Predicting function and structure using bioinformatics protocols: study of the intracellular regions of the Jagged and Delta protein families

Thesis

Neli Ivanova

Predicting function and structure using bioinformatics protocols: study of the intracellular regions of the Jagged and Delta protein families.

Thesis submitted for

Master of Philosophy in Life Sciences

Open University, U.K.

Date of submission: 15 May 2006

International Centre for Genetic Engineering and Biotechnology

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

CONTENTS

ABSTRACT

Title: **Predicting function and structure using bioinformatics protocols: study of the intracellular regions of the Jagged and Delta protein families.**

Author: Neli Ivanova, B.Sc.

Director of Studies: Sándor Pongor, Ph.D., D. Sc.

External supervisor: Martin J. Bishop

Study was carried out at: International Centre for Genetic Engineering and Biotechnology (ICGEB), Trieste

The type I membrane-spanning proteins Jagged (Jagged-1 and -2) and Delta (Delta-1, -3 and -4) are the human ligands of Notch receptors, which mediate key signaling events in cell differentiation and morphogenesis. The Jagged and Delta proteins are composed of a relatively large extracellular region and of a 100-150 residue, yet uncharacterized cytoplasmic tail, which has been recently found to be important in Notch bi-directional signaling. We applied bioinformatics methods to analyze the intracellular region of human Notch ligands, and to predict their structural and functional properties. We searched databases for orthologues, and found that while the intracellular region is evolutionary well conserved within the same ligand type, a wide variability is observed in different ligands. No significant similarity was found between the intracellular region of Jagged and Delta and proteins of known 3D structure. Globularity and disorder predictions indeed suggest that these regions are largely unstructured. However, secondary structure predictions show that these regions have some propensity to form local secondary structure elements. Functional predictions based on pattern recognition imply that the specificity in the Notch machinery response might be related to specific post-translational modifications and binding motifs in the ligand cytoplasmic tail,

rather than to specific interactions between the receptors and the extracellular region of the ligands. We also speculate that, given the unusual amino acid composition, the cytoplasmic tail of Jagged and Delta might be involved in zinc binding.

LIST OF ABBREVIATIONS

ADAM, a disintegrin and metalloproteinase

AF6, human afadin

AGS, Alagille syndrome

BLAST, basic local alignment search tool

BLOSUM, blocks substitution matrix

CADASIL, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy

DisEMBL, intrinsic protein disorder prediction

DisPhos, disorder enhanced phosphorylation sites predictor

Dlg1, human homologue of *Drosophila* Discs Large protein

DLL1-4, human homologues of *Drosophila* Delta

DSL, Delta/Serrate/Lag-2 domain

EGF, epidermal growth factor

ELM, eukaryotic linear motif

GlobPlot, predictor of intrinsic protein disorder & globularity

HMM, hidden Markov models

Jag1-2, Jagged proteins

IUP, intrinsically unstructured protein

IUPRED, prediction of intrinsically unstructured regions

MAGI, membrane associated guanylate kinase with inverted architecture

MAGUK, membrane associated guanylate kinase

MIM, Mendelian Inheritance in Man

NetOGlyc, predictions of β-N-acetylglucosamine O-glycosylation

NetPhos, neural network-based predictor of phosphorylation sites

NTC1-4, Notch receptors 1-4

PDZ, PSD-95/Dlg/ZO-1,2 domain

PONDR, predictor of naturally disordered regions

PSI-BLAST, position specific iterative BLAST

PSI-PRED, protein structure prediction

SD, spondylocostal dysostosis

SEG, filtering of low complexity segments

T-ALL, T cell acute lymphoblastic leukemia

TOF, familial form of tetralogy of Fallot

3D-PSSM, fold recognition using position specific scoring matrix

INTRODUCTION

Notch signaling

Mechanism of the core signaling pathway. Notch mediated signal transduction controls cell fate (specification, differentiation, proliferation and survival) and is a key process in tissue patterning and morphogenesis in developing vertebrates and invertebrates (Artavanis-Tsakonas et al., 1999; Kadesch, 2004). The main players in this signaling network are Notch receptors, four members of which have been identified in humans (NTC1, NTC2, NTC3, NTC4), and their corresponding ligands, belonging to two distinct families: homologues of *Drosophila* delta protein (DLL1, DLL3, DLL4) and homologues of *Drosophila* Serrate, Jagged-1 and -2 (JAG1, JAG2).

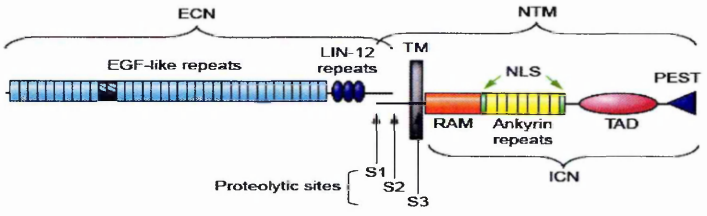
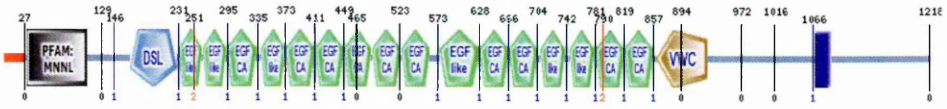


Figure 1. Domain organization of Notch receptors. Human Notch1 (NTC1) is shown as an example. Proteolytic cleavage by furin at site S1 produces two subunits, ECN and NTM, which remain non-covalently associated at the cell surface. EGF-like modules 11 and 12, implicated in ligand binding in *Drosophila* Notch, are shaded. S2 and S3 identify the sites of proteolytic cleavage induced upon activation by the ligand. ICN, intracellular domain of Notch; NLS, nuclear localization signal; PEST, proline, glutamate, serine, threonine rich sequence; TAD, transactivation domain; TM, transmembrane.

Notch receptors are membrane-spanning glycoproteins assembled in a non-covalent heterodimeric complex. (**Figure 1**) The polypeptide encoded by Notch genes is proteolytically cleaved in the Golgi during the transport to the cell surface, to give an extracellular (ECN) and a transmembrane subunit (NTM). The ECN contains an array of 29-36 EGF tandem repeats, followed by three LIN-12 repeats

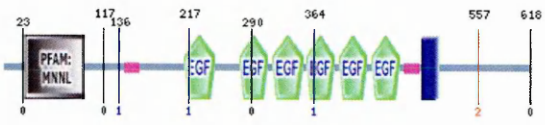
that maintain Notch in a resting state. The intracellular region of the NTM includes a RAM domain, followed by seven ankyrin repeats, a TAD domain, and a PEST region. All the ligands of the DSL (Delta/Serrate/Lag-2) family share the same architecture (Letunic et al., 2004) (**Figure 2**). They are type I membrane spanning proteins composed of a N-terminal, cysteine rich region that includes a DSL domain, a variable number of EGF-like repeats, a transmembrane segment, and a relatively short (~100-150 amino acids) cytoplasmic tail. Ligands of the Jagged group (JAG1 and JAG2) have also a juxtmembrane additional region that is not present in the Delta group ligands.



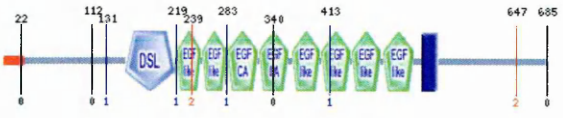
JAG1_HUMAN

JAG2_HUMAN

DLL1_HUMAN



DLL3_HUMAN



DLL4_HUMAN

Figure 2. Domain architecture of human Notch ligands as depicted by SMART. **MNNL**, N-terminal region of Notch ligands (Pfam); **DSL,** Delta/Serrate/lag-2 domain; **EGF-like** - epidermal growth factor (EGF) domain, unclassified subfamily; **EGF_Ca** - Calcium-binding EGF-like domain; **VWC** - von Willebrand factor (VWF) type C domain; the transmembrane region is shown as a blue rectangle; low-complexity regions in magenta.

Notch signaling is initiated by receptor-ligand interactions between two distinct cells. The receptor/ligand interaction has not been characterized in detail yet. From deletion studies, it has been found that a couple of tandem EGF repeats in the receptor (EGF-11 and -12) (Rebay et al., 1991) and the DSL domain in the ligand (Shimizu et al., 1999) are the minimal requirement for the binding to occur. In response to ligand binding, the transmembrane subunit of the receptor (NTM) is cleaved by an extracellular ADAM type metalloproteinase, 12 residues upstream of the membrane-spanning region. This cleavage facilitates a further cleavage of NTM, on the cytoplasmic side. This cleavage is carried out by the presenilin/γ-secretase protease and releases the intracellular domain (ICN) from the membrane (Weinmaster, 2000). This series of controlled proteolytic events is referred to as "regulated intramembrane proteolysis" or RIP, and is a signal transduction mechanism shared with the adhesion molecules CD44 and nectin-1, the amyloid β-A4 protein, the ErbB-4 receptor tyrosine protein kinase, and others. Once translocated into the nucleus, the ICN interacts with nuclear factors that activate transcription, the main target being a transcription factor (CSL) called CBF1/RBP in mammals, Suppressor of Hairless in *Drosophila*, and LAG-1 in *C. elegans* (**Figure 3**).

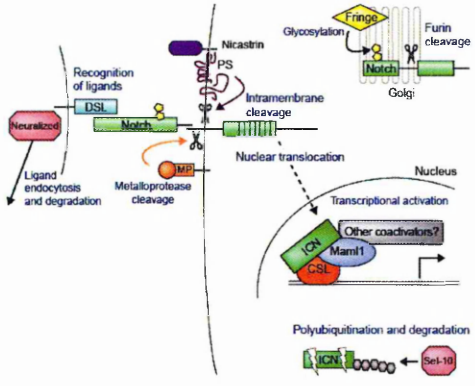


Figure 3. Key biochemical events in the Notch signal transduction pathway.

