 3_10-helix was registered. To avoid multiple comparisons of the same 3_10-helices, comparisons were limited between 3_10-helices situated nearby on sequence coordinates. This procedure was repeated for all the pairs of proteins in the top 20 Pfam families, where at least one of the sequence contains at least one 3_10-helix.

The algorithm narrated above counts and characterizes the substitutions of a rare secondary structure (namely 3_10 and π-helix in the present case, but it can be applied to poly-Proline helix, β-bulge, α-turn, β-turn, γ-turn, π-turn, ω-loop etc.. too) on sequence space. We devised this strategy because we could not find any study in which such kind of rigorous evolutionary analysis has been performed. Although related analysis on sequence/structure evolution have been carried out earlier, motivations for those studies differed notably from the scope and depth of the present problem. Most importantly, a methodology that investigates similarity between entire 3D-domains across major Pfam sequences on a statistical scale, building upon thorough knowledge of the effects of evolution on secondary structures, was not found in existing works.
Results and Discussions:
Section -1

Results obtained from Methodology Section 2.1.1 with Discussion:
The comparison results (between the predicted value of $\gamma_x$ (mean of the distribution) from our algorithm and values provided by DSSP) are extremely interesting. These results (fig. 2)a-2)g) show that the trend of obtained values of $\gamma_x$ predicted from our algorithm. The ordinate magnitude in these figures describes the error (in probability units, between 0.00 to 1.00) in prediction, whereas the difference in abscissa describes the absolute number of cases where a particular error of has happened. (The sorted profile of these numbers, when used as axes, therefore, described how much of error was committed, for how many number of times, for a particular SCOP class.) While these trends can be observed to distinctly follow the actual occurrence profile of $3_{10}$-helices in proteins, they under-predict the $\gamma_x$ magnitude on $3_{10}$-helices consistently. A careful observation of the ordinate magnitudes shows that the maximum margin of error committed for all-$\alpha$, $\alpha/\beta$ proteins is (merely) 0.09 (probability) unit; for membrane proteins and small proteins it is 0.08 (probability) unit only; for $\alpha + \beta$ proteins, it is an even less, 0.06 (probability) unit; while for multidomain proteins the error is an (absolutely negligible) 0.05 (probability) unit. When applied on $\pi$-helices, this algorithm could predict the $\gamma_x$ magnitudes with extreme reliability (maximum error margin < 0.025 (probability) unit). Still, striving for further accuracy we wanted to ascertain the reason for the (small margin of) under-prediction. For this investigation, we randomly selected 20% of the proteins from our dataset manually, to repeat the experiment. This examination revealed that due to tremendous variation in the range of torsion angles in the $3_{10}$-helices (cutting across the SCOP classes), the filtering criterion for detection of $3_{10}$-helices (based on torsion angles) fell short in identifying them. Position-specific shifts in $\phi - \psi$ of the $3_{10}$-helices in proteins reported elsewhere[40], supports this argument. (Nonetheless, we could not find any algorithm to detect the $3_{10}$-helices with better efficiency from existing literature than our procedure.) Hence our program could not detect ~2%, ~4% and (7-9)% of the residues belonging to $3_{10}$-helices in the multidomain-proteins, $\alpha + \beta$ and membrane proteins and $\alpha/\beta$, all-$\alpha$ and small proteins, respectively. This consistent error in detection of residues belonging to $3_{10}$-helices explains the small yet measurable differences between 'predicted' and 'real' trends in (fig. 2)a-2)g).

Leaving the speculative thoughts regarding how closely the predicted patterns of occurrence profile of $3_{10}$-helices would have matched that in protein crystal structures, if the torsion angle based detection scheme for $3_{10}$-helices had worked properly, certain subtle trends from the obtained results (fig. 2)a-2)g) can clearly be noticed. The results, without the torsion angle related correction, tends to suggest that the basic pattern of the occurrence of $3_{10}$-helices on the primary structures can be reliably reproduced although the extents of such occurrences might not match. Reproduction of the basic trend in their occurrence (extremely well for multidomain proteins; moderately well for $\alpha + \beta$ and membrane proteins; a touch poorly for all-$\alpha$ and $\alpha/\beta$ and small proteins) opens up a debate with two-pronged possible explanations; viz. :

1) Some minute aspects of pure Poisson flow might not be pertinent when attempting to understand nature’s plan to place $3_{10}$-helices on the primary structure. This implies that although in some proteins $3_{10}$-helices occur as accidents and can well be considered as fossils of folding pathway exemplifying nature’s faulty (yet, quickly rectified) plan; in case of some other proteins, perhaps, nature associates some subtle yet definite plans behind the construction and placement of $3_{10}$-helices on the primary structure. While the first part of this argument forms the basis of the present work, tiny deviations of predicted magnitudes from the observed ones (for all-$\alpha$, $\alpha/\beta$ and small proteins), can possibly be explained with the second part of it.

