An updated evolutionary study of the Notch family, reveals a new ancient origin and novel invariable motifs as potential pharmacological targets

Dimitrios Vlachakis*, 1, 2, 3, Louis Papageorgiou 4, Ariadne Papadaki 1, Maria Georga 1, Sofia Kossida 5, Elias Eliopoulos 5

1 Laboratory of Genetics, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, Athens, Greece
2 University Research Institute of Maternal and Child Health & Precision Medicine, and UNESCO Chair on Adolescent Health Care,”Aghia Sophia” Children’s Hospital, National and Kapodistrian University of Athens, Athens, Greece
3 Division of Endocrinology and Metabolism, Center of Clinical, Experimental Surgery and Translational Research, Biomedical Research Foundation of the Academy of Athens, Athens, Greece
4 Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
5 IMGT, the international ImMunoGeneTics information system, Université de Montpellier, Laboratoire d'ImmunoGénétique Moléculaire and Institut de Génétique Humaine, University of Montpellier, Montpellier, France

Corresponding Author: Dimitrios Vlachakis
Email address: dimvl@icloud.com

Notch family proteins play a key role in a variety of developmental processes by controlling cell fate decisions and operating in a great number of biological processes in several organ systems, such as hematopoiesis, somatogenesis, vasculogenesis, neurogenesis and homeostasis. The Notch signaling pathway is crucial for the majority of developmental programs and regulates multiple pathogenic processes. Notch family receptors’ activation has been largely related to its multiple effects in sustaining oncogenesis. The Notch signaling pathway constitutes an ancient and conserved mechanism for cell to cell communication. Much of what is known about Notch family proteins function comes from studies done in Caenorhabditis Elegans and Drosophila Melanogaster. Although, human Notch homologs had also been identified, the molecular mechanisms which modulate the Notch signaling pathway remained substantially unknown. In this study, an updated evolutionary analysis of the Notch family members among 603 different organisms of all kingdoms, from bacteria to humans, was performed in order to discover key regions that have been conserved throughout evolution and play a major role in the Notch signaling pathway. The major goal of this study is the presentation of a novel updated phylogenetic tree for the Notch family as a reliable phylogeny "map", in order to correlate information of the closely related members and identify new possible pharmacological targets that can be used in pathogenic cases, including cancer.
An updated evolutionary study of the Notch family, reveals a new ancient origin and novel invariable motifs as potential pharmacological targets.

Dimitrios Vlachakis¹,²,³,* Louis Papageorgiou⁴, Ariadne Papadaki¹, Maria Georga¹, Sofia Kossida⁵, Elias Eliopoulos¹,

1. Laboratory of Genetics, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, Athens, Greece
2. University Research Institute of Maternal and Child Health & Precision Medicine, and UNESCO Chair on Adolescent Health Care, “Aghia Sophia” Children's Hospital, National and Kapodistrian University of Athens, Athens, Greece
3. Division of Endocrinology and Metabolism, Center of Clinical, Experimental Surgery and Translational Research, Biomedical Research Foundation of the Academy of Athens, Athens, Greece
4. Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
5. IMGT, the international ImMunoGeneTics information system, Université de Montpellier, Laboratoire d'Immunogénétique Moléculaire and Institut de Génétique Humaine, University of Montpellier, France

*Correspondence to: Dr Dimitrios Vlachakis, dimvl@aua.gr
Abstract
Notch family proteins play a key role in a variety of developmental processes by controlling cell fate decisions and operating in a great number of biological processes in several organ systems, such as hematopoiesis, somatogenesis, vasculogenesis, neurogenesis and homeostasis. The Notch signaling pathway is crucial for the majority of developmental programs and regulates multiple pathogenic processes. Notch family receptors’ activation has been largely related to its multiple effects in sustaining oncogenesis. The Notch signaling pathway constitutes an ancient and conserved mechanism for cell to cell communication. Although, human Notch homologs had also been identified, the molecular mechanisms which modulate the Notch signaling pathway remained substantially unknown. In this study, an updated evolutionary analysis of the Notch family members among 603 different organisms of all kingdoms, from bacteria to humans, was performed in order to discover key regions and important motifs that have been conserved throughout evolution and play a major role in the Notch signaling pathway. The major goal of this study is the presentation of a novel updated phylogenetic tree for the Notch family as a reliable phylogeny "map", in order to correlate information of the closely related members and identify new possible pharmacological targets that can be used in pathogenic cases, including cancer.

Introduction
The Notch gene was originally discovered by Dexter in 1914 and it was named from the irregular notched wing phenotype of Drosophila melanogaster, caused by the loss-of-function in the responsible gene’s after a point mutation (Dexter, 1914). Since then, Notch protein and its homologs, Notch1, Notch2, Notch3, Notch4, LIN-12 and GPL-1 have been identified in genomes from all kingdoms, indicating the progressive differentiation of Notch family. Their length varies from ≈110 amino acids in bacteria (Durieux, et al., 2019) to ≈4500 amino acids (aa) in animals (Fairclough, et al., 2013). Notch family members are evolutionary conserved, type-1 transmembrane glycoproteins, that function both as transmembrane receptors for ligands and transcription factors (Kopan and Ilagan, 2009). They regulate cell fate determination and promote cell differentiation, maintenance and survival. These proteins have either overlapping or unique cellular functions, but these functions remain quite unclarified, in the majority of the organisms found (Hogan and Bautch, 2004).

In mammals, there are four genes that encode four parologue Notch transmembrane receptors, Notch1 to 4. The Notch1 gene is essential for developmental processes, while loss-of-function mutations in result to early fetal death, due to dysfunctional angiogenesis, organogenesis and cardiology. Moreover, it plays a key role in the definitive formation of Hematopoietic Stem Cells (HSCs), responsible for the production of all mature blood cells during adulthood (Tanigaki and Honjo, 2007). Notch1 protein consists of approximately 2627 aa and its signaling pathway regulates the development of B and T lymphocytes (Gerhardt, et al., 2014). The Notch2 gene encodes a ≈2471aa receptor and determines the cell fate in the heart, liver, kidneys, teeth, bones as well as other cell types that are being developed in the fetus. After birth, Notch2 signaling is involved in the immune system’s function, tissue repair, homeostasis and bone reshaping when needed. The Notch3 gene encodes a ≈2321aa receptor, which determines the fate of Vascular Smooth Muscle Cells (VSMCs) in the arterial network of the brain and finally the Notch4 gene
encodes a \( \approx2059\text{aa} \) receptor, that determines fetal vascular morphogenesis and remodeling (Krebs, et al., 2000; Vlachakis, et al., 2014).