Notch signaling is regulated at different levels (**Figure 4**): glycosylation of receptors and ligands is tuning receptor/ligand recognition (Haines and Irvine, 2003), cytoplasmic proteins like Numb and Deltex play a role in suppressing Notch signal, E3 ubiquitin ligases regulate the level of Notch signal by targeting its components for degradation (Lai, 2002), and several nuclear proteins take part to the activation of transcription.

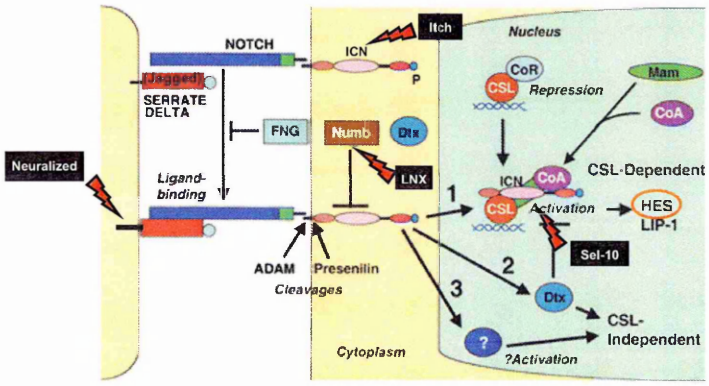


Figure 4. Regulation of Notch signaling.

Bi-directional signaling. Recent reports show that Notch ligands undergo a proteolytic processing that is strikingly similar to that reported for Notch receptors (**Figure 5**).

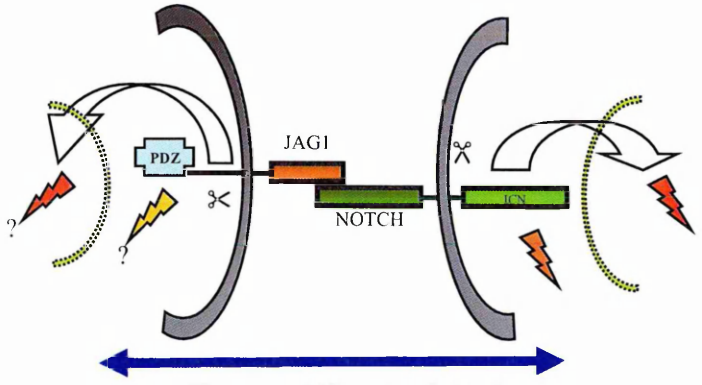


Figure 5. Bidirectional signaling.

Delta and Jagged undergo ADAM-mediated ectodomain processing followed by presenilin/γ-secretase-mediated intramembrane proteolysis to release signaling fragments (Ascano et al., 2003; Ikeuchi and Sisodia, 2003; LaVoie and Selkoe, 2003; Six et al., 2003). In these events, Jagged and Delta compete with Notch and might thus antagonize Notch signaling *in vivo*. The intracellular region of these ligands released from the cell membrane can be found in the cytoplasm as well as in the nucleus, where it can activate gene expression via the transcription factor AP1 (p39 jun). Notch-related signal transduction pathways are thus active not only in the receptor bearing cell, but also in the ligand bearing one. The molecular mechanism of the latter, however, remains largely uncharacterized, and its role in Notch signaling feed-back and cell differentiation is still unknown.

Cross-talk with other signaling pathways. A PDZ binding motif has been identified in the cytoplasmic tail of some, but not all Notch ligands. The C-terminus of Jagged-1 has a highly evolutionarily conserved sequence (RMEYIV) that comprises a PDZ Class II recognition motif (φ-X- φ -COOH, where φ is a hydrophobic residue and X is any residue). Jagged-1 has been shown indeed to interact with the PDZ domain of the protein AF6 in a PDZ-dependent manner (Ascano et al., 2003). The C-terminal region of Delta-1 and -4 (VIATEV) also contain a PDZ binding motif, although of a different type (S/T-X- φ - COOH, a ligand for Class I PDZ domains). There is recent evidence that Delta-1 and -4 interact with the PDZ domains of Dlg1, the human homolog of the *Drosophila* Discs Large protein (Six et al., 2004). In other studies, the interaction between Delta-1 and members of the MAGUK family (Membrane Associated Guanylate Kinases) has been reported (Pfister et al., 2003; Wright et al., 2004). In contrast, the C-terminus of Delta-3 (ILSVK) and Jagged-2 (YAGKE) does not resemble PDZ

ligands. The presence of PDZ binding motifs, together with the experimentally confirmed interaction of Jagged-1, Delta-1 and -4 with PDZ containing proteins, suggest that Notch ligands are involved in a cell-autonomous, Notch-independent signal transduction pathway or, more intriguingly, that Notch signaling is coupled to other signaling networks (**Figure 5**). Dlg1 is a membrane-associated guanylate kinase involved in the maintenance of cell adhesion, cell polarity, growth control and cell invasion, and is essential for the assembly of multiprotein complexes at cell-cell junctions. AF6, together with E-cadherin/catenin belongs to an adhesion system that plays a role in the organization of cell-cell junctions. It can be then speculated that Notch ligands might also be involved in the cell adhesion system. How the RIP mechanism of proteolytic cleavage occurring in Jagged and Delta proteins can affect their interaction with the partner PDZ proteins remains unknown, as well as the role of Notch receptors in these interactions.

Notch signaling and endocytosis. The cytoplasmic tail of Notch ligands is involved not only in bi-directional signaling and interaction with PDZ containing proteins, but also in ligand internalization. Although in some instances soluble forms of DSL ligands can activate Notch signals, normally an intact membrane anchored ligand is required for full activation (**Figure 6**). The current hypothesis is that after a receptor/ligand interaction is established, "receptor shedding" is required to expose the juxtmembrane region of the receptor to proteolytic cleavage (Kanwar and Fortini, 2004; Le Borgne et al., 2005; Le Borgne and Schweisguth, 2003). Receptor shedding would be promoted by endocytosis of the ligand/ECN complex, which is in turn triggered by mono-ubiquitination of the Delta ligand by the E3 ubiquitin ligase Neuralized. The precise role of ligand endocytosis in the context of Notch signaling however remains unclear. More E3 ubiquitin ligases are being identified, and it is possible that the different Notch ligands are specifically

recognized by different E3 ubiquitin ligases.

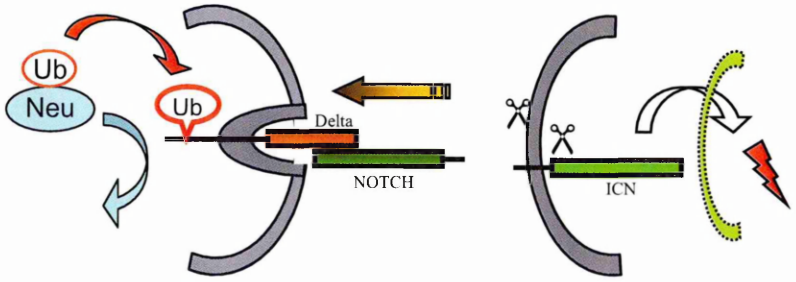


Figure 6 . Ligand endocytosis.

Notch signaling and cell-fate decisions. Notch signaling can have many different, if not opposite effects depending on the timing and the tissue context (Radtke and Raj, 2003; Weng and Aster, 2004). For example, while the maintenance of stem cells or progenitor cells in an undifferentiated state have been observed in the hematopoietic system and in the pancreas, terminal differentiation is induced in the skin by DLL1 or Jagged. In general, Notch signaling is acting on cell fate decisions either through lateral signaling or through inductive signaling (Artavanis-Tsakonas et al., 1999). In lateral signaling, equivalent, equipotent cells initially express both Notch receptors and their ligands, but the concentrations of these proteins start to differ between neighboring cells perhaps due to fluctuations in the steady-state expression levels. Small differences in receptor and/or ligand concentrations in cells are amplified over time, leading to cells that exclusively express either the receptors or their ligands, thus guiding the specification of the cell fate and cell differentiation. In inductive signaling, the interaction occurs between two developmentally distinct cells expressing exclusively either the receptor or the ligand. The fate of the bi-potential precursor cell is decided by the occurence of this interaction, while in the absence of Notch signal the precursor cell would follow another fate. The cell expressing the receptor, and therefore the recipient of the Notch signal, is induced

to differentiate into a particular cell lineage.

Notch signaling in development. Notch receptors and ligands are widely expressed during organogenesis in mammalian embryos, where they play a key role in establishing cell-lineage decisions in tissues derived from all the three primary germ layers: the endoderm (for ex. the pancreas), the mesoderm (skeleton, mammary gland, the vascular system and hematopoietic cells), and the ectoderm (neuronal cell lines) (Harper et al., 2003). In the pancreas, where different cell types appear with different timing, yet stemming from the same early cells, Notch-1 appears to delay both endocrine and exocrine development trapping progenitor cells in an undifferentiated state. In the presomitic mesoderm that will differentiate into the axial skeleton, muscles, tendons and dermis, Notch signaling plays a role as a molecular clock that controls regular segmentation of the mesoderm. Notch is also required in the later steps of vascular development, which includes proliferation and branching of the newly formed vessels. In the hematopoietic system, enforced activation of Notch-1 suppresses the differentiation of stem cells into myeloid, erythroid, or lymphoid lineages, and plays a role at a number of stages of lymphocyte development in the bone marrow and thymus. One of the essential functions of Notch-1 is the suppression of B cell development in the thymus. In the nervous system, Notch activation is required for the self-renewal of neural stem-cells, although it is not necessary for their generation.

Furthermore, Notch signaling controls the differentiation of glial cells and the length and organization of dendritic extensions from neurons (neurite arborization).