2) If the same calculations are performed with an algorithm that detects $3_{10}$-helices reliably (as and when it is constructed), the predicted trends will converge to the actually observed trends; suggesting pointedly that the trends in occurrences of $3_{10}$-helices on primary structure are indeed accidental and rare; which, in turn, establish the hypothesis of possible mistakes of nature while making the protein fold, with $3_{10}$ and $\pi$-helices as impeccable examples of it.

Results presented under 'Result Section-2' resolves this debate.
Surprisingly, the detection scheme for $\pi$-helices did not suffer from such errors. The distribution of $\pi$-helices on the sequence axis (fig. - 2.3) provided a clear indication that their occurrences on the sequence axis takes place purely by chance and at random. As explained in the introduction section, such an observation indicates strongly that (probably), nature does not attach priority to the construction of $\pi$-helices. Poisson occurrence of $\pi$-helices hints (almost definitively) at implying that nature does not have a categorical plan in identifying what to do with them (either structurally or functionally) and they exemplify the errors committed by nature while making the primary structure fold. Probably the $\pi$-helices are true fossils of nature’s plan while constructing $\alpha$-helices.

Section-2 :
Results obtained from Methodology Section 2.1.2 with Discussion :
Although the principal cause of our algorithm’s subtle yet tangible under-prediction of $3_{10}$-helices could be established, in order to verify the basic hypothesis (that is, nature did not attach much of an importance with the construction of $3_{10}$-helices and $\pi$-helices; and probably, occurrences of these, merely depict nature’s mistakes and not her plan) we undertook a study to probabilistically model the patterns of sequence intervals between consecutive occurrences of $3_{10}$-helices and $\pi$-helices across all the SCOP domains of all the protein crystal structures (resolution < 3Å) contained in the non-redundant PDB set classified as per SCOP classes. This study was significant because of the inconclusive nature of findings obtained from the torsion angle based study on the occurrence profile of the $3_{10}$-helices.

Study of distribution of sequence intervals between primary structure coordinates for $3_{10}$-helix occurrences presents us with unexpected (with the existing knowledge of protein structure) and an enormously interesting set of data (described in Table-1). The distribution for all-$\alpha$ and membrane proteins had a $\chi^2$- best-fit with Johnson-SB distribution; for $\alpha/\beta$, $\alpha+\beta$ and multidomain proteins, the Weibull and exponential distributions served as ideal templates; whereas for the small-proteins, it was the log-Weibull distribution. Implications of these best-fit results are unexpected and deep. On the other hand, it provided an ideal construct to validate the data presented in the Section-1 of the Result.

| Class of Proteins      | Statistical distributions  |
|------------------------|-----------------------------|
| All-$\alpha$           | Johnson-SB Distribution     |
| Membrane               | Johnson-SB Distribution     |
| $\alpha + \beta$       | Exponential Distribution    |
| $\alpha/\beta$         | Weibull distribution        |
| Multi-domain           | Weibull distribution        |
| Small                  | Exponential Distribution    |
| $\pi$-helices(all classes) | Pareto Distribution     |

Table-1) : $\chi^2$-best-fit distribution to model inter-arrival sequence length between consecutive $3_{10}/\pi$ helices.

Exponential distribution routinely describes the lengths of the inter-arrival duration in a homogeneous Poisson process [44-46]. Weibull distribution is known to be a special case of exponential distribution. Exponential distribution's role as the best template for the $\chi^2$- best-fit for $\alpha+\beta$ domain of proteins and third best-fit for small proteins, point unambiguously towards a Poisson profile of occurrence of $3_{10}$-helices in the primary structure of $\alpha+\beta$ and small proteins. The Weibull distribution describes the phenomena space in-between that of exponential distribution and Rayleigh distribution [61,62], both of which are best known for their description of Poisson process related phenomena. The fact that $\alpha/\beta$
and multidomain proteins have Weibull distribution as the best-fit template for chi-square goodness of fit, suggests in an indirect (yet affirmative) way that the pattern in $3_{10}$'s occurrence on the primary structure in these structural families to be related to Poisson school of statistical distributions. These results implied further that the difference between the 'predicted' and 'experimentally observed' worms in figures [3-8] are purely due to failure of detection of $3_{10}$-helices in the primary structures and not due to erroneous and/or simplistic aspects in our basic hypothesis. (Cause behind $3_{10}$-helix detection failure with torsion angle based algorithm has been discussed beforehand).