Notch family protein domains have been conserved throughout evolutionary history, from invertebrates to humans (Hori, et al., 2013). They are composed of an extracellular domain (Notch extracellular domain, NECD), a transmembrane domain and an intracellular domain (Notch intracellular domain, NICD)(Yavropoulou, et al., 2015). The NECD contains 29 to 36 Epidermal Growth Factor-like Repeats (EGF-like domain), depending on the type of receptor and a Negative Regulatory Region (NRR). The NRR is composed of three cysteine-rich Notch / LIN-12 repeats (LNRs) and a Heterodimerization Domain (HD). Each EGF-like repeat has 6 cysteines, which form 3 disulfide bonds, contributing to the 3D structure of the protein (Muiño, et al., 2017). The NICD has a RBPJk-associated molecule domain (RAM), nuclear localization sequences (NLS), 7 ankyrin repeats (ANK) domain, a transcriptional activation domain (TAD) and a C-terminal Pro Glu Ser Thr (PEST) domain (Hori, et al., 2013). Although Notch receptors are highly conserved, they have some structural variation mainly in the number of EGF-like repeats, in the presence of the TAD domain, and the length of the segment between the ANK repeats and C-terminal (Sander, et al., 2006). Notch signaling pathway is highly conserved, with a key role in cell-cell communication. It regulates the vascular development and physiology, as well as multiple developmental processes of the Central Nervous System (CNS). The pathway is involved in the regulation of Nerve Stem Cells’ (NCSs) proliferation, survival, self-renewal and differentiation and has been associated with early neurodevelopment, learning and memory, as well as neurodegeneration (Mizutani, et al., 2007; Polychronidou, et al., 2015). Thus, mutations in Notch pathway participants, that lead to defective signaling, cause a variety of human diseases including neurodegenerative diseases, developmental disorders and cancer (Table 1).

**Methods**

**Dataset collection and filtering**

Data was collected from the NCBI database (ncbi.nlm.nih.gov) as previously described in Mitsis et al. (2020), towards to extracting the amino acid sequences that are related to the Notch proteins using the keyword “Notch” (Mitsis, et al., 2020). Protein sequences that responded to the query but did not include the Notch family members were eliminated from the primary dataset, by using related keywords and regular expressions techniques in the header information, and local alignments with reference protein sequences. Furthermore, a final dataset for each species class was produced by using internal protein alignments and protein identity score. Duplicated protein sequences in each species that were found share 95% > protein identity within the dataset were removed. In total, 25,761 Notch family related protein sequences were identified from several species, and a dataset containing 603 unique, non-duplicate protein sequences was created (Sup. Dataset 1).

**Multiple sequence alignment**

Multiple sequence alignment (MSA) was executed using the MATLAB Bioinformatics Toolbox, utilizing a guide tree and the progressive MSA method as previously described in several studies (Mitsis, et al., 2020; Papageorgiou, et al., 2016; Sobie, 2011). Pairwise distances among sequences were estimated based on the pairwise alignment with the “Gonnet” method and followed by
calculating the differences between each pair of sequences. The Neighbor-Joining method was
used towards to estimating the guide tree by assuming equal variance and independence of
evolutionary distance estimates (Sup. Dataset 2). Finally, consensus sequence was calculated and
visualized through the JalView platform (Waterhouse, et al., 2009) using the multiple sequences
alignment results and parameters including amino acid conservation. The commentary section of
Jalview, which presents the amino-acid conservation using logos and histograms, was further
observed to uncover innovative motifs.

Notch family protein clusters

Notch family protein clusters were identified by using phylogenetic analysis. The phylogenetic
analysis was performed using the MATLAB Bioinformatics Toolbox (Kufareva and Abagyan,
2012) utilizing the Unweighted Pair-Group Method (UPGMA) (Michener and Sokal, 1957;
Pavlopoulos, et al., 2010; Sneath, 1973) while the matrix of the pairwise distances was calculated
using the protein-adapted Jukes-Cantor statistical method (Papageorgiou, et al., 2016; Yang and
Zhang, 2008). The constructed phylogenetic tree visualized using MEGA radiation option and the
final Notch family protein clusters separated in different sub-datasets using a threshold. In total,
eight major sub-clusters including Notch Bacteria, Notch Plants, Notch Invertebrates,
Notch Protist, Notch1, Notch2, Notch3 and Notch4 were identified (Sup. Dataset 3).

Consensus sequences and a specialized phylogenetic analysis

Representative consensus protein sequences of the Notch family sub-clusters were estimated
according to the clustering results of the Notch family protein clusters (Mitsis, et al., 2020). The
representative protein sequences for each sub-cluster were calculated using the MATLAB
Bioinformatics Toolbox (Nanni, et al., 2014). Two representative protein sequences of the GLP1
and LIN12 homologs of the Caenorhabditis elegans species (Girard, et al., 2007; Sorkac, et al.,
2018) were also included in the final dataset of the consensus protein sequences (Sup. Dataset 4).
Last but not least, a specialized phylogenetic analysis was performed using the MATLAB
Bioinformatics Toolbox (Kufareva and Abagyan, 2012) utilizing the Unweighted Pair-Group
Method (UPGMA) (Michener and Sokal, 1957; Pavlopoulos, et al., 2010; Sneath, 1973) with 100
bootstrap replicates. Finally, the constructed topology of the phylogenetic tree was visualized with
MEGA traditional option (Sup. Dataset 5).