Notch signaling in cancer. At least two direct links between alterations in Notch signaling and human cancer have been established to date. A rare form of T cell acute lymphoblastic leukemia (T-ALL) is associated with a translocation that fuses the intracellular portion of Notch-1 with the promoter/enhancer region of the T-cell receptor beta locus, leading to constitutive activation of Notch-1 signaling (Screpanti et al., 2003). The majority of T-ALL cases have been recently asssociated with activating mutations in Notch-1 (Pear and Aster, 2004; Weng et al., 2004). Another chromosomal translocation, which is altering the function of Mastermind, a nuclear regulatory protein in the Notch signaling pathway, has been linked to mucoepidermoid carcinoma, a common type of malignant salivary gland tumor. High levels of the Notch ligand DLL1 have been observed in neuroblastoma cell lines. High expression levels of Notch have also been reported in some breast cancers and in human colon adenocarcinomas. Intriguingly, Notch can behave both as an oncogene or a tumor suppressor, depending on the cellular context and on the interactions with other signaling pathways.

Notch signaling in genetic disorders. The importance of the Notch pathway in cell fate control and development is further confirmed by the association of several diseases with mutations in genes involved in this complex signaling network (Gridley, 2003).

Alagille syndrome (AGS, MIM #118450) is a rare autosomal dominant disorder characterized by a variety of clinical abnormalities, including a reduction in the number of bile ducts eventually leading to the obstruction of biliary flow, and cardiac, musculoskeletal, ocular, facial defects. Although no clear genotype-phenotype correlation has been defined, AGS is caused by mutations in JAG1. While the majority of the mutations causing AGS are related to the generation of stop codons leading to unstable mRNA or truncated proteins, many missense

point mutations either introduce or delete cysteine residues that are critical for proper folding of the mature protein. Most of these mutations are located in the DSL domain and in the EGF tandem repeats.

Familial tetralogy of Fallot (TOF, MIM #187500) is the most common form of complex congenital heart disease (~1/3000 births). It is characterized by ventricular septal defects, obstruction to right ventricular outflow, aortic dextroposition and right ventriculat hypertrophy. A familial form of TOF was found to be associated with a missense G274D mutation occurring in the second EGF repeat of JAG1.

Spondylocostal dysostosis (SD, MIM #277300) is a vertebral malsegmentation syndrome characterized by multiple hemivertebrae, rib fusions and deletions. Mutations correlated with autosomal recessive SD have been identified in DLL3. Two of these mutations are expected to lead to truncated forms of the protein, while the third is a missense mutation in one of the EGF tandem repeats, G385D. Interestingly, this is the same kind of mutation observed in JAG1 and for which the genotype has been correlated to the TOF phenotype.

Cerebral autosomal dominant arteriopathy, with subcortical infarcts and leukoencephalopathy (CADASIL, MIM #125310) is associated with strokes and dementia. It is caused by mutations in the NTC3 member of the Notch receptor family. Most of the mutations involve the removal or insertion of cysteine residues in the EGF repeats and are likely to affect receptor folding, trafficking, maturation, or signaling.

Disease	Target	Description
Tetralogy of Fallot	JAG1	heart malformation: ventricular septal defect, pulmonary stenosis, displaced aorta, right ventricular hypertrophy
Alagille syndrome	JAG1	arteriohepatic dysplasia: paucity of biliary ducts in the liver, cardiovascular abnormalities affecting the

		great vessels
Spondylocostal dysostosis	DLL3	Jarcho-Levin syndrome: vertebrae and rib malformations
CADASIL	NTC3	cerebral autosomal dominant arteriopathy with subcortical infarcts, dementia
T-cell acute lymphoblastic leukemia	NTC1 NTC3	chromosomal translocation: TCR promoter – truncated Notch; mutations
Mucoepidermoid (salivary gland) carcinoma	MECT1 MAML2	chromosomal translocation: mect1-mastermind

Structural biology of Notch signaling. Very little is known about the detailed molecular mechanisms involved in Notch signal transduction. The structure of a NL (Notch/Lin12) repeat (Vardar et al., 2003), and the structure of the ligand binding region of Notch, encompassing three epidermal growth factor repeats (Hambleton et al., 2004), have been determined by NMR. The structure of Notch ankirin repeats have also been solved (Ehebauer et al., 2005; Lubman et al., 2005). Of the effector proteins, the structure of CSL bound to DNA has been recently solved by X-ray crystallography (Kovall and Hendrickson, 2004). Notch ligands are still awaiting structure determination. Most of the structural aspects that determine Notch functions remain as well uncharacterized. The interaction of Notch ligands with their receptors requires the DSL (Delta/Serrate Ligand) domain, but neither the structure of this domain nor the mechanism of binding has been determined. Notch signaling is sensitive to the concentration of extracellular calcium, but the effect of calcium ions on receptor and ligand structure have not been studied yet. Notch receptor/ligand recognition is modulated by glycosylation, but the structural determinants that regulate this interaction are not known. Other post-translational modifications, like beta-hydroxylation at aspartic or asparagine residues have been identified, but their role remains unclear.

AIM OF THE WORK

The signal transduction cascade initiated in the Notch bearing-cell by the proteolytic cleavage of the receptor and the release of the ICN from the membrane has been studied in detail, and several regions of Notch receptors, as well as some of the binding partners have been structurally characterized. On the contrary, very little is known on the side of the ligand-bearing cell. Most recent work has raised many issues about the role of the ligand-bearing cell in Notch signaling, and on the role of the cytoplasmic tail of Notch ligands in bi-directional signaling, in the cross-talk with other signaling pathways, in cell-autonomous, Notch-independent signaling, and in endocytosis-mediated receptor shedding. As experimentally derived structural data that could be give insight into the role of the Notch ligands intracellular region in signaling are still lacking, we applied bioinformatics methods to predict their structural and functional properties.

METHODS

General description of prediction approaches

Most problems in biological sequence analysis are related to the general approach of "prediction" in which we attempt to predict a property of a new sequence given a set of (positive and negative) examples. From the logical point of view this is a classification problem. Early methods of protein classification relied on pair-wise comparison of sequences, based on the alignment of sequences using exhaustive dynamic programming methods (Needleman-Wunsch, Smith-Waterman), or faster, heuristic algorithms (FASTA, BLAST). Pair-wise comparison yields a similarity measure that can be used to classify proteins on an empirical basis. The next generation of methods used generative models for the protein classes and similarity of a sequence to a class was assessed by a score computed between the model and the class. Hidden Markov Models (HMMs) are now routinely used in protein classification (SAM, HMMER). Discriminative models (such as artificial neural networks, support vector machines etc.) are used in a third generation of protein classification methods in which the goal is to learn the distinction between class members and non-members. Roughly speaking, 80-90 % of new protein sequence data can be classified by simple pair-wise comparison. The other, more complicated techniques are used mostly to verify if a new sequence is a novel example of an existing class or it represents a truly new class in itself. As the latter decisions refer to the biological novelty of the data, there is a considerable interest in new, improved classification methods.

While multiple-alignments, HMM models are immensely useful for analyzing evolutionarily related sequences, other fields of pattern classification mostly use simple vector/based descriptions. In this generalized framework a property is called

a "feature", and feature vectors are the structures that summarize the frequency (or % frequency [0,1], occurrence [0 or 1]) of the selected property within an object. As opposed to sequences and 3D descriptions, vectors provide an unstructured description of the objects, which is highly dependent on the – often arbitrary – choice of the components. Nevertheless, vector computations are fast and well elaborated, so vector descriptions are used for problems where structured descriptions can not be provided. Simple descriptions like amino acid composition or dipeptide compositions give surprisingly good classification performance in a number of applications.