However, (in the midst of all these Poisson school of random and rare occurrences) the all-$\alpha$ and membrane protein distribution profiles ($\chi^2$- best-fit distribution template 'Johnson SB', a distribution closely related to the classical normal distribution), points clearly to some prudent plan of nature behind constructing and placing the $3_{10}$-helices in these two structural families. Although an exact outline of this plan is difficult for us to envisage, the internal organization of helices within all-$\alpha$ and membrane proteins provides us with significant clues for that pursuit. The fact that the predicted occurrence profile of the $3_{10}$-occurrences profile for all-$\alpha$ and membrane proteins differed markedly (fig-3.a and fig-3.d, respectively), finds a support in these findings. It is interesting to note that a number of studies [63-65] have hinted at the presence of conducive environment for presence of helical structures in the membrane proteins. They (membrane proteins) constitute a structural class where there are fewer opportunities to destabilize the helical hydrogen bonds. Hence even if the preferred hydrogen bonding pattern, viz. $[(i+4) \rightarrow i]$ is not satisfied, the (strenuous) $[(i+3) \rightarrow i]$ hydrogen bonding pattern amongst residues can well be accommodated within the membrane environment. On the other hand, since the constraint of using solely the helical structures as building blocks for the all-$\alpha$ proteins(3.6 residues per turn with a translation of 1.5Å along the helix axis, hydrogen bonding pattern $[(i+4) \rightarrow i]$) can be daunting under every circumstances and since the transition to $3_{10}$-helices (they have 3 residues per turn with a translation of 2Å along the helix axis, hydrogen bonding pattern $[(i+3) \rightarrow i]$) - might not be extremely difficult to accommodate for small lengths of sequence, a non-trivial probability can be attached to the formation of $3_{10}$-helices within all-$\alpha$ domains, which may well show non-random and non-rare characteristics. Thus, although mutually different, the causality behind all-$\alpha$ and membrane protein's not following Poisson family of distributions (w.r.t occurrence profile of $3_{10}$-helices in them) - can well be understood from the framework of proposed logic of the present work. These results tend to imply that merely the lack of detection accuracy ($\sim$8% for all-$\alpha$ and $\sim$4% for membrane proteins) is not the sole reason for $3_{10}$-helix occurrence profile in these two structural classes to deviate from Poisson school of reasoning that govern the same in other classes.

A (near) perfect $\chi^2$-goodness of fit result for the sequence-interval profile of the $\pi$-helices could be obtained with the second order Pareto (lomax) distribution. Since Pareto distribution is a power-law distribution, it essentially models a stochastic process. Hence occurrences of $\pi$-helices in the primary structures tend to suggest that a stochastic process, instead of a deterministic process, can best describe their occurrence profiles. This finding bolsters our assertion (stated earlier) about them.

To summarize the results and at the same time, to compare and contrast them with the premise of starting hypothesis, we can enumerate the principal findings; as :
1) number of residues forming a part of $3_{10}$-helices are less than 3% of the total number of protein residues in non-redundant PDB structures.
2) not a single study could be found that reports a statistically significant number of cases (even, as low as five cases) where a $3_{10}$-helix could be found to perform a significant role in either stabilizing the protein or performing a tangible function, where an $\alpha$-helix would have failed to perform these jobs.
3) the occurrence profile of $3_{10}$-helices on the primary structure has been found to be irregular and non-repetitive for some protein structural classes to the extent of being almost random and rare for some and completely random and rare for some.
4) the torsion angle range for $3_{10}$-helices are farthest from being considered as structured.
5) lengths of $3_{10}$-helices are typically small, ranging from 3-5 residues on an average. Indeed all these arguments (with much more veracity) can be put forward to describe the case of occurrence of $\pi$-helices.

Section 3):

Results from phylogenetic analysis (Methodology section 2.2) with Discussion:
The hypothesis was put to test on top 20 Pfam families. The Pfam database links of protein sequence profiles with their structures wherever available. For two out of twenty families (PF01535, PF00361) structures were not available, while for another (PF07690) the number of structures were inadequate to perform any statistical test. For six families (PF00077, PF00516, PF00560, PF0023, PF00400, PF06817), the mean length was found to be very small and no $3_{10}$ and $\pi$ helix could be identified. Hence analysis was performed on the rest (PF00005, PF00078, PF00115, PF00072, PF00033, PF02518, PF00528, PF00069, PF00032, PF00106, PF00583).