Results

Dataset

The primary NCBI dataset contained 26646 entries, related to Notch family members. This dataset
consisted of not only Notch protein and its homologs, but also unrelated proteins, such as the
example of strawberry notch homolog (SNO). These irrelevant proteins, along with synthetic,
hypothetical, partial, low quality, predicted proteins and which were referred as to noisy data and
were removed from our dataset. Since the NCBI database provides partial duplicates sequences,
the sequences with greater than 95% similarity were also removed, retaining the one with the
longer length. Thus, the final dataset involved 603 Notch proteins and specifically, 90% of these
corresponded to Notch protein, 4% to Notch1, 3% to Notch2, 2% to Notch3, and 1% to Notch4
(Figure 1). Herein, representatives of all kingdoms from simpler organisms such as bacteria, to
more complex organisms including *Homo sapiens* were detected. The length of the Notch family proteins ranges between $\approx 71$ aa (*bacteria*'s Notch-like protein) and $\approx 4835$ aa (*invertebrates* Notch). As for the Notch1-4 paralogues, an exponential pattern of length was observed, with N4 being the shortest, followed by Notch3, Notch2 and finally Notch1 being the longest.

**Multiple sequence alignment and motifs**

Multiple sequence alignment (MSA) of protein sequences from the Notch family was performed to identify highly conservative regions within all organisms, from *monera* to *invertebrates*. As known, Notch receptors consist of specific conservative domains including EGF-like domain, NRR, TMD, RAM, NLS, ANK, NOD/NODP and PEST, that could be identified in the MSA (Aster, et al., 2017; Sander, et al., 2006) (Figure 2). Monitoring and analyzing the MSA using Jalview resulted in identifying EGF, LNR and NOD/NODP domains appearance in most kingdoms, though a variation of the number of repeats and length is noticeable (Gordon, et al., 2008). Even *bacteria* Notch-like proteins share some of the characteristic motifs of those regions in their receptors; therefore, we can speculate that lateral gene transfer might have occurred via bacterial transfers (Gazave, et al., 2009; Ponting, et al., 1999). Two other members of the Notch pathway, including Fringe (*Irving and Wieschaus, 1994*) and Strawberry Notch (Sno) (Gazave, et al., 2009) show the same scenario, in which the eventuality of lateral gene transfer cannot be excluded.

EGFs repeats are crucial components of the Notch signaling. Thus, it was expected to identify them as the most conserved site. EGF’s 6 Cysteine residues, being the key element of the domain, are responsible for the formation of disulfide bonds and influence the native 3D structure of the Notch members (Figure 3). The importance of these key residues is also apparent in the pathological phenotype. In the case of a Notch3 mutation, if any of the 6 Cys is mutated into another amino acid, it leads to a rare neurodegenerative syndrome, called CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy) (Papakonstantinou, et al., 2019; Vlachakis, et al., 2014). It is worth mentioning that the EGF-domain is also characterized by the conserved Glycine residues, as shown by the current multiple alignments (Figure 3). Furthermore, the highly conserved EGFs, led us to suggest that two new motifs A and B are crucial for the Notch family evolution. The CXNGGXC motif (Figure 3/ Motif A) consisting of two highly conserved Gly residues and a second motif (Figure 3 / Motif B), CXCXXG[FY]XG, characterized by a conserved Cys and Gly and a non-polar aromatic amino acid (F/Y) among two conserved Gly (Figure 3). These motifs could possibly be EGF-domain’s precursors, since they are repeated with a different number of repetitions depending on the respective species, where the number of repetitions is directly related to the complexity of the organism (Kovall, et al., 2017).

Additionally, a segment of the LNR-domain appears to be highly conserved in organisms from all kingdoms and the characteristic motif of the LNR-repeat was also found in *bacteria*. Motif C could be considered as the LNRe’s single unit of short peptide that has been repeated via internal tandem duplications through evolution in *eukaryotes* (Bjorklund, et al., 2006) (Figure 4 / Motif C). NOD/NODP domain seems to have maintained throughout the evolutionary history of all eukaryotes and it contains two conserved motifs (Motifs D and E). Remarkably, the conserved
motif D has been identified also in bacteria’s Notch-like proteins, which are significantly smaller than the homolog proteins of the other species.

**Structural characterization of Notch Receptor**

Despite the progress achieved recently, in the structure determination of limited fragments of the Notch receptor, its quaternary 3D structure remains undetermined (Figure 5). Notch receptor 3D models are based on the structural features of the long NECD region, since it contains many calcium binding EGF-like domains (Figure 5). The NECD differs between species. *Drosophila* and *mammalian* Notch receptors are much larger than their counterparts from other *invertebrate* species like *Caenorhabditis elegans*, although each invariably maintains the same molecular architecture (Kopan and Ilagan, 2009). On the other hand, in *monera like bacteria*, and in *protists*, the NECD size is much shorter and compact (Figure 2). Based on previous studies it has been observed that the NECD region of the Notch receptor is expected to have a rigid near-linear 3D structure, however potential sites of flexibility may occur at the 3D structure of the EGF domain which is less conserved between species / kingdoms. (Figure 2 and 3) (Kopan and Ilagan, 2009; Morgan, et al., 1999).

At the N-terminal end, for most species, the Notch receptor contains 36 EGF-like domains, a region of containing calcium-binding sites. Next to the EGF region there are three LNR and a hydrophobic region which has been shown to mediate heterodimerization (HD) (Figure 5). Together, the LNR repeats and the heterodimerization domain form the NRR, adjacent to the cell membrane (Figure 5). This region prevents ligand-independent activation of the Notch receptor by concealing and protecting from metalloproteases. When Notch activation is achieved via ligand binding to repeats within the EGF-like domain, then two sequential proteolytic events termed S2 and S3 cleavages are induced (Sanchez-Irizarry, et al., 2004). The S3 cleavage site lies within the transmembrane segment and is cleaved by the γ-secretase complex to liberate NICD. NICD comprises one RAM domain, seven ANK, one TAD and one PEST domain (Figure 5). Both the RAM domain and ANK repeats have been identified as regions involved in the interaction with CSL transcription factors (Chillakuri, et al., 2012). The TAD region is found in Notch-1 and -2 but not in -3 and -4 in *mammals* (Chillakuri, et al., 2012). The C-terminal PEST domain is involved in NICD degradation by proteolysis.