The general scheme followed in this work is depicted in **Figure 7.**

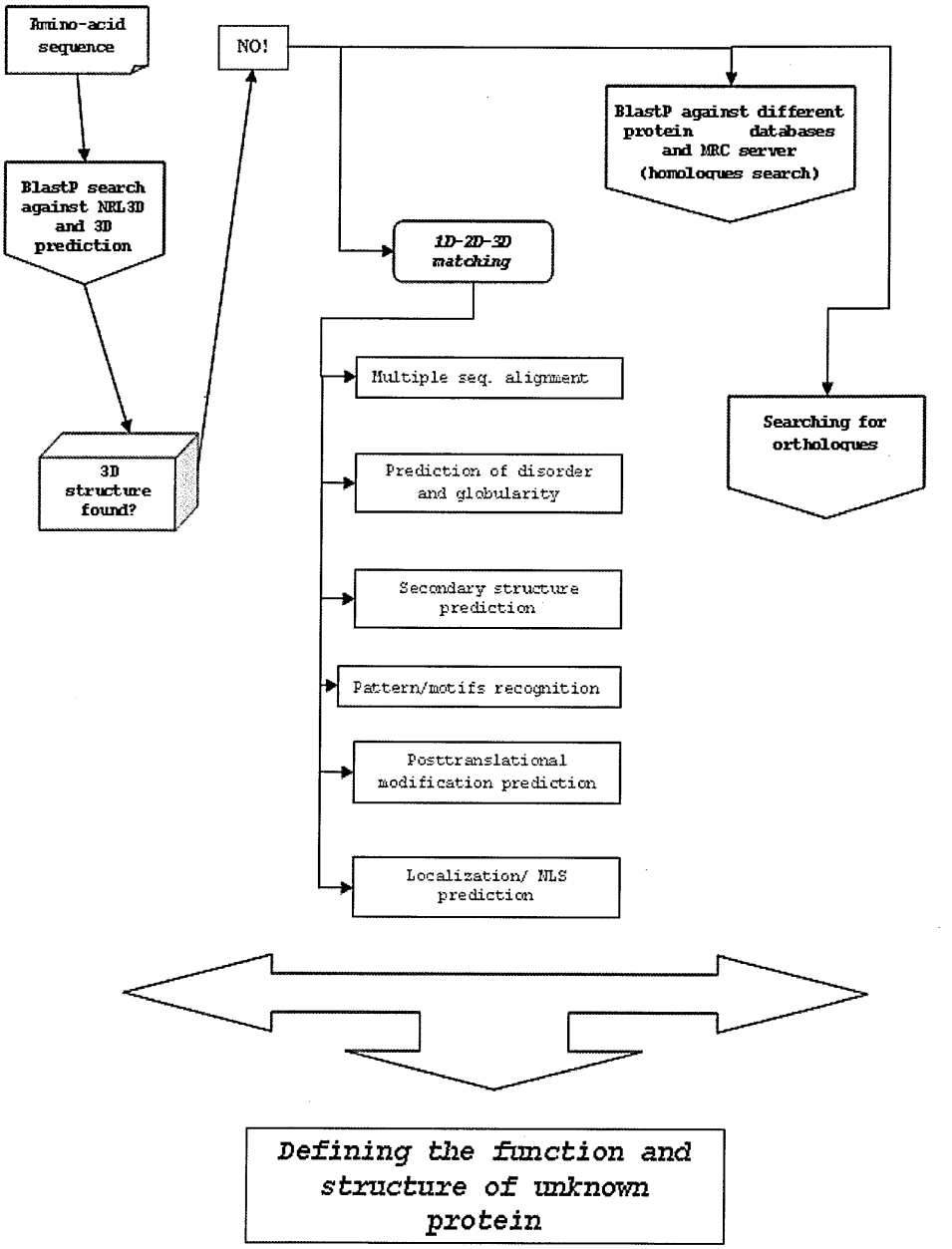


Figure 7. A flowchart for predicting structure and function from protein sequences by using bioinformatics techniques.

Identification of Jagged and Delta ligands.

The intracellular region of the human proteins (SW: DLL1_HUMAN, DLL4_HUMAN, DLL3_HUMAN, JAG1_HUMAN, JAG2_HUMAN) were used as seeds for BLASTP searches in the genomic databanks at NCBI and EMBL to find mammalian homologues (*Mus musculus, Rattus norvegicus, Bos taurus, Pongo pygmaeus, Pan troglodytes, Macaca fascicularis, Felis catus, Canis familiaris, Ovis aries*). Other entries were found searching organism-specific (*Gallus gallus, Xenopus laevis, Cynops pyrrhogaster, Brachidanio rerio, Tetraodon nigroviridis, Drosophila melanogaster, Glomeris marginata, Apis mellifera, Anopheles gambiae, Strongylocentrotus purpuratus, Lytechinus variegatus, Ciona savignyi, Halocynthia roretzi*) protein databases (RefSeq at NCBI; Swiss-Prot + trEMBL at EXPASY) using default BLASTP parameters (BLOSUM62 score matrix (among the best for detecting most weak protein similarities), SEG filter for low complexity regions (Low-complexity sequence can often be recognized by visual inspection. Filters are used to remove low-complexity sequence because it can cause artifactual hits), Expect value cut-off: 10. This setting specifies the statistical significance threshold for reporting matches against database sequences. The value (10) means that 10 such matches are expected to be found merely by chance, according to the stochastic model of Karlin and Altschul (1990). If the statistical significance ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported). Only sequences that could be aligned over the full length of the intracellular region were retained. Additional entries were found searching the Pfam database for all proteins containing either the MNNL (Notch ligand, N-terminal) or the DSL (Delta/Serrate/Lag-2) domain and cross-checking with the entries found in the sequence databanks (Appendix 1).

Multiple sequence alignment and phylogenetic analysis.

The intracellular regions of Jagged and Serrate proteins were aligned using ClustalW (score matrix: Gonnet 250, penalty for gap opening, -10; penalty for gap closing, -1; penalty for gap extension, 0.2; penalty for gap separation, 4) run from the EBI web server (Appendix 1). Phylogenetic trees were generated using the neighbor joining algorithm as implemented in ClustalW and drawn using PhyloDraw (Choi et al., 2000).

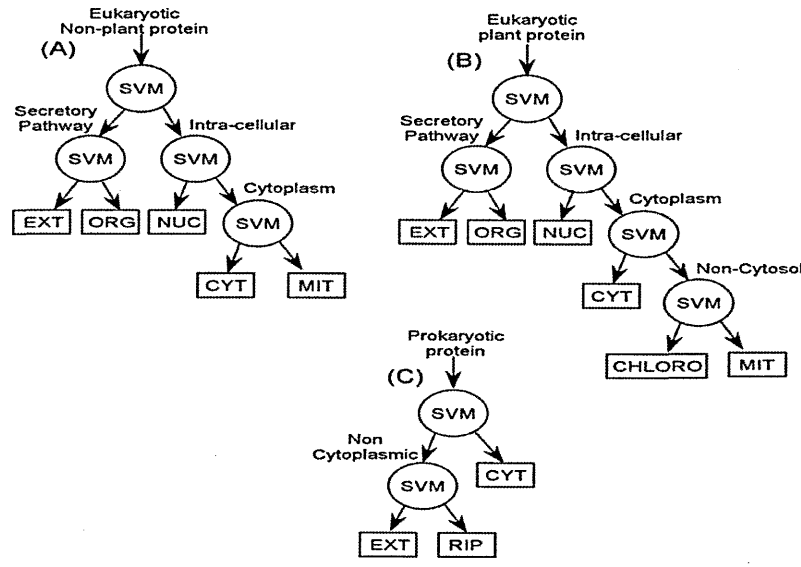
PhyloDraw is a unified viewing tool for phylogenetic trees. PhyloDraw supports various kinds of multi-alignment formats (and pairwise distance matrix) and visualizes various kinds of tree diagrams, e.g. rectangular cladogram, slanted cladogram, phylogram, unrooted tree, and radial tree. By using several control parameters, users can easily and interactively manipulate the shape of phylogenetic trees.

Cellular localization.

The prediction of cellular localization is a very typical example of a difficult biological prediction problem. Sorting of proteins into cytoplasmic, membrane-bound or extracellular compartments has different rules in various organisms, and there is no reason to suppose that proteins targeted to the same compartment would share evolutionary origins. So from this point of view, this is a typical field where high-dimensional unstructured descriptions can be used, and in fact many methods use amino acid compositions and other simplistic feature vectors for

classifying proteins. On the other hand, protein sorting is based on well/known molecular signals, such as signal peptides, nuclear localization signals, which are more-or less defined in terms of their sequence even though quite viable between organisms. Identification of such signals with structured models (sequence patterns etc) is an approach that is different from the unstructured models mentioned above. Current methods use a combination of unstructured and structured descriptions. We used a family of these servers developed by Burkhart Rost and collaborators at Columbia University.

LOCtree (Nair and Rost, 2005) is based on a multidimensional, unstructured feature-vector description of the proteins combined with Support Vector Machine (SVM) learning algorithms. The input is a sequence which is described in terms of i) amino acid composition (20 units), ii) composition of the 50 N-terminal residues (20 units), and iii) amino acid composition in the three secondary structure states (60 units). For the eukaryotic plant and non-plant systems, raw output from the SignalP signal-peptide prediction server is used as an additional input, so the final input is a blend of unstructured (amino acid composition) and structured (signal peptide) information.



The prediction is based on SVM, and the originality of the algorithm is the use of a hierarchical decision scheme that follows the logic of sorting pathways in binary decisions ("hierarchical SVM"). At each point of the hierarchy there is a binary decision taken by an SVM learner as shown in the sketch above. The system is first trained by a well-selected set of sequences of known cellular localization. The selection of training sequences is a key element, training itself is time-consuming but is essentially automated. Prediction on the other hand is very fast, secondary structure prediction, signal peptide prediction is not time consuming, amino acid compositions are rapidly computed and also SVM classification has a very low time requirement.

PredictNLS (Nair and B Rost, 2005) is an automated tool for the analysis and determination of Nuclear Localization Signals (NLS).

Nuclear localization signals (NLSs) are semi-conserved short stretches of amino acids known to be associated with nuclear import. Even though one can construct simple sequence motifs that will identify some of the known NLS sequences, the accuracy is not sufficient. The PredictNLS server of Rost and associates shows an

interesting approach to solve this difficult biological prediction problem. In addition to NLS sequences being very diverse, few of them are well characterized by experiment. Same as with LocTree, Rost and associates used additional biological knowledge to improve the prediction. The collected the experimentally sequences associated with the importin and transportin pathways of nuclear transport, grouped them into sequence families. These families were then extended based on sequence similarity using a strict criterion so that a high similarity to the experimentally tested sequences remains conserved. These collections were then based to extract local sequence features that can be used to scan sequence for potential NLS signals.

Fold recognition.

Fold recognition trials for the intracellular region of human Jagged and Delta proteins were run from the 3D-PSSM web server (Kelley et al., 2000) and its more recent version **PHYRE**.

Structural characterization of proteins is one of the ultimate goals of protein research. Current methods of protein structure determination such as X-ray crystallography and NMR are not ready yet to analyze multidomain proteins similar to the ones studied here. Single-domain proteins and expressed domains of multidomain proteins are relatively easily amenable to structural analysis so there is a rapidly growing body of data on protein domain structures. Simply put, the shape of the main-chain of a domain type is called a "fold", and classification of protein structures into folds is one of the traditional research areas of structural bioinformatics, characterized by such landmark databases as SCOP and CATH. Folds are characterized based on secondary structure and size (e.g. the SCOP

hierarchy includes alpha, alpha and beta, alpha/beta, small protein categories). The classification of folds is hierarchical, e.g. the main levels of CATH are class, architecture, topology, homology and sequence similarity. The lower levels of the hierarchy contain evolutionarily related groups that can be linked with the homologous protein families known in sequence classification, so structural families can be easily expanded to include sequence homologs presumably adopting the same fold. Given the wealth of information on fold groups one can design a large variety of structured and unstructured descriptions that will allow fold prediction at varying levels of accuracy. The main problem of this prediction task is that common folds are known to occur in many, evolutionarily divergent protein families, so there may be very little sequence similarity between proteins having the same fold. The default solution to this problem is to collect more and more sequences for all sequence groups adopting the same fold and so a similarity/based prediction can be relatively easily designed to cover all known variants of a given fold. Naturally, the generalization to novel sequences is not guaranteed with this approach.