Although the details of the results are provided in fig-4, fig-5.a, fig-5.b, and in Suppl. Mat. 1&2; here we report the significant trends. In six out of the eleven families (PF00005, PF00115, PF00072, PF02518, PF00069, PF00106, PF00583), the events of retention ($\hat{\text{Retn}}$) of a $3_{10}$-helix was found to be less than both the events of their replacement ($\hat{\text{Rplc}}$) and events of their new insertion ($\hat{\text{Insr}}$); no cases was observed where ($\hat{\text{Retn}}$) is found to be greater than ($\hat{\text{Rplc}}$) and ($\hat{\text{Insr}}$); while for three (PF00005, PF00033, PF00032) out of eleven families ($\hat{\text{Rplc}}$) outnumbered ($\hat{\text{Retn}}$) and ($\hat{\text{Insr}}$) both. These are significant findings (unforeseen hitherto). Together, they show nature’s apathy for retention of $3_{10}$-helices (as have been categorically demonstrated in the obtained results) across sequences of all the Pfam families of interest. The other observation regarding $3_{10}$ replacements outnumbering those with $3_{10}$ retentions and $3_{10}$ insertions in non-trivial number of cases, takes the non-retention trend to a logical finish. (Managers do not want the sacking of an inefficient employee merely, they want him to be sacked for a better replacement). Hence results obtained from this part of the analysis is found to be in complete accordance with our hypothesis that existence of $3_{10}$-helices as nature’s fault during protein folding. Interestingly, no correlation was found between SCOP classes with trends in the occurrence of anyone of ($\hat{\text{Rplc}}$), ($\hat{\text{Retn}}$) and ($\hat{\text{Insr}}$) (for example, while for PF00005 (the ABC transporter with structural domain $\alpha/\beta$) ($\hat{\text{Rplc}}$) > ($\hat{\text{Retn}}$) > ($\hat{\text{Insr}}$) is observed, for PF00106 (short chain dehydrogenase) with same structural domain $\alpha/\beta$ the trend reversed a little, viz. ($\hat{\text{Retn}}$) < ($\hat{\text{Rplc}}$) < ($\hat{\text{Insr}}$). Absence of any particular pattern confirms that the trends reported earlier in this paragraph are completely general and henceforth, vindicates our hypothesis on a global scale.

But results from the same analysis also reveals that in seven out of eleven classes (PF00078, PF00115, PF00072, PF02518, PF00069, PF00106, PF00583) ($\hat{\text{Insr}}$) magnitude is greater than ($\hat{\text{Rplc}}$) magnitude ( ($\hat{\text{Retn}}$, in any case, is lower than either of them). Although the maximum is a staggering 9231 out of 12795 cases for protein kinase domain ($\alpha + \beta$), for four out of seven Pfam classes the trend ($\hat{\text{Rplc}}$) ~ ($\hat{\text{Insr}}$) could be observed. Quite unambiguously, this is in contradiction to our hypothesis. However, an in-depth literature search and further analysis of the data resolves the crisis. Evolution is not an event but a process. The very fact that evolution does not proceed at a constant rate, but depends upon various types of environmental selective pressures on the individuals of any species at any given time is established conclusively in number of recent works[66-68]. The fact that one kind of secondary structure may experience replacements at a higher rate than another is known
too[69]. That being the case, evolution can well be compared to a unit-pipeline in a workshop set-up, where the supply-line can contain (occasional) faulty materials too; but all that is known is, during the formation of the finished product, these faulty elements will either be corrected to the efficient ones, or they would be thrown out of any further consideration. Viewing the $3_{10}$-helices from this perspective solves all the riddles in one unified way. Hence, in the present context, we hypothesized that while many $3_{10}$-helices are being constantly inducted into the quality control pipeline (that is process of evolution), many of them will either be replaced by $\alpha$-helices or loops, or else, they will be disposed off during the process. Selection of probable secondary structures for this hypothesis wasn’t difficult, in order to release the steric constraints, it is easier for a $3_{10}$-helix to transform itself to either $\alpha$-helix or a loop, than to undergo a large-scale rearrangement to achieve the same in the form of $\beta$-sheet. It is interesting to note that although there is no disallowed region of Ramachandran map between the preferred $\phi - \psi$ magnitude for $3_{10}$ and $\alpha$-helices[70], the $3_{10}$ block of $\phi - \psi$ is less favorable[71]. This smart quality control mechanism ensures a smooth $3_{10} \longrightarrow \alpha$ transition, but opposes $\alpha \longrightarrow 3_{10}$ transition. Our assertion from evolutionary analysis, viz. $3_{10}$-helices are unrealized possibilities in their route to become $\alpha$-helices, find support from the findings of some recent MD studies, where $3_{10}$and $\pi$-helices are typified as “transient” and “defective” $\alpha$-helices[72-74].

Results obtained from this analysis (Table #) vindicated the correctness of our hypothesis, completely. In all the eleven Pfam classes, $3_{10}$-helices are found to be transforming into either $\alpha$-helices or loops. Dominant $3_{10} \longrightarrow \alpha$ transformation was found in five out of eleven Pfam classes, whereas for another five, a dominant $3_{10} \longrightarrow$ loop transformation could be observed. In a single case(PF00072, response regulator receiver domain, possessing interestingly, an all-$\beta$ structural domain) $3_{10}$-helices could be observed to prefer $\alpha$-helices and loops equally.