Recently, they have been determined the complete 3D structures of the extracellular domain (ECD) of the *Drosophila* Notch receptor and the *human* Notch1 receptor, by using single particle electron microscopy and antibody labelling (Chillakuri, et al., 2012; Kelly, et al., 2010). Since the inherent resolution of the method is poorly corresponding to the small size of the domains involved, these are significant challenging experiments.

**Ca**$^{2+}$ signaling activation and digital emulation

Notch family proteins, being capable of recognizing and binding to the neighbor cell’s ligand, achieve selective cell-cell adhesion initiated by protein-protein interaction. They comprise an extracellular region, composed of EGF-repeats, LNR repeats and NOD/NODP. EGF-repeats act as transmitters that promote the signal to the LNRs and along those extracellular regions, intracellularly to the ANK repeats. In this process, calcium plays a crucial role and is required for...
the biological activation of transmembrane proteins that contain EGF and LNR repeats. More precisely, Ca\(^{2+}\) binds to cbEGFs, a large subgroup of EGFs. CbEGFs have essentially similar structure as EGFs, but they are so conserved, that they cluster together in EGF dendrograms (Stenflo, et al., 2000). The central role of Ca\(^{2+}\) has been demonstrated through crystallization experiments, where Ca\(^{2+}\) seems to stabilize the N-terminal of tandem EGFs, establishing a stable surface for the protein-protein interaction (Downing, et al., 1996). The Ca\(^{2+}\) binding coordination is related to cbEGFs’ amino acid modifications, including hydroxylation and glycosylation of an asparagine or aspartic acid residue. Moreover, when a Ca\(^{2+}\) is binding, the extracellular structure is locally changing, making it versatile and difficult to model. Ca\(^{2+}\) role is also observed in the LNR repeats, as each cysteine-rich Lin-12/Notch repeat binds one calcium ion to achieve the correct folding and maintenance of its 3-D structure. Mutations in calcium binding sites lead to dysregulated pathways, introducing pathogenic phenotypes. A substantial case of defective Ca\(^{2+}\) binding is the Marfan syndrome and Hemophilia B.

The extracellular region of a Notch family precursor is a mix of calcium-loaded and calcium-free sites, with each EGF repeat forming a beta-sheet and the LNR repeats forming loops. Simulating each beta-sheet as a hoop, the whole EGF region could be presented as a strongly-connected chain. Furthermore, emulating the chain into a digital format, each hoop, depending on its Ca\(^{2+}\) binding or Ca\(^{2+}\) binding disability, is emulated by a sequence of 1 and 0, respectively. In that way, simulating each Lin12/Notch repeat as a gear (are more specialized forms than EGF repeats), the complex of LNRs could be emulated by a compact gear mechanism, which produces a more specialized signal based on the calcium interactions (Figure 6). Combining all this mechanism together we can easily understand its specialization in producing different unique codes (signals) to the cell in order to start different biological tasks. This proposed emulation could represent a way to present if the signal is going to be promoted from the EGF repeats to the LNRs etc. and finally to the nucleus, completing a the signaling pathway. In case of mutation in these residues due to defective Ca\(^{2+}\) binding, the corresponding digit changes and an error in the code occurs, indicating signal transduction failure associated with the possibility of a functional abnormality (Figure 6).

### Notch family protein clusters

Notch family protein clusters were estimated using a phylogenetic analysis (Figure 7). A phylogenetic tree could provide clustering information for the Notch family members, so that representative consensus sequences could be properly defined. Having such a large dataset with proteins of variant sequence length, we initiated the construction of an unrooted tree. Phylogenetic analysis revealed a distinct separation of Notch family members into eight monophyletic branches, the Notch bacteria cluster, the Notch plants cluster, the Notch protist cluster, the Notch invertebrates cluster, as well as the Notch1, Notch2, Notch3 and Notch4 clusters. Each cluster’s content was carefully examined and reviewed (proteins, classes and kingdoms it contained), eventually leading to clustering the data into the eight groups (Table 2). Moreover, it was clear that the resulting clusters of the phylogenetic tree were formed by related protein members and not by classes/organisms, even if we included the same species in different clusters for the Notch1 to 4 homologs. Furthermore, considering the conservation and similarity of the characteristic domains presented in each Notch cluster, it was verified that the Notch family members
classification was accurate because they belong in the same phylum / kingdom (Figure 7). In this study, using protein sequences from representatives of all kingdoms, it was necessary to observe the agreement of the emerging phylogenetic tree with the tree of life. Thus, a bibliographic examination was carried out, which confirmed the correct and proper positioning and clustering of organisms according to evolution, by examining the position of the invertebrates, such as cnidarians, flatworms, nematodes, arthropods, chordates, as well as the phylogenetic position of the protists and plants, in the tree of life (Keeling, 2019) (Table 2).

**Notch family evolution**

In this study, we identified novel Notch protein clusters from various kingdoms extending from bacteria to chordates and conducted a comprehensive phylogenetic analysis in order to confirm and expand the evolutionary history of the Notch family (Figure 8). Notch family evolution was examined in a more specialized phylogenetic analysis with 100 bootstrap replicates using representative consensus protein sequences of each identified cluster from the previous phylogenetic analysis. Each consensus sequence, with its amino acids corresponding to the most frequently encountered, fully represented the organisms and the members of the NOTCH family, accordingly. Hence, the specialized phylogenetic tree contained meaningful information, as needed for further evolutionary research (Papageorgiou, et al., 2018). Based on results, both constructed phylogenetic trees of the Notch family members were found to share similar topology.