The **3D-PSSM/Phyre** servers of Lawrence Kelley and Mike Sternberg are a good example of using highly structured data for prediction. The basis of the prediction are "profiles", multiple alignments obtained for known folds using sequence and 3D alignment methods using, in addition to 3D superposition, also secondary structure and 3D solvation potential (solvent accessibility) information. The 3D alignments are then complemented with unambiguously selected sequence homologs, and the resulting alignments are converted to profiles that are easily amenable to sequence similarity searching. Owing to the carefully build 3D-alignments this method can predict folds in cases when traditional search programs such as PSI-BLAST are of no help.

Globularity prediction.

Disordered regions and their prediction are relatively new additions to the repertoire to the scope of bioinformatics. While most of structural research and the associated prediction methods concentrate on well characte4rized globular proteins, it is well known that a large percentage of proteins does not adopt a detectable structure in solution. Apart from the well known fibrillary proteins characterized by characteristic repetitive sequences (such as collagen, keratin, etc.), there are non-globular parts in a large variety of proteins, and the sequence of theses non globular segments is highly variable between protein families. From the point of view of prediction, the problem is roughly analogous to cellular localization prediction, since the sequences are varied and there are only broad compositional principles that distinguish the sequences from those of globular proteins. Nevertheless the prediction can be approached by the same principles. Additionally one can use information on sequence complexity (the Seg program of John Wootton) because disordered sequences are also known to be of low complexity.

Predictions of globularity and order/disorder for the intracellular region of human Jagged and Delta proteins are run using GLOBPLOT (Linding et al., 2003b), PONDR® ® (Romero et al., 2004), DISEMBL (Linding et al., 2003a), IUPRED (Dosztanyi et al., 2005) and COILS (Lupas, A.,at all,1996).

The **GlobPlot** server of Rune Linding et al uses a variant of traditional secondary structure prediction that is based on amino acid propensities. A propensity of an amino acid residue can be calculated from the frequency of the given residue type within a given structure. Linding and associates have used the traditional Chou Fasman approach to calculate propensities for the disordered state. In the first approximation, the random coil state of the Chou Fasman algorithm may be used to

predict disordered region, and this approach was improved by the authors by combining the propensities into "secondary structure" (helix, strand, turn), and "disordered" (coil). The algorithm produces plots for various propensities and makes predictions by identifying the peaks within the plots. The prediction can be improved by also analyzing the known globular domains at the same time and limiting the prediction to those areas where globular motifs are not found. VSL1 combines two predictors optimized for long (>30 residues) and short (<=30 residues) disordered regions, respectively; VL3 is a neural network predictor trained on 152 long regions of disorder that were characterized by various methods and a set of ordered proteins consisting of 290 PDB-Select-25 chains having no disordered residues; VL-XT integrates three feed-forward neural networks: VL1, the N-terminus predictor (XN), and the C-terminus predictor (XC); XL1 is a neural network predictor optimized to predict regions of disorder greater than 39 amino acids, and was trained on 7 disordered regions identified from missing electron density in X-ray structures; CaN is a neural network predictor that was trained on regions of 13 homologous calcineurin proteins.

The **PONDR®** server is based on a machine learning algorithm, feed-forward neural networks that use sequence information from windows of generally 21 amino acids. Attributes, such as the fractional composition of particular amino acids or hydropathy, are calculated over this window, and these values are used as inputs for the predictor. The neural network, trained on a specific set of ordered and disordered sequences, then outputs a value for the central amino acid in the window.

DisEMBL uses different order/disorder definitions. The Loops/Coils definition is based on the assignment of a secondary structure state other than helix or strand as disordered; the Hot Loop definition is based on Loops/Coils residues that display a

high crystallographic B factor; the Remark-465 definition (missing coordinates in the PDB file) is based on residues that show no electron density in X-ray structures.

The **IUPRED** algorithm is a propensity-plot type predictor, which is technically similar to the GlobPlot server, however it uses a different amino acid scale that estimates the interaction-forming i.e. structure/stabilizing propensity of the amino acids. This property – the interaction propensity -can be estimated for amino acid pairs in the globular protein structures using distance cutoff limits. When plotting interaction propensities along the proteins, ordered and experimentally known disordered regions give different pictures which allows one to predict these regions with some confidence

The COILS server predicts coiled-coil regions characteristic of many protein families. COILS is a program that compares a sequence to a database of known parallel two-stranded coiled-coils and derives a similarity score. By comparing this score to the distribution of scores in globular and coiled-coil proteins, the program then calculates the probability that the sequence will adopt a coiled-coil conformation.

Secondary structure prediction.

Protein secondary structure prediction is one of the traditional fields of bioinformatics which has been tackled by a very large variety of computational tools. From the conceptual point of view secondary structure elements are one of the most difficult to predict since the have no appreciable sequence conservation. On the other hand, there are large numbers of experimentally known structures that make the development of SS prediction a challenging field of research. The early methods

used "propensities" – numerical constants derived from the frequency of an amino acid or amino acid pair to occur in

JPRED is a web server that takes a protein sequence or multiple alignment of protein sequences, and from these predicts secondary structure using a neural network called Jnet. The prediction is the definition of each residue into either alpha helix, beta sheet or random coil secondary structures. For single sequences a multiple alignment is constructed. It is created by the PSI-BLAST algorithm with 3 iterations. The prediction algorithms use two tandem/connected neural networks that scan the alignment with a window and output the secondary structure prediction for each window position. The algorithm uses a jury of neural networks for decision.

PSI-PRED is similar to **JPRED** in as much as it is a secondary structure prediction method based on two feed-forward neural networks which run on a PSI-BLAST alignment. The current version of PSI-PRED includes a new algorithm which averages the output from up to 4 separate neural networks in the prediction process to increase prediction accuracy.

SSpro secondary structure prediction is based on an ensemble of bidirectional recurrent neural networks (BRNNs). BRNNs are graphical models that learn from data the transition between an input and an output sequence of variable length. The model is based on two hidden Markov chains, a forward and a backward chain, that transmit information in both directions along the sequence, between the input and the output sequences. Three neural networks are then used to analyze the signals and output the predictions.

Pattern recognition and Phosphorylation.

Predictions of functional sites for the intracellular region of human Jagged and Delta proteins are obtained from ELM (Puntervoll et al., 2003) restricting the search to *Homo sapiens* and the cellular compartment to either *plasma membrane, cytoplasm,* or *nucleus.* Potential phosphorylation sites are identified using DISPHOS (Iakoucheva et al., 2004), NetPhos (Blom et al., 2004) , Yin-Yang sites (R. Gupta, S. Brunak and J. Hansen, 2003) .

Sequence patterns are perhaps the simplest and the first important representation tools to describe conserved sites within sequences. The resulting descriptions are in the form of regular expressions and are loosely termed as motifs or patterns. Since the publication of the first collection of patterns, PROSITE, there were many different methods designed for extracting motifs from sequences or finding them in sequences. This subject belongs to one of the best elaborated fields of computer science in general and bioinformatics in particular, so its full description would be beyond the scope of this thesis. Regular expressions are extremely efficient tools but have the well known draw-back that a single mismatch can either block the prediction or bring in a very large number of false positives. Nevertheless, may simple sites in proteins, like those of posttranslational modification and enzymatic digestion, can be quite accurately found using regular expressions.

ELM is an Internet resource for predicting functional sites in eukaryotic proteins. Putative functional sites are identified by patterns (regular expressions). Context-based rules and logical filters are applied to reduce the amount of false positives.

Phosphorylation sites can be regarded as one of the posttranslational modification sites that are located with regular expression search. Phosphorylation sites are quite variable, which results in a low prediction accuracy. Because of the pivotal

role of phosphorylation in signal transduction and other biological processes, there are a number of dedicated methods that serve the prediction of phosphorylation sites.

DISPHOS uses a machine learning algorithm called Logical Regression (LogReg) to predict phosphorylation site. This is a discriminative method that is able to learn differences between positive and negative instances, in this case phophorylated S,T or Y residues and their non phophorylated counterparts. The input is a 25 residues window centered around an S,T or Y residue, and is encoded in terms of residue occurrences within individual positions of the window (24x20=480 binary features), 20 relative amino acid frequencies, as well as the prediction results for the window calculated by various disorder and secondary structure algorithms . This is a highly varied feature set and LogReg is an efficient tool to handle such varied input.

NetPhos is an artificial neural network based method that analyzes a 25 residue window centered on a potential phosphorylation site represented in terms of amino acid frequency and positional information analogous to that used by DisPhos in conjunction with the LogReg algorithm.

Yin-Yang sites are those that can be alternatively phosphorylated or glycosylated. The YinOYang WWW server produces neural network predictions for O-ß-GlcNAc attachment sites in eukaryotic protein sequences. The principle is similar to NetPhos and in fact this server can also use the NetPhos server for the analysis. O-(beta)-GlcNAcylation is a dynamic post-translational modification that affects a large number of nuclear and cytoplasmic proteins. Such sites may be reversibly and dynamically modified by O-GlcNAc or Phosphate groups at different times in the cell. In some cases, a reciprocal relationship may exist with phosphorylation on the same Ser/Thr residues. The spread of O-(beta)-GlcNAcylation is known to be reciprocal with phosphorylation. Predicted O-(beta)-

GlcNAc sites were found in over half of all SwissProt human sequences, 65% of which were nuclear or cytoplasmic.

All used methods are presented in Table 1.