**Conclusion:**
We have shown here that results from three different and rigorous investigations, with contrasting algorithms, converge to confirm the basic hypothesis of ours; namely, nature does not attach any well-cut-out plans with the construction of $3_{10}$ and $\pi$-helices. We think that the apparently disparate array of observations regarding various known facts about $3_{10}$-helices (detailed in introduction section), can be explained from a common basic platform if it is hypothesized that they symbolize nature’s mistake during its attempt to make the primary structure fold to its native state. Of course these mistakes take place rarely, which explains the rare occurrences of $3_{10}$-helices. Going back to an analogy used previously, one can view nature as a manager in a hurry while making the primary structure fold; probably that is why the temptation of constructing the first hydrogen bond with one fewer unit than that required to construct an $\alpha$-helix, scores better of him. But before long, the smart manager realizes his(her, at any rate) mistake. So he drops the plan to persist with the mistake and hence the faulty scheme of hydrogen bonds are stopped from becoming longer. Since no particular plan was associated with the construction of these structures, no notable example of their involvement in either stabilizing the protein or helping it with certain functionality can be found too. The Ramachandran-map for these structures, not surprisingly, demonstrates the widest possible spectrum of $\phi - \psi$ combinations. Such (wild) variability can never be observed for the $\phi - \psi$ variability range for $\alpha$-helices or $\beta$-sheets; structures with which nature attaches definite importance in lending the proteins with stability and/or functionality. The very fact that barring ignorable tendencies in minuscule number of cases, occurrence profiles of $3_{10}$-helices across all the protein structural families, tend to follow a Poisson family of distribution; suggests that their occurrence in primary structure is indeed random and rare; that is, without a specific plan. Finally, extensive evolutionary analysis shows how the new insertions of $3_{10}$-helices, do not contradict our hypothesis and how all the top Pfam classes show no interest in retaining the $3_{10}$-helices and how nature makes it a point to transform $3_{10}$-helices to $\alpha$-helices and loops. - This entire body of evidences, taken as a whole, seem to point to the definite (non-accidental) inference of nature’s committing mistakes when constructing the $3_{10}$ and $\pi$-helices.
We attempted to address a basic question in this work. Instead of justifying the stance of nature on the construction and placement of $3_{10}$ and $\pi$-helices with isolated cases and accidental observations that conforms to no general pattern; we questioned, with what probability, can we assert that they exemplify nature's mistakes while performing protein folding? It turned out that our (somewhat sacrilegious) question did find a reasonable ground when describing the occurrence profile of $3_{10}$-helices and $\pi$-helices on the primary structure. The dominating trend of results across the various structural classes, from statistical, mathematical and evolutionary aspects, suggested that these (rare) secondary structures should be viewed as evidences of nature's mistakes with regard to making the primary structure fold to the native state of a folded protein.

Acknowledgment: One of the authors, Anirban, would like to thank the COE-DBT scheme for supporting him during the tenure of this work. He would also like to thank the Director of Bioinformatics Center, University of Pune; Dr. Urmila Kulkarni-Kale, for her support during the work, though this work is not a part of his PhD project.

References:
[1] Huggins ML. 1943. The structure of fibrous proteins. Chem Rev 32:195-218.
[2] Pauling L, Corey RB, Branson HR. 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 37:205-211.
[3] Lipika Pal, Pinak Chakrabarti and Gautam Basu; Sequence and Structure Patterns in Proteins from an Analysis of the Shortest H elices: Implications for Helix Nucleation; J. Mol. Biol. (2003) 326, 273-291.
[4] D.J. Barlow, J.M. Thornton; Helix geometry in proteins; J. Mol. Biol.; 201 (1988) 601-619.
[5] Karpen M.E., Haseth P.L.D, Neet K.E.; Differences in the amino acid distributions of $3_{10}$-helices and $\alpha$-helices; Protein Science (1992), I, 1333-1342.
[6] Janani Venkatraman, Sasalu C. Shankaramma, and Padmanabhan Balaram; Design of Folded Peptides; Chem. Rev. 2001, 101, 3131-3152.
[7] CAROL A. ROHL AND ANDREW J. DOIG; Models for the 310-helix/coil, Pi-helix/coil, and a-helix/310-helix/coil transitions in isolated peptides; Protein Science (1996), 5:1687-1696.
[8] L.J. Smith, K.A. Bolin, H. Schwalbe, M.W. MacArthur, J.M. Thornton, C.M. Dobson; Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations, J. Mol. Biol. 255 (1996) 494-506.
[9] Y.D. Wu, Y. Zhao, A theoretical study on the origin of cooperativity in the formation of $3_{10}$ and a-helices, J. Am. Chem. Soc. 123 (2001) 5313-5319.
[10] R. Gratias, R. Konat, H. Kessler, M. Crisma, G. Valle, A. Polese, F. Formaggio, C. Toniolo, Q.B. Broxterman, J. Kamphuis; First step towards the quantitative identification of peptide $3_{10}$-helix conformation with NMR spectroscopy : NMR and X-ray diffraction structural analysis of a fully-denatured $3_{10}$-helical peptide standard; J. Am. Chem. Soc. 120 (1998) 4763-4770.
[11] Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968;23:283-439.
[12] Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 1984;44:97-179.
[13] Tirado-Rives J, Maxwell DS, Jorgensen WL. Molecular dynamics and Monte Carlo simulations favor the alpha-helical form for alanine-based peptides in water. J Am Chem Soc 1993;115:11590.
[14] Smythe ML, Huston SE, Marshall GR. Free energy profile of a 310-to a-helical transition of an oligopeptide in various solvents. J Am Chem Soc 1993;115:11594.
[15] Zhang L, Hermans J. 310 helix versus a-helix: a molecular dynamics study of conformational preferences of aib and alanine. J Am Chem Soc 1994;116:11915.
[16] Millhauser GL. Views of helical peptides: a proposal for the position of 310-helix along the ther-
[17] Mück MS, Martinez GV, Fiori WR, Todd AP, Millhauser GL. Short alanine-based peptides may form 310-helices and not α-helices in aqueous solution. Nature 1992; 359:633.