Due to the small number of sequences which were available the past decade, all previously performed evolutionary studies have a significantly smaller dataset. In this study, a more comprehensive phylogenetic analysis was achieved and it was important to compare the evolutionary scenarios presented here with the previous ones. The evolutionary distances between Notch2 and Notch3 groups have been found shorter than Notch, Notch1 and Notch4. This finding is correspondingly consistent with the evolutionary analysis performed by Gazave, E., et al. (Gazave, et al., 2009). On the other hand, Notch4 group found to be evolutionary older than Notch1, Notch2 and Notch 3, and with this result being consistent with the evolutionary study performed by Theodosiou, A., et al. (Theodosiou, et al., 2009). However, in this study we have introduced and present for the first time, the plant, protist and bacteria groups which are in the early Notch evolution (Figure 8). Another remarkable observation is that Notch protein groups are positioned before Notch 1-4 groups, indicating the origin of the Notch family. In consideration of the tree of life, bacteria, protists and plants, appeared before the rest of the eukaryotes and therefore their position in the constructed tree is validated. Bacteria belongs to prokaryotes, while protists and plants are eukaryotic organisms and thus, much more complex than prokaryotes. The fact that the latter two are eukaryotes, combined with the absence of evidence for the existence of Notch proteins in Archaea, partially modifies the obtained results, compared to the current evolutionary theories. In addition, bearing in mind the limited number of representatives given in the kingdom of plants, this slight divergence in the present evolutionary study of the Notch family is justified. However, future enrichment of the dataset, with additional Notch family protein sequences from other species, could provide an improved and even more distinct phylogenetic analysis. Lastly, in the phylogenetic tree, Notch paralogs of Caenorhabditis elegans appear distinctly before Notch1-4, as expected, given the fact of an independent genomic duplication, leading to two Notch genes (Theodosiou, et al., 2009).
Central nervous system (CNS) has evolved in early metazoans (such as marine organisms) as a simple neural network. This precursor system, necessary for cells’ electrical signaling and cell-to-cell interaction, became more complicated through evolution (Akanuma, et al., 2002). Notch family members form a key component of an evolutionarily conserved signaling mechanism, involved in the regulation of the CNS. Indeed, Notch signaling takes part in Neural Stem Cell (NSC) proliferation, survival, self-renewal, differentiation, apoptosis and cell fate choices (Gazave, et al., 2009). In the CNS, Notch members are present during the entire lifetime, from embryonic stages to adult nervous system, controlling neurogenesis, axons’ and dendrites’ growth and CNS plasticity (Presente, et al., 2001).

Several pathogenic mechanisms have been observed from different mutation cases in Notch receptor protein domains. Recent studies suggest that Notch receptors play correspondingly a crucial role in neurodegenerative disorders, including Alzheimer’s disease, Down syndrome and CADASIL (Polychronidou, et al., 2015). Notch family members are estimated that they existed a billion years before. Previous studies have shown that Notch signaling in Metazoa is a subsequent evolutionary result, as it uses ancient protein domains and mechanisms. Thus, it was assumed that this pathway may exist in Úrmetazoa or even in more inferior organisms, such as bacteria which also have Notch-like proteins (Gazave, et al., 2009). Therefore, new insights have been extracted from the present in silico study, where it could form the basis for detecting unrelated functions in this receptor family. So far, the evolutionary history of the Notch family seems to be directly linked to the tree of life. In the same direction, the complexity of these proteins could be proportional to the complexity of the species and the nervous system, as well. Last but not least, bacteria’s undifferentiated Notch-like proteins seem to have been evolved and differentiated in all other types of paralogues Notch’s as the complexity of the species increases. Similar evo-devo scenarios are claimed generally in proteins families, which are important for survival and evolution, such as the example of the nuclear receptor family (Mitsis, et al., 2019).

Despite the enormous scientific interest in Notch family proteins, current knowledge of their involvement in biological pathways and their function is still quite limited. Certainly, this knowledge could be significantly enriched with the potential determination of the Notch receptor 3D structure. Considering the cell-fate-determination and the cell communication, enforced by these proteins as well as the proteolytic procedure that they undergo in the signaling pathway, Notch receptors could be used as promising therapeutic targets for several diseases, including Cancer. In the present study, a series of the most highly conserved motifs that have arisen through evolution is presented. These motifs could be used as innovative pharmacological targets, through the development of new technologies concerning the numerous critical pathways, affected by this protein family. All in all, useful beneficial insights are provided, concerning the Notch family’s evolution. A comprehensive evolutionary analysis that confirms the existence of several kingdoms among Notch family members is provided. Notch genes were duplicated several times during evolution, leading to four genes in chordates, including Homo sapiens. If we accept that more than
the half of our body is not human, then the most logical scenario for the ancient origin of Notch are the bacteria.

**Figures legend**

**Table 1**: Diseases which are caused by NOTCH1-4 mutations. The research was done using the DisGeNET database ([www.disgenet.org](http://www.disgenet.org)).

**Table 2**: Categorization of the NOTCH family members using identified clusters, per kingdom and phylum.

**Figure 1**: Statistical analysis of the Notch family members based on the sequence annotation.

**Figure 2**: Architecture of Notch family receptors based on the Interpro Database ([Mitchell, et al., 2015](#)) results for all identified kingdoms. Notch family receptors are represented with major domains annotated, including Notch extracellular domain (NECD), Notch intracellular domain (NICD), transmembrane domain (TM), Epidermal growth factor (EGF), Cysteine-rich LNR repeats (LNR), Notch domain present in many Notch proteins (NOD/NODP), RBPJκ-associated molecule domain (RAM), Nuclear localization sequences (NLS), Ankyrin repeat domain (ANK), and the Domain rich in proline, glutamine, serine and threonine residues (PEST). Protein domains marked with (*) represent domains that have only been observed in the MSA, and those marked with (?) represent domains for which no other information is available.

**Figure 3**: Highly conserved Notch family motifs based on the 603 Notch sequences MSA results. Motif-A and Motif-B have been highlighted (colored red) in the EGF-domain, based on the Notch consensus sequence from the MSA. The characteristic EGF-Cys-repeats are illustrated in with large red boxes.

**Figure 4**: Highly conserved Notch family regions based on the 603 Notch sequences MSA results. (A) Conserved region of the LNR domain (B). Conserved regions of the NOD/NODP domain.

**Figure 5**: Structural architecture of Notch family receptors based on the available fragments from the Protein Data Bank (PDB: 5FM9, 3ETO, 3V79).