Sequence databases

	Database
Protein databases:	Swissprot - the main curated protein database
	SPTR- non-redundant set of Swissprot & TrEMBL
	TrEMBL - automatic translation of EMBL based on the annotation ofcoding regions
	IPI- complete sets of human, mouse and rat proteins
	PIR- functionally annotated protein sequences
	NRL-3D- sequences of known 3D structures
	RefSeq Protein- a biologically non-redundant collection of proteinsequences
Nucleic Acid databases	EMBL - the complete set of known sequences including HTGs, ESTs, STSs, GSSs
	RefSeq - biologically non-redundant set of DNA and RNA sequences
	Tetraodon nigroviridis Genome
Sections of EMBL	Bacteriophage
	Fungi
	Invertebrates
	Other Mammals
	Other Vertebrates
	Patent Sequences
	Plants
	Viral
	ESTs
Species extracted from EMBL	Oryctolagus cuniculus (Rabbit)
	Rattus spp. (Rat)
	Bos taurus (Cow)
	Ovis aries (Sheep)

Domain databases

Software	Reference	URL
Pfam	Alex Bateman et al,2004	http://www.sanger.ac.uk/Software/Pfam/
SMART	Letunic I, eta al , 2004; Schultz, J., et al ,1995	http://smart.embl-heidelberg.de/
Prodom	F Corpet et al,2000 Catherine Bru et al,2005	http://protein.toulouse.inra.fr/prodom/
SBASE	Vlahovicek et al, 2003	http://hydra.icgeb.trieste.it/sbase/

Homology searches

Software	URL
NCBI blast against PDB	http://www.ncbi.nlm.nih.gov/BLAST/

NCBI blast against nr(all databases)	http://www.ncbi.nlm.nih.gov/BLAST/
Uniprot blast against UniRef100	http://www.expasy.org/tools/blast/
Uniprot blast against the UniProt knowledgebase	http://www.expasy.org/tools/blast/
Uniprot blast against against all EMBL + GSS (without GTG and ESTs)	http://www.expasy.org/tools/blast/

Multiple sequence alignments

Software	Reference	URL
ClustalW	Thompson, J. D. et al,1997	http://www.ebi.ac.uk/clustalw/

Cellular localization

Software	Reference	URL
LOCtree	R Nair and B Rost, 2005	http://cubic.bioc.columbia.edu/services/loctree/
PredictNLS	R Nair and B Rost, 2005	http://cubic.bioc.columbia.edu/predictNLS

Fold recognition

Software	Reference	URL
3D-PSSM	Kelley LA et al , 2000	http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html

Globularity/disorder prediction

Software	Reference	URL
GLOBPLOT	Rune Linding ,et al,2003	http://globplot.embl.de/
PONDR®	Romero, P., et al,2001 Li, X., et al,1999	http://www.pondr .com/
DISEMBL	Rune Linding[1] et al,2000	http://dis.embl.de/
IUPRED	Veronika Csizmók, et all, 2005	http://iupred.enzim.hu/
Coils	Lupas, A., at all,1996	http://www.ch.embnet.org/software/COILS

Secondary structure prediction

Software	Reference	URL
PHYRE(Protein Homology/analogY Recognition Engine)		http://www.sbg.bio.ic.ac.uk/phyre/
PsiPred		McGuffin LJ,et al.2000; Jones DT,et al, 1999
Jnet		Cuff J. A and Barton G.J , 1999
SSpro		J. Cheng, et al, 2005

Pattern recognition

Software	Reference	URL
ELM	Puntervoll, P., et al,2003	http://elm.eu.org/
Prosite	Hulo N., et al,2004	http://www.expasy.org/prosite/

Phosphorylation sites

Software	Reference	URL
DISPHOS	Lilia Lakoucheva et al,2004	http://core.ist.temple.edu/pred/
NetPhos	Blom, N., et al,1999	http://www.cbs.dtu.dk/services/NetPhos/
Yin-Yang prediction	R. Gupta, S. Brunak and J. Hansen, 2003	http://www.cbs.dtu.dk/services/YinOYang/
SignalP	Henrik Nielsen, et al,1995	http://www.cbs.dtu.dk/services/SignalP/
Metal binding potential	Fredj Tekaia, Edouard Yeramian and Bernard Dujon, 2002	http://www.pasteur.fr/~tekaia/aafreq.html

Table 1. Methods. Databases and bioinformatics tools used

RESULTS*

IDENTIFICATION OF JAGGED AND DELTA LIGANDS

Searches of databases for homologues of human Jagged and Delta intracellular region and orthologues of human Notch ligands led to a collection of sequences shown in Appendix 3.

As expected, Notch ligands can be found in all phyla of multicellular organisms, including mammals, birds, amphibians, fishes, insects, echinoderms, chordates, and nematodes.

One can see that the Jagged family appeared in Metazoa and there is only one type, the Jagged 1 / Jagged 2 division. The intercellular part of the ligand doesn't exist (or has not been found) till the appearance of Insects, and even after this point, the protein exists only in one type. Jagged1 and Jagged 2 appeared for the first time in Fish. The Delta family also seems eto have appeared in Metazoa as one type. The difference is that before the emergence of Insects this ligand has only an intercellular part. A division is visible in Fish, but only two types – Delta 1 and 4 seem to exist. As one can notice, the sequence of Delta 3 is obviously shorter and has quite adifferent amino acid composition. It is highly probable that this difference affects both its structure and function, as well as its evolutionary fate. Delta3 appeared in Fish, and is not well conserved.

* All further results and predicted figures by the used programs, which are not shown in the section Results are in Appendix 4.

MULTIPLE SEQUENCE ALIGNMENTS AND PHYLOGENETIC ANALYSIS

One of the first steps of protein family analysis is to find common elements (conserved regions, common motifs, conserved residues) that are shared by the majority or by all the members of a protein family. This strategy is more promising if the proteins studied are closely related, i.e. there are no major differences between them such as domain deletions, additions, etc. For such simple cases multiple alignment programs represent a good approach. The proteins we studied are widely distributed in eukaryotes and their overall structure and function are seemingly conserved. So, we decided to use the CLUSTALW (complete results are in Appendix 4) algorithm as the first approximation since this program is known to perform well on related sequences.

The relationship between the sequences corresponding to the intracellular region of all identified Notch ligands, after multiple sequence alignment and clustering, are summarized in the form of a Cladogram in **Figure 8**. From this representation, the presence of relatively well distinct groups can be identified. The first group includes sequences similar to human Jagged, and can be divided into two sub-groups, the first comprising Jagged-1 (J1) and the second Jagged-2 homologues (J2). The second group includes sequences similar to human Delta-3 (D3). This group also includes *Drosophila* Serrate and Delta. The third group includes sequences similar

to human Delta-1 and -4, and can be divided into two sub-groups, the first comprising Delta-1 (D1) and the second Delta-4 homologues (D4). A fourth group includes sequences that are related to Delta, but that seem to be more distantly related (DX). The evolutionary distance between the different sequences can be better represented as a phylogenetic tree (**Figure 9**).

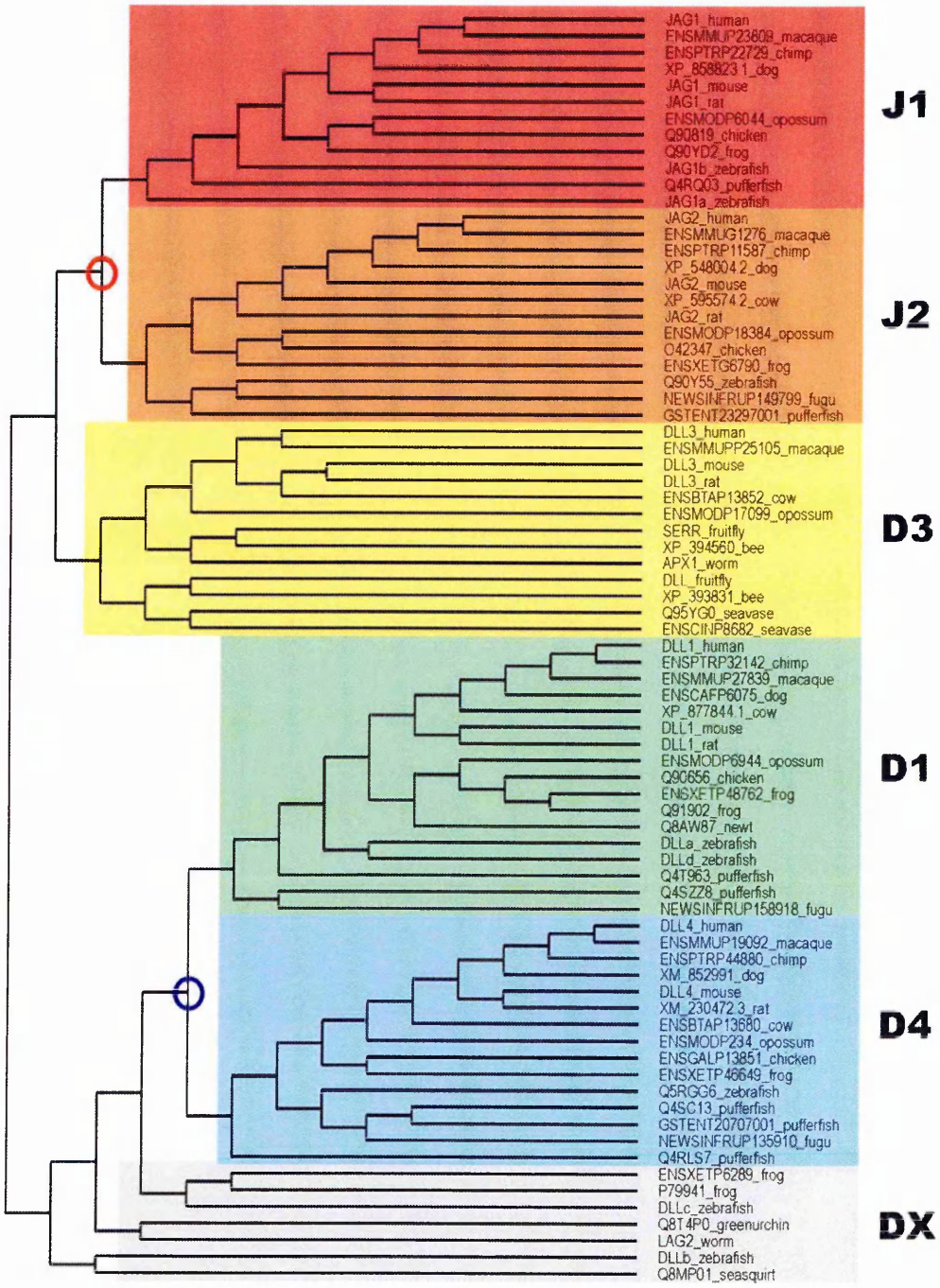


Figure 8. Phylogenetic analysis of Notch ligands. A cladogram representation generated from the multiple sequence alignment of Notch ligands intracellular region. Identified groups are labelled as J1, J2, D1, D4, D3, and DX, and colored accordingly. The branching points between J1 and J2 and between D1 and D4 groups are also labeled.