[18] Andrei L. Lomize, Henry I. Mosberg; Thermodynamic Model of Secondary Structure for alpha-Helical Peptides and Proteins; Biopolymers. 1997 Aug;42(2):239-269.

[19] Aleksandr V. Mikhonin and Sanford A. Asher; Direct UV Raman Monitoring of 310-Helix and p-Bulge Premelting during alpha-Helix Unfolding; J. AM. CHEM. SOC. 2006, 128, 13789-13795.

[20] Glenn L. Millhauser, Chris J. Stenland, Kimberly A. Bolin and Frank J. M. van de Ven; Estimating the Relative Populations of 310-helix and α-helix in Ala-rich Peptides: A Hydrogen Exchange and High Field NMR Study; J. Mol. Biol. (1997) 267, 963-974.

[21] Sudha, T. S., Vijayakumar, E. K. S., and Balaram, P. Circular dichroism studies of helical oligopeptides: can 310 and α-helical conformations be chiroptically distinguished? (1983) Int. J. Pept. Protein Res. 22, 464-468.

[22] Sheinerman, F.B., Brooks, C.L. (1995). 310-helices in peptides and proteins as studied by modified Zimm-Bragg theory. J. Am. Chem. Soc. 117, 10098-10103.

[23] Gerstein, M., Chothia, C. (1991). Analysis of protein loop closure: two types of hinges produce one motion in lactate dehydrogenase. J. Mol. Biol. 220, 133-149.

[24] McPhalen, C. A., Vincent, M. G., Picot, D., Jansonius, J. N., Lesk, A. M., Chothia, C. (1992). Domain closure in mitochondrial aspartate aminotransferase. J. Mol. Biol. 227, 197-213.

[25] Tirado-Rives, J., Jorgensen, W. L. (1991). Molecular dynamics simulations of the unfolding of an α-helical analogue of ribonuclease A S-peptide in water. Biochemistry, 30, 3864-3871.

[26] Sung, S. S. (1995). Folding simulations of alanine-based peptides with lysine residues. Biophys. J. 68, 826-834.

[27] Worthylake DK, Wang H, Yoo S, Sundquist WI, Hill CP; Structures of the HIV-1 capsid protein dimerization domain at 2.6 Å resolution; Acta Crystallogr D Biol Crystallogr. 1999;55(Pt 1):85-92.

[28] Tani K, Hiroaki Y, Fujiyoshi Y.; Aquaporin-4: Rinsho Shink eigaku. 2008; 48(11):941-944.

[29] Manjasetty BA, Niesen FH, Scheich C, Roske Y, Goetz F, Behlke J, Sievert V, Heinemann U, Büßow K; X-ray structure of engineered human Aortic Preferentially Expressed Protein-1 (APEG-1); BMC Struct Biol. 2005 Dec 14;5:21.

[30] Hara T, Kato H, Katsube Y, Oda J.; A pseudo-michaelis quaternary complex in the reverse reaction of a ligase: structure of Escherichia coli B glutathione synthetase complexed with ADP, glutathione, and sulfate at 2.0 Å resolution. Biochemistry. 1996; 35(37):11967-11974.

[31] Worthylake DK, Prakash S, Prakash L, Hill CP.; Crystal structure of the Saccharomyces cerevisiae ubiquitin-conjugating enzyme Rad6 at 2.6 Å resolution; J Biol Chem. 1998; 273(11):6271-6276.

[32] Xu RM, Jokhan L, Cheng X, Mayeda A, Kainer AR.; Crystal structure of human UP1, the domain of hnRNP A1 that contains two RNA-recognition motifs. Structure. 1997;5(4):559-570.

[33] Li T, Horan T, Osslund T, Stearns G, Arakawa T.; Conformational changes in G-CSF/Receptor complex as investigated by isotope-edited FTIR spectroscopy; Biochemistry. 1997; 36(29):8849-8857.

[34] Tirado-Rives J, Jorgensen WL. Molecular dynamics simulations of the unfolding of aprotinin in water. Biochemistry 1993;32:4175-4184.

[35] Freedberg DI, Venable RM, Rossi A, Bull TE, Pastor RW. Discriminating the helical forms of peptides by NMR and molecular dynamics simulation. J. Am Chem Soc.; 2004; 126 : 10478-10484.

[36] Sung S, Wu X. Molecular dynamics simulations of synthetic peptide folding. Proteins 1996;25:202.