**Figure 6**: Structural visualization of the ENCD of a Notch family receptor and its digital format.

**Figure 7**: The unrooted phylogenetic tree of Notch family 603 members. The tree was constructed utilizing the UPGMA method in MATLAB and visualized using the TreeExplorer tool of MEGA. Eight distinct monophyletic branches are visible. The phylogenetic trees confidently separate the Notch1 cluster (branch colored Pink), the Notch2 cluster (branch colored orange), the Notch 3 cluster (branch colored red), the Notch 4 cluster (branch colored yellow), the Notch Bacteria cluster (branch colored light blue), the Notch Plants cluster (branch colored green), the Notch Protist cluster (branch colored dark blue) and the Notch *invertebrates* cluster (branch colored gray).
Figure 8: The evolutionary history of Notch family members in a specialized phylogenetic tree. The tree was constructed with UPGMA method in MATLAB for 100 bootstrap replicates and was visualized in MEGA.

References

Akanuma, T., Hori, S., Darras, S. and Nishida, H. (2002) Notch signaling is involved in nervous system formation in ascidian embryos, Development genes and evolution, 212, 459-472.

Aster, J.C., Pear, W.S. and Blacklow, S.C. (2017) The Varied Roles of Notch in Cancer, Annual review of pathology, 12, 245-275.

Bjorklund, A.K., Ekman, D. and Elofsson, A. (2006) Expansion of protein domain repeats, PLoS computational biology, 2, e114.

Chillakuri, C.R., Sheppard, D., Lea, S.M. and Handford, P.A. (2012) Notch receptor-ligand binding and activation: insights from molecular studies, Seminars in cell & developmental biology, 23, 421-428.

Dexter, J.S. (1914) The analysis of a case of continuous variation in Drosophila by a study of its linkage relations, The American Naturalist, 48, 712-758.

Downing, A., Knott, V., Werner, J., Cardy, C., Campbell, I. and Handford, P. (1996) Solution structure of a pair of calcium-binding epidermal growth factor-like domains: implications for the Marfan syndrome and other genetic disorders, Cell, 85, 597-605.

Durieux, I., Ginevra, C., Attaiech, L., Picq, K., Juan, P.A., Jarraud, S. and Charpentier, X. (2019) Diverse conjugative elements silence natural transformation in Legionella species, Proceedings of the National Academy of Sciences of the United States of America, 116, 18613-18618.

Fairclough, S.R., Chen, Z., Kramer, E., Zeng, Q., Young, S., Robertson, H.M., Begovic, E., Richter, D.J., Russ, C., Westbrook, M.J., Manning, G., Lang, B.F., Haas, B., Nusbaum, C. and King, N. (2013) Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta, Genome biology, 14, R15.

Gazave, E., Lapebie, P., Richards, G.S., Brunet, F., Ereskovsky, A.V., Degnan, B.M., Borchiellini, C., Vervoort, M. and Renard, E. (2009) Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes, BMC evolutionary biology, 9, 249.

Gerhardt, D.M., Pajcini, K.V., D’altri, T., Tu, L., Jain, R., Xu, L., Chen, M.J., Rentschler, S., Shestova, O. and Wertheim, G.B. (2014) The Notch1 transcriptional activation domain is required for development and reveals a novel role for Notch1 signaling in fetal hematopoietic stem cells, Genes & development, 28, 576-593.

Girard, L.R., Fiedler, T.J., Harris, T.W., Carvalho, F., Antoshechkin, I., Han, M., Sternberg, P.W., Stein, L.D. and Chalfie, M. (2007) WormBook: the online review of Caenorhabditis elegans biology, Nucleic acids research, 35, D472-475.

Gordon, W.R., Arnett, K.L. and Blacklow, S.C. (2008) The molecular logic of Notch signaling—a structural and biochemical perspective, Journal of cell science, 121, 3109-3119.

Hogan, K.A. and Bautch, V.L. (2004) Blood vessel patterning at the embryonic midline. In, Current topics in developmental biology. Elsevier, pp. 55-85.

Hori, K., Sen, A. and Artavanis-Tsakonas, S. (2013) Notch signaling at a glance. The Company of Biologists Ltd.

Irvine, K.D. and Wieschaus, E. (1994) fringe, a Boundary-specific signaling molecule, mediates interactions between dorsal and ventral cells during Drosophila wing development, Cell, 79, 595-606.
Keeling, P.J. (2019) Combining morphology, behaviour and genomics to understand the evolution and ecology of microbial eukaryotes, *Philosophical Transactions of the Royal Society B*, 374, 20190085.

Kelly, D.F., Lake, R.J., Middelkoop, T.C., Fan, H.Y., Artavanis-Tsakonas, S. and Walz, T. (2010) Molecular structure and dimeric organization of the Notch extracellular domain as revealed by electron microscopy, *PloS one*, 5, e10532.

Kopan, R. and Ilagan, M.X. (2009) The canonical Notch signaling pathway: unfolding the activation mechanism, *Cell*, 137, 216-233.

Kovall, R.A., Gebelein, B., Sprinzak, D. and Kopan, R. (2017) The Canonical Notch Signaling Pathway: Structural and Biochemical Insights into Shape, Sugar, and Force, *Developmental cell*, 41, 228-241.

Krebs, L.T., Xue, Y., Norton, C.R., Shutter, J.R., Maguire, M., Sundberg, J.P., Gallahan, D., Closson, V., Kitajewski, J. and Callahan, R. (2000) Notch signaling is essential for vascular morphogenesis in mice, *Genes & development*, 14, 1343-1352.

Kufareva, I. and Abagyan, R. (2012) Methods of protein structure comparison, *Methods in molecular biology*, 857, 231-257.

Michener, C.D. and Sokal, R.R. (1957) A QUANTITATIVE APPROACH TO A PROBLEM IN CLASSIFICATION, *Evolution*, 11, 130-162.

Mitchell, A., Chang, H.Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.Y., Bateman, A., Punta, M., Attwood, T.K., Sigrist, C.J., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D.A., Wu, C.H., Oregno, C., Silitoie, I., Mi, H., Thomas, P.D. and Finn, R.D. (2015) The InterPro protein families database: the classification resource after 15 years, *Nucleic acids research*, 43, D213-221.