Given the above relationships e.g. midpoint rooting network places the root halfway between the two most distinct taxa. This method is based on the assumption that the amount of evolutionary change is proportional to time.

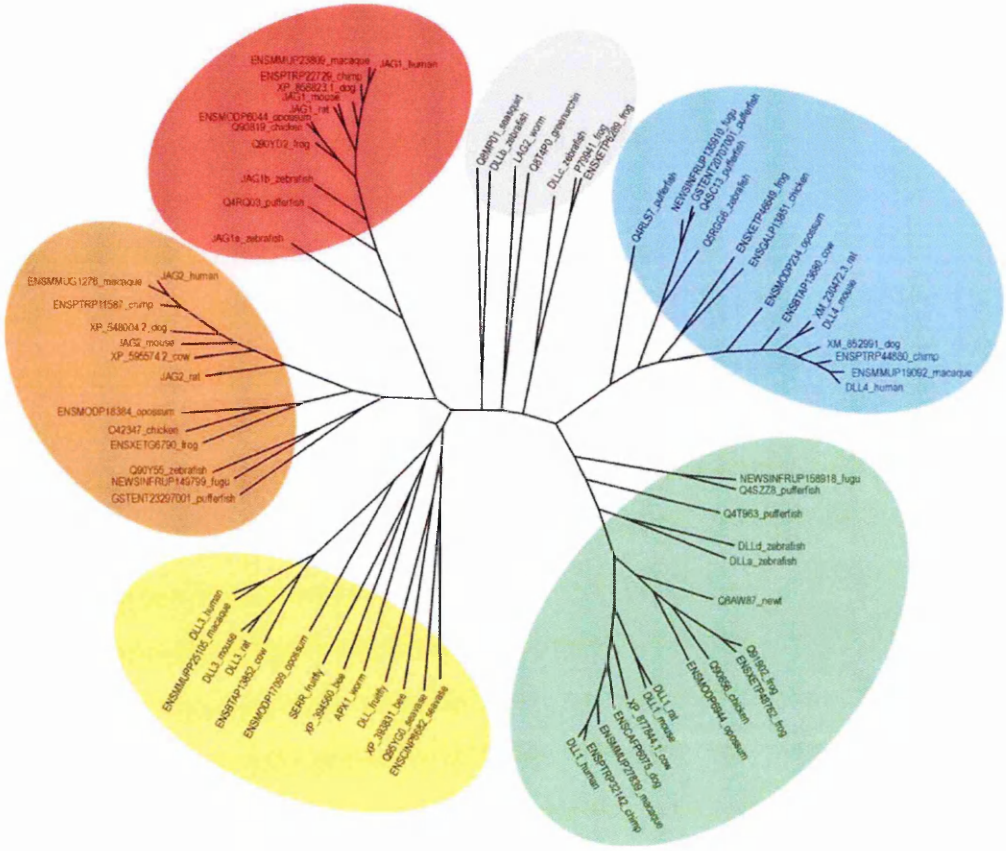


Figure 9. Phylogenetic analysis of Notch ligands. A phylogenetic tree representation generated from the multiple sequence alignment of Notch ligands intracellular region. Identified groups are colored as in Figure 8.

A more detailed analysis represented as a ClustalW output in Appendix 4 , confirms that the intracellular region of Notch ligands is evolutionary very well

conserved within the same ligand type, although the degree of conservation is more pronounced in the J1, D1, and D4 groups then, for example, in the J2 and D3 groups. In the J1 group (orthologues of human Jagged-1) the sequence conservation is very strict over the entire sequence length and through all the species from man to zebrafish. In the J2 group (orthologues of human Jagged-2) some degree of divergence can be observed going from mammals to fishes, but the sequence is still very well conserved within mammals, in primates as well as in mouse, rat, dog and cow.

In Jagged family – Jagged - 1 is very well conserve form Fishes to Human especially in few regions. One could speculate that few domains with a different structure and function could exist and define the entire role of Jagged -1 in the cell. The conservation in Jagged -2 is not that much. It is most obvious at the very beginning of the sequence. This could be cleavage site or signal for cellular localization..

In the D1 and D4 groups there is again a remarkable degree of conservation. The D3 group is the most divergent, together with the DX group, which includes some outliers. This is not surprising, however, because the D3 group includes not only mammalian ligands but also evolutionary very far phyla like insects and chordata. In fact, within the mammalian group of D3 ligands, the sequence is again rather well conserved. We expect that the similarity between Delta - 1 and -4 will be visible as similarity of the function and behavior in the cell for these two, much different from the ones of Delta -3.

LOCtree

Knowledge of cellular localization is a key element in characterizing a protein family. Prediction methods may not be absolutely accurate, but they still can be expected to provide a coherent picture on a protein family. I.e. members of the family would be predicted to be similar in terms of their target compartments and their localization. The intracellular region of Notch ligands is expected to protrude from the inner side of the plasma membrane into the cytoplasmic space. However, experimental reports suggest that these ligands are proteolytically cleaved and released from the membrane. Hence the interest of analyzing these fragments in terms of potential cellular localization (**Figure 10**).

Jagged -1 and -2 are predicted to be localized in the nucleus, but they are not listed as DNA-binding proteins, according to the R-index. For Jagged -1 - as far as predictions can be trusted) - the possibility to find it in the nucleus is almost 100%, the R index is 10, the reliability for the "No-DNA binding" is the same. The predictions for Jagged -2 are the same, but with low R-index (Figure 10). Nuclear Localization signal predicted by PredictNLS (Appendix 4) is found only for Jagged -2. At the beginning of the sequence this pattern is detected as - RKRRKE. Even if a low R-index used for the LOCtree program, the results allows one to speculate that the intercellular part of Jagged -2 could be cleaved from the membrane and then could bind to an importin, which could then carry it to the nucleus. Although Jagged-2 is not predicted to be a DNA- binding protein, it may still be part of a DNA-binding complex.

Delta -1 and -4 could be localized in the nucleus, even though this is more possible for Delta -1 then -4. The function is defined as DNA-binding and just the opposite, this is more reliable for Delta4 then -1, even if this is not so reasonable. Delta-3 seems to behave differently (as was expected), with a high score obtained for the cytoplasmic localization not secreted and not nuclear with not so high R index.

It may be in different cells, in different conditions all the ligands have different function and structure, that's why the localization prediction are so doubtful (**Figure 10**).

Protein	Localization	R-Index	Intermediate localization prediction	R-index of intermediate localization predictions
Jagged-1	Not DNA-binding	10	Not Secreted Nuclear Not DNA-binding	6 10 10
Jagged-2	Not DNA-binding	4	Not Secreted Nuclear Not DNA-binding	1 7 4
Delta-1	DNA-binding	1	Not Secreted Nuclear DNA-binding	1 9 1
Delta-3	Cytoplasmic	10	Not Secreted Not Nuclear Cytoplasmic	6 3 10
Delta-4	DNA-binding	4	Not Secreted Nuclear DNA-binding	6 4 4

Figure 10. Predicted sub-cellular localization of the human Jagged and Delta intracellular region. Both the final and the intermediate localization are shown, together with the corresponding reliability index (1 is min and 10 is max score).

FOLD RECOGNITION

Identification of known folds in a protein or protein family is one of the first steps of protein family analysis since there is wealth of structural, functional and biophysical data available for the various protein folds. Even though the domain composition of the proteins studied by us are generally known, we decided to run fold prediction programs in order to see whether or not the previously not annotated regions can be

assigned to one of the newly characterized folds. We used the 3D/PSSM/Phyre system of Lawrence Kelley available on line.

Similarity searches in the PDB and threading trials using 3D-PSSM gave no results. Both similarity scores from BLAST and E-values from 3D-PSSM were not significant.

GLOBULARITY

While identifying globular regions is an essential preliminary step in addressing structural studies of new multidomain proteins, intrinsically unstructured proteins and disordered regions are increasingly acknowledged to play an important functional role, especially in signaling networks. Because the intracellular regions of Notch ligands do not display any significant similarity with other known proteins, we used the programs that are currently available to predict globularity and order/disorder to detect any globular region in the cytoplasmic tails of human Notch ligands

As it was mentioned above the predictions were made using GLOBPLOT (**Figure 11**), PONDR® **(Figure 12)**, DISEMBL, IUPRED, COILS. The outputs of the last three methods are shown in Appendix 4.