[37] Shirley WA, Brooks CL 3rd. Curious structure in canonical alanine-based peptides. Proteins: Structure, Function, and Bioinformatics 64:691-699 (2006).
[41] P. Kolandaivel, P. Selvarengan, K.V. Gunavathy; Structure and potential energy surface studies on 310-helices of hen egg white lysozyme and Phaseolus vulgaris arcelin-1 proteins; Biochimica et Biophysica Acta 1764 (2006) 138-145.
[42] H.M. Berman, K. Henrick, H. Nakamura Announcing the worldwide Protein Data Bank. Nature Structural Biology 10 (12), p. 980 (2003).
[43] Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.
[44] Kovalenko I.N., Kuznetsov N.Yu. and V.M. Shurenkov; Models of random processes: a handbook for mathematicians and engineers; 1996; Boca Raton: CRC Press, pp 75-78.
[45] Balakrishnan, N. and Basu, A. P. The Exponential Distribution: Theory, Methods, and Applications. New York: Gordon and Breach, 1996.
[46] Gupta, R. D. and Kundu, D. (2001), Exponentiated exponential family: an alternative to gamma and Weibull, Biometrical Journal, vol. 43, 117 - 130.
[47] R.D. Finn, J. Tate, J. Mistry, P.C. Coggill, J.S. Sammut, H.R. Hotz, G. Ceric, K. Forslund, S.R. Eddy, E.L. Sonnhammer and A. Bateman; The Pfam protein families database: Nucleic Acids Research (2008) Database Issue 36:D281-D288.
[48] JL Thorne, N Goldman and DT Jones; Combining protein evolution and secondary structure; Molecular Biology and Evolution, Vol 13, 666-673, 1996.
[49] Lio, P., Goldman, N., Thorne, J.L. and Jones, D.T. (1998) PASSML: combining evolutionary inference and protein secondary structure prediction. Bioinformatics 14:726-733.
[50] Fornasari MS, Parisi G, Echave J: Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. Mol Biol Evol 2002, 19:352-356.
[51] Matthias Haimel, Karin Pro, Michael Rebhan; ProteinArchitect: protein evolution above the sequence level; 2009; PLoS ONE 4(7): e6176. doi:10.1371/journal.pone.0006176.
[52] Babajide A, Farber R, Hofacker I, Inman J, Lapedes A, Stadler P, Exploring protein sequence space using knowledge-based potentials. J Theor Biol 2001, 212:35-46.
[53] Dokholyan NV, Shakhnovich EI, Understanding hierarchical protein evolution from first principles, J Mol Biol 2001, 312:289-307.
[54] Lio, P., Goldman, N., Thorne, J.L. and Jones, D.T. (1998) PASSML; combining evolutionary inference and protein secondary structure prediction. Bioinformatics 14:726-733
[55] Nick Goldman, Jeffrey L. Thorne, and David T. Jones; Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution; Genetics 149: 445-458, 1998.
[56] Juliette TJ Lecomte, David A Vuletich and Arthur M Lesk; Structural divergence and distant relationships in proteins: evolution of the globins; Current Opinion in Structural Biology 2005, 15:290-301.
[57] Wolfgang Kabsch and Chris Sander; Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637.
[58] Taniura K, Dudley J, Nei M & Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1396-1399.
[59] Masatoshi Nei, Jianzhi Zhang. ; Evolutionary distance estimation. ENCYCLOPEDIA OF LIFE SCIENCES. 2005.
[60] Jones DT, Taylor WR and Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences 8: 275-282.
[61] Johnson, Norman L.; Kotz, Samuel; Balakrishnan, N. (1994), Continuous univariate distributions. Vol. 1, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics (2nd ed.), New York: John Wiley & Sons.
[62] Weibull, W. (1951), A statistical distribution function of wide applicability, J. Appl. Mech.-Trans. ASME 18 (3): 293-297.
[63] H. Luecke, B. Schober, H.-T. Richter, J.-P. Cartailleur, J.K. Lanyi, Structure of bacteriorhodopsin at 1.55 A resolution, J. Mol. Biol. 291 (1999) 899-911.
[64] S. Kim, T.A. Cross, Uniformity, ideality, and hydrogen bonds in transmembrane a-helices, Biophys. J. 83 (2002) 2084-2095.
[65] Sanguk Kim and T.A. Cross; 2D solid state NMR spectral simulation of 310, alpha and pi-helices; Journal of Magnetic Resonance 168 (2004) 187-193.

[66] Roman A. Laskowski, Janet M. Thornton and Michael J.E. Sternberg; Protein Evolution: Sequences, Structures and Systems; Biochem. Soc. Trans. (2009) 37, 723-726.

[67] Tawfik, D.S. (2009) Protein evolution: a reconstructive approach, http://www.biochemistry.org/Portals/0/Conferences/abstracts/SA099/SA099S007.pdf (Abstract).

[68] Studer, R.A. and Robinson-Rechavi, M. (2009) Evidence for an episodic model of protein sequence evolution. Biochem. Soc. Trans. 37, 783-786.