Mitsis, T., Papaageorgiou, L., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, G.P. and Eliopoulos, E. (2019) A comprehensive structural and functional analysis of the ligand binding domain of the nuclear receptor superfamily reveals highly conserved signaling motifs and two distinct canonical forms through evolution, *World Academy of Sciences Journal of Life Sciences*, 1, 264-274.

Mizutani, K.-i., Yoon, K., Dang, L., Tokunaga, A. and Gaiano, N. (2007) Differential Notch signalling distinguishes neural stem cells from intermediate progenitors, *Nature*, 449, 351.

Morgan, W.D., Birdsall, B., Frenkkel, T.A., Gradwell, M.G., Burghaus, P.A., Syed, S.E., Uthaipibull, C., Holder, A.A. and Feeney, J. (1999) Solution structure of an EGF module pair from the Plasmodium falciparum merozoite surface protein 1, *Journal of molecular biology*, 289, 113-122.

Muiño, E., Gallego-Fabrega, C., Cullell, N., Carrera, C., Torres, N., Krupinski, J., Roquer, J., Montaner, J. and Fernández-Cadenas, I. (2017) Systematic review of cysteine-sparing NOTCH3 missense mutations in patients with clinical suspicion of CADASIL, *International journal of molecular sciences*, 18, 1964.

Nanni, L., Lumini, A. and Brahanam, S. (2014) An empirical study of different approaches for protein classification, *TheScientificWorldJournal*, 2014, 236717.

Papaageorgiou, L., Eleni, P., Raftopoulou, S., Mantaioiu, M., Megaloooknomou, V. and Vlachakis, D. (2018) Genomic big data hitting the storage bottleneck, *EMBnet journal*, 24.

Papaageorgiou, L., Loukatou, S., Sofia, K., Maroulis, D. and Vlachakis, D. (2016) An updated evolutionary study of Flaviviridae NS3 helicase and NS5 RNA-dependent RNA polymerase reveals novel invariable motifs as potential pharmacological targets, *Molecular bioSystems*, 12, 2080-2093.

Papakonstantinou, E., Bacopoulou, F., Brouzas, D., Megaloooknomou, V., D’Elia, D., Bongcam-Rudloff, E. and Vlachakis, D. (2019) NOTCH3 and CADASIL syndrome: a genetic and structural overview, *EMBnetjournal*, 24.

Pavloupolou, G.A., Soldatos, T.G., Barbosa-Silva, A. and Schneider, R. (2010) A reference guide for tree analysis and visualization, *BioData mining*, 3, 1.

Polychronidou, E., Vlachakis, D., Vlamos, P., Baumann, M. and Kossida, S. (2015) Notch signaling and ageing. In, *GeNeDis 2014*. Springer, pp. 25-36.
Ponting, C.P., Aravind, L., Schultz, J., Bork, P. and Koonin, E.V. (1999) Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer, Journal of molecular biology, 289, 729-745.

Presente, A., Andres, A. and Nye, J.S. (2001) Requirement of Notch in adulthood for neurological function and longevity, Neuroreport, 12, 3321-3325.

Sanchez-Irizarry, C., Carpenter, A.C., Weng, A.P., Pear, W.S., Aster, J.C. and Blacklow, S.C. (2004) Notch subunit heterodimerization and prevention of ligand-independent proteolytic activation depend, respectively, on a novel domain and the LNR repeats, Molecular and cellular biology, 24, 9265-9273.

Sander, G.R., Krysinska, H. and Powell, B.C. (2006) Developmental Signaling Networks: The Notch Pathway, Physiology of the Gastrointestinal Tract, 287.

Sneath, P.H.A., Sokal R.R (1973) Unweighted pair group method with arithmetic mean, Numerical Taxonomy, 230–234.

Sobie, E.A. (2011) An introduction to MATLAB, Science signaling, 4, tr7.

Sorkac, A., Dilorio, M.A., O'Hern, P.J., Baskoylu, S.N., Graham, H.K. and Hart, A.C. (2018) LIN-12/Notch Regulates GABA Signaling at the Caenorhabditis elegans Neuromuscular Junction, G3, 8, 2825-2832.

Stenflo, J., Stenberg, Y. and Muranyi, A. (2000) Calcium-binding EGF-like modules in coagulation proteinases: function of the calcium ion in module interactions, Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 1477, 51-63.

Tanigaki, K. and Honjo, T. (2007) Regulation of lymphocyte development by Notch signaling, Nature immunology, 8, 451.

Theodosiou, A., Arhondakis, S., Baumann, M. and Kossida, S. (2009) Evolutionary scenarios of Notch proteins, Molecular biology and evolution, 26, 1631-1640.

Vlachakis, D., Tsaniras, S.C., Ioannidou, K., Papageorgiou, L., Baumann, M. and Kossida, S. (2014) A series of Notch3 mutations in CADASIL; insights from 3D molecular modelling and evolutionary analyses, Journal of Molecular Biochemistry, 3.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. and Barton, G.J. (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench, Bioinformatics (Oxford, England), 25, 1189-1191.

Yang, K. and Zhang, L. (2008) Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction, Nucleic acids research, 36, e33.

Yavropoulou, M.P., Maladaki, A. and Yovos, J.G. (2015) The role of Notch and Hedgehog signaling pathways in pituitary development and pathogenesis of pituitary adenomas, Hormones, 14, 5-18.
Table 1 (on next page)

Diseases which are caused by NOTCH1-4 mutations. The research was done using the DisGeNET database ( www.disgenet.org ).

Diseases which are caused by NOTCH1-4 mutations. The research was done using the DisGeNET database ( www.disgenet.org ).
Table 1: Diseases which are caused by NOTCH1-4 mutations. The research was done using the DisGeNET database (www.disgenet.org).

| Diseases                  | NOTCH1                                      | NOTCH2                                      | NOTCH3                                      | NOTCH4                                      |
|---------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|
|                           | T-cell acute lymphoblastic leukemia          | Hajdu-Cheney Syndrome                       | CADASIL                                     | Unclarified                                 |
|                           | Adams-Oliver Syndrome                        | Alagille Syndrome                           | Infantile Myofibromatosis                   |                                             |
|                           | Aortic Valve Disease                         | Cancer                                      | Early-onset arteriopathy with cavitating leukodystrophy |                                             |
|                           | Cancer                                       | Cancer                                      | Lateral meningocele Syndrome                 |                                             |
|                           |                                             |                                             | Cancer                                      |                                             |
Table 2 (on next page)

Categorization of the NOTCH family members using identified clusters, per kingdom and phylum.

| Kingdom/Phylum | Cluster 1 | Cluster 2 | Cluster 3 |
|---------------|-----------|-----------|-----------|
| Eubacteria    |           |           |           |
| Archaea       |           |           |           |
| Eukarya       |           |           |           |

Categorization of the NOTCH family members using identified clusters, per kingdom and phylum.
Table 2: Categorization of the NOTCH family members using identified clusters, per kingdom and phylum.

| Kingdoms     | Eukaryotes | Single Cell | BACTERIA | PROTISTS | ANIMALS | PLANTS | FUNGI |
|--------------|------------|-------------|----------|----------|---------|--------|-------|
| Phylum       |            |             |          |          |         |        |       |
|              |            |             |          |                               |         |        |       |
|              |            |             |          | Placezoans | Cnidarians | Flatworms | Nematodes | Arthropods | Mammals | Fishes | Birds | Turtles | Amphibians | Lizards |        |       |
|              |            |             |          | Placozoa | Cnidarians | Flatworms | Nematodes | Arthropods | Mammals | Fishes | Birds | Turtles | Amphibians | Lizards |        |       |
|              |            |             |          | -        | -        | -        | -         | -         | -        | -      | -     | -      | -        | -        | 5      | 1     |
| NOTCH        | ✓          | ✓          | ✓        | ✓        | ✓        | ✓        | ✓        | ✓         | ✓        | ✓      | ✓     | ✓      | ✓        | ✓        | ✓      | ✓     |
| 40           | 8          | 112        |          |          |          |          |          | 5        |         |        |       |        |          |          |        |       |
| NOTCH1       | -          | -          | -        | -        | ✓        | ✓        | ✓        | ✓         | ✓        | ✓      | ✓     | ✓      | ✓        | ✓        | -      | -     |
|              |            |             |          |          |          |          |          | 152       |          |        |       |        |          |          |        |       |
| NOTCH2       | -          | -          | -        | -        | ✓        | ✓        | ✓        | ✓         | ✓        | ✓      | ✓     | ✓      | ✓        | ✓        | -      | ✓     |
|              |            |             |          |          |          |          |          | 114       |          |        |       |        |          |          | 1      |       |
| NOTCH3       | -          | -          | -        | -        | ✓        | ✓        | ✓        | ✓         | ✓        | ✓      | ✓     | ✓      | ✓        | ✓        | -      | -     |
|              |            |             |          |          |          |          |          | 83        |          |        |       |        |          |          |        |       |
| NOTCH4       | -          | -          | -        | -        | ✓        | ✓        | ✓        | ✓         | ✓        | ✓      | ✓     | ✓      | ✓        | ✓        | -      | -     |
|              |            |             |          |          |          |          |          | 87        |          |        |       |        |          |          |        |       |
Figure 1

Statistical analysis of the Notch family members based on the sequence annotation

Statistical analysis of the Notch family members based on the sequence annotation
Figure 2

Architecture of Notch family receptors based on the Interpro Database (Mitchell, et al., 2015) results for all identified kingdoms.

Architecture of Notch family receptors based on the Interpro Database (Mitchell, et al., 2015) results for all identified kingdoms. Notch family receptors are represented with major domains annotated, including Notch extracellular domain (NECD), Notch intracellular domain (NICD), transmembrane domain (TM), Epidermal growth factor (EGF), Cysteine-rich LNR repeats (LNR), Notch domain present in many Notch proteins (NOD/NODP), RBPJκ-associated molecule domain (RAM), Nuclear localization sequences (NLS), Ankyrin repeat domain (ANK), and the Domain rich in proline, glutamine, serine and threonine residues (PEST). Protein domains marked with (*) represent domains that have only been observed in the MSA, and those marked with (?) represent domains for which no other information is available.
Figure 3

Highly conserved Notch family motifs. Motif-A and Motif-B have been highlighted (colored red) in the EGF-domain, based on the Notch consensus sequence from the MSA.

Highly conserved Notch family motifs. Motif-A and Motif-B have been highlighted (colored red) in the EGF-domain, based on the Notch consensus sequence from the MSA. The characteristic EGF-Cys-repeats are illustrated in with large red boxes.
Figure 4

Highly conserved Notch family regions based on the 603 Notch sequences MSA results. (A) Conserved region of the LNR domain (B). Conserved regions of the NOD/NODP domain.

Highly conserved Notch family regions based on the 603 Notch sequences MSA results. (A) Conserved region of the LNR domain (B). Conserved regions of the NOD/NODP domain.
Figure 5

Structural architecture of Notch family receptors based on the available fragments from the Protein Data Bank (PDB: 5FM9, 3ETO, 3V79).

Structural architecture of Notch family receptors based on the available fragments from the Protein Data Bank (PDB: 5FM9, 3ETO, 3V79).
Figure 6

Structural visualization of the ENCD of a Notch family receptor and its digital format.
The unrooted phylogenetic tree of Notch family members. The tree was constructed utilizing the UPGMA method in MATLAB and visualized using the TreeExplorer tool of MEGA. Eight distinct monophyletic branches are visible. The phylogenetic trees confidently separate the Notch1 cluster (branch colored Pink), the Notch2 cluster (branch colored orange), the Notch 3 cluster (branch colored red), the Notch 4 cluster (branch colored yellow), the Notch Bacteria cluster (branch colored light blue), the Notch Plants cluster (branch colored green), the Notch Protist cluster (branch colored dark blue) and the Notch multicellular cluster (branch colored gray).
Figure 8

The evolutionary history of Notch family members in a specialized phylogenetic tree. The tree was constructed with UPGMA method in MATLAB for 100 bootstrap replicates and was visualized in MEGA.

The evolutionary history of Notch family members in a specialized phylogenetic tree. The tree was constructed with UPGMA method in MATLAB for 100 bootstrap replicates and was visualized in MEGA.