[69] JL Thorne, N Goldman and DT Jones; Combining protein evolution and secondary structure; Molecular Biology and Evolution, Vol 13, 666-673, 1996.

[70] Toniolo C, Benedetti E. 1991. The polypeptide 310-helix. Trends Biochem Sci 16:350-353.

[71] Michael Feig, Alexander D. MacKerell, Jr., and Charles L. Brooks, 3rd; Force Field Influence on the Observation of Pi-Helical Protein Structures in Molecular Dynamics Simulations; J. Phys. Chem. B 2003, 107, 2831-2836.

[72] Aleksandr V. Mikhonin and Sanford A. Asher; Direct UV Raman Monitoring of 310-Helix and p-Bulge Premelting during alpha-Helix Unfolding; J. AM. CHEM. SOC. 2006, 128, 13789-13795.

[73] Sorin, E. J.; Rhee, Y. M.; Shirts, M. R.; Pande, V. S. J. Mol. Biol. 2006, 356, 248-256.

[74] Freedberg, D. I.; Venable, R. M.; Rossi, A.; Bull, T. E.; Pastor, R. W. J. Am. Chem. Soc. 2004, 126, 10478-10484.

Figures and Legends:

Figure Title 1)a) : The Ramachandran Map of 310-helices.
Legend for Figure 1)a) : Ramachandran-Map of all the 310-helices present in non-redundant PDB.

Figure Title 1)b) : The Ramachandran Map of π-helices.
Legend for Figure 1)a) : Ramachandran-Map of all the π-helices present in non-redundant PDB.
A - F: (3/10 helices) for different SCOP classes
A: All-Alpha Proteins
B: Alpha / Beta
C: Alpha + beta
D: Membrane Proteins
E: Multidomain Proteins
F: Small Proteins

G: pi Helices

X-axis: Absolute number of cases with a particular magnitude of probability
Y-axis: Absolute magnitude of probability
Figure Title 2: Observed-vs-Predicted occurrence of $3_{10}$ and $\pi$-helices

Legend for Figure 2: For the concerned class of protein, the number of $3_{10}$ and $\pi$-helices in its crystal structure were counted. For two points chosen arbitrarily, the graph can be interpreted as the difference in probability with which certain number of $3_{10}$-helices were predicted and certain were observed. Say, two points with ordinate: 0.02 and abscissa: 33 (predicted) and 43 (observed), implies that there are 10 (43-33=10) cases where the difference between magnitude of observed probability of occurrence $3_{10}$-helices and predicted probability of occurrence $3_{10}$-helices, is 0.02 probability units. In other words, the ordinate magnitude describes a particular magnitude of probability; whereas, the abscissa denotes the number of cases with that particular probability. Study was carried out on all the non-redundant PDB structures across all the SCOP classes.
A - F: (i/i0 helices) for different SCOP classes
A: All-Alpha Proteins
B: Alpha / Beta
C: Alpha + beta
D: Membrane Proteins
E: Multidomain Proteins
F: Small Proteins
G: pi Helices

X-axis: Sequence interval between 3-10 helices
Y-axis: No. of 3-10 helices with particular sequence interval
Figure Title 3) : Modeling the inter-arrival interval for $3_{10}$ and $\pi$-helices on all the primary structures
Legend for Figure 3: Occurrence of $3_{10}$-helices on the primary structure do not follow a regular and repetitive pattern. In fact these occurrences appear to be random. The sequence intervals between consecutive occurrences of $3_{10}$-helices (abscissa) are plotted against the number of $3_{10}$-helices with a particular length of sequence interval magnitude (ordinate). Absolute number of $3_{10}$-helices with certain sequence-interval was normalized by the maximum length of the primary structure with at least two $3_{10}$-helices, from the entire non-redundant PDB. Plots have been generated for every SCOP class containing $3_{10}$-helices. Data fitting is done with the software 'R'.

Figure Title 4) : Distribution of replacement, retention and insertion across evolutionary distance
Legend for Figure 4: Distributions of substitution, retention and insertion over evolutionary distances across families, showed that absolute number of replacements and new $3_{10}$-helix formation is almost same. However, their trends show that with less evolutionary distance(< 4-5), no. of retained and new formations are more than no. of substitutions; with no. of new formations being significantly higher. But as evolutionary distances increases no. of new and retained $3_{10}$-helix decreases drastically and no. of substitutions increases. This result validates our methodology (to count substitutions and replacements) and hypothesis as it is clear that event of retention drastically decrease increasing evolutionary distance.
Figure Title 5) : Results of evolutionary analysis
Legend for Figure 5)a) : Describes the number of cases of replacement, retention and insertion of 3_10-helices across 20 top Pfam classes.
Legend for Figure 5)b) : Surveying the replacement cases further, this figure describes the cases where 3_10-helices are being replaced by α-helices, β-sheets and loops, across 20 top Pfam classes.