For Jagged-1 and -2, results from different predictions methods consistently point to a disordered nature of the intra-cellular region. For Jagged -1 all the methods agree disorderness at the beginning of the sequence (10-30 AA) and for sure at the

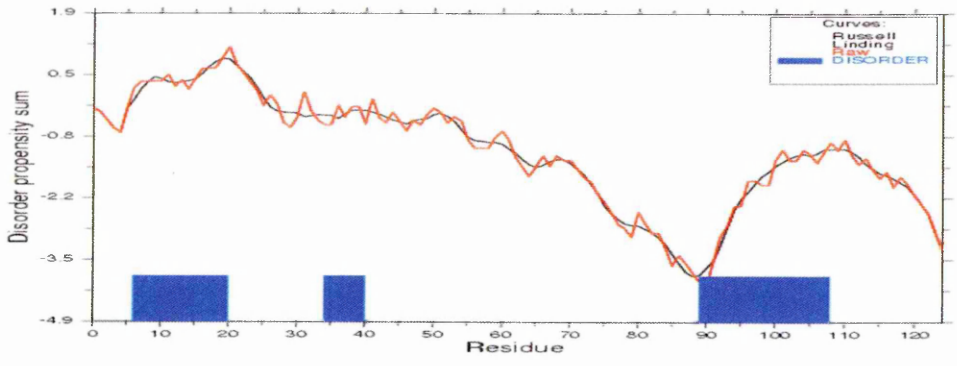
end of the sequence (90-110). For Jagged -2 the "agreement" shows three disorder regions beginning (20-40), middle (70-80) and again the end (95-110).

For Delta proteins, prediction results are more complex. Delta-1 is predicted to be mainly disordered in its C-terminal half (residues ~75-150), but to have several globular regions in its N-terminal part. Delta-3 is predicted to be mainly unstructured in its 1-70 regions, but its mean charge/mean hydropathy ratio is compatible with values found in globular proteins, and the C-terminal region is likely to be less disordered. Also Delta-4 is predicted to be largely unstructured (residues ~10-80), with perhaps the exception of its C-terminal region. Combined results are visualized in Appendix 4.

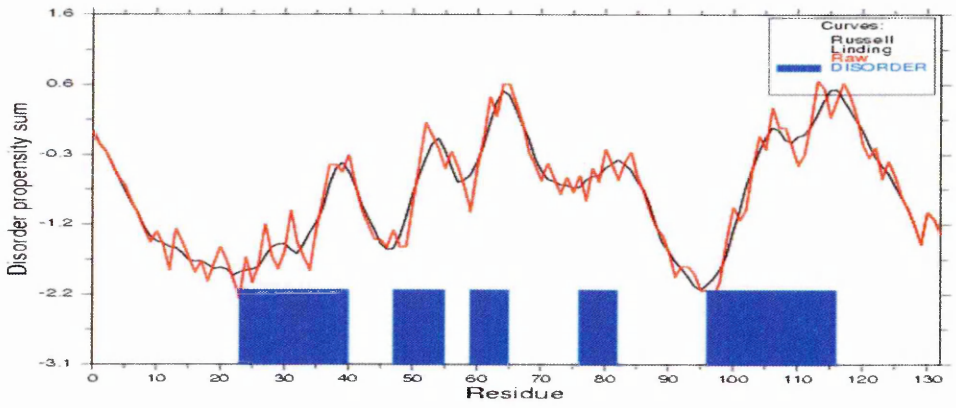
GLOBPLOT

Disorder score is calculated using the Russel-Linding disorder propensity (red) and plotted against the residue number. A smoothed curve (black) is also shown. Uphill regions are predicted to be disordered and are highlighted in blue, downhill regions are predicted to be globular and are highlighted in green. No threshold is defined.

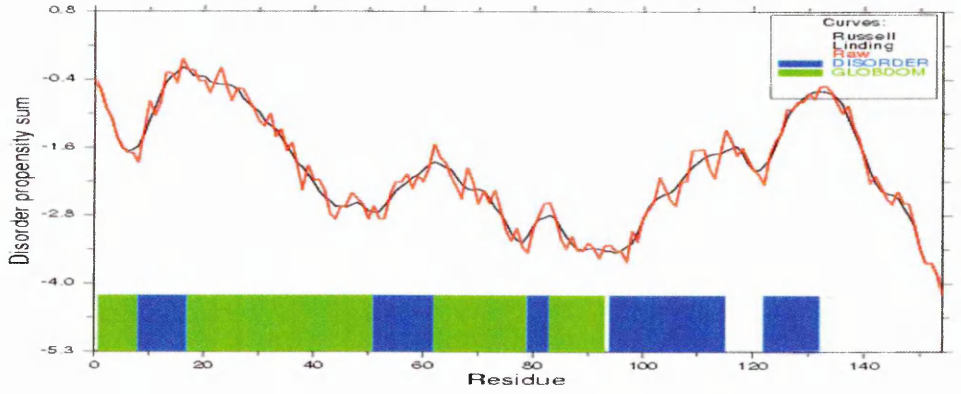
Jagged -1



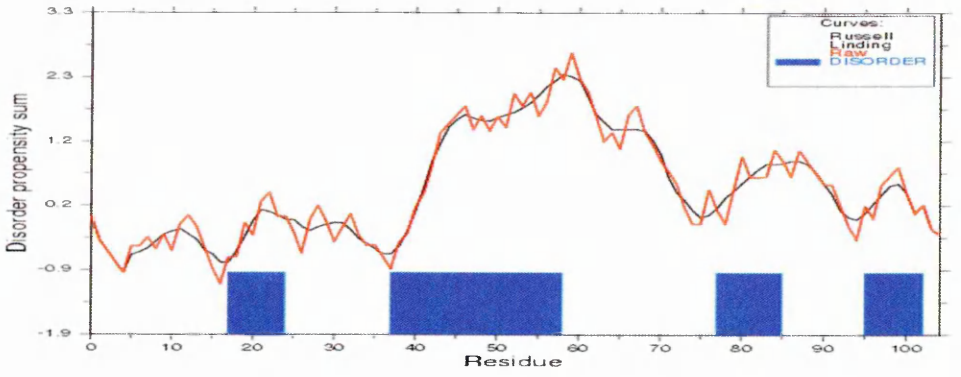
Jagged -2



Delta -1



Delta -3



Delta-4

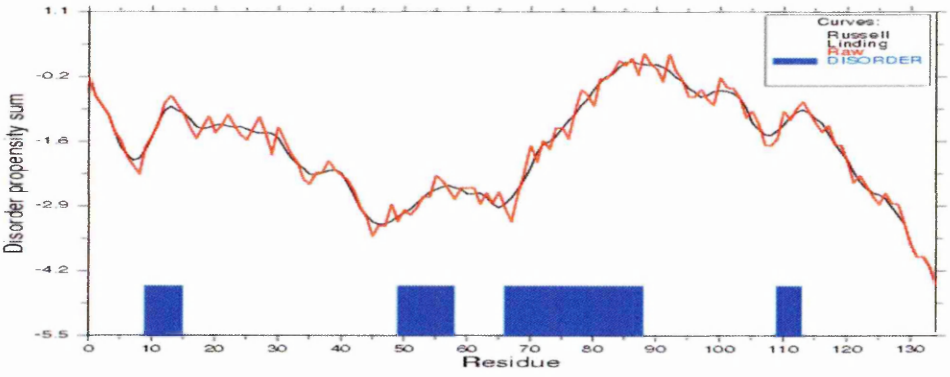


Figure 11. Protein disorder predicted by GLOBPLOT.

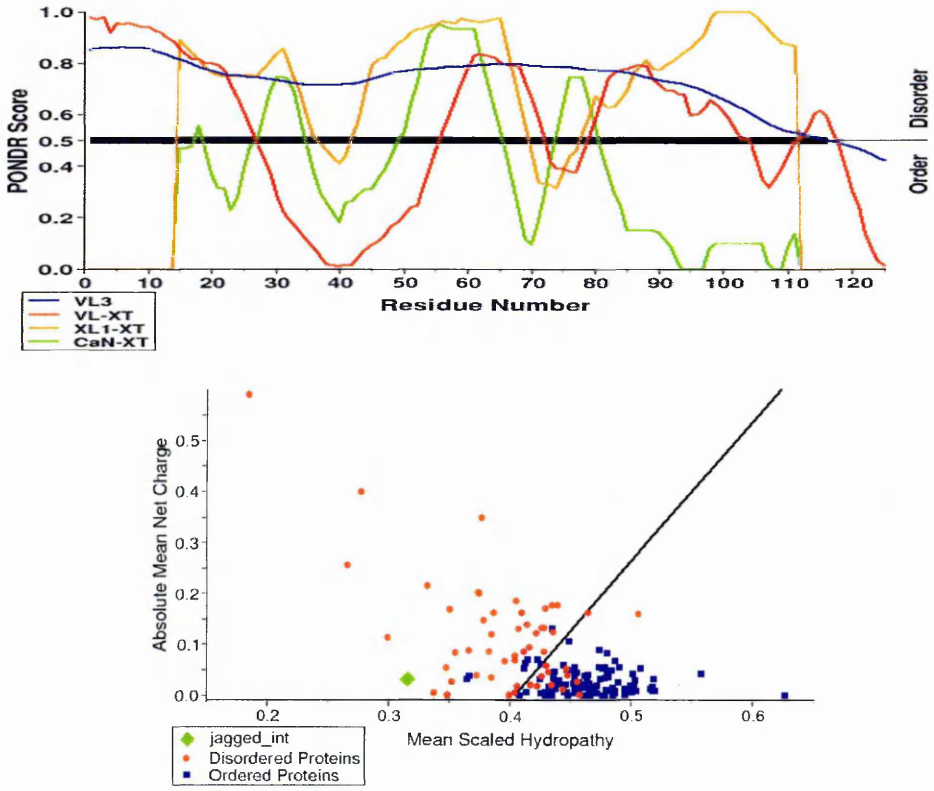
PONDR®

The score is calculated using different predictors and plotted against the residue number. The charge-hydropathy plots compare the absolute, mean net charge and the mean, scaled hydropathy.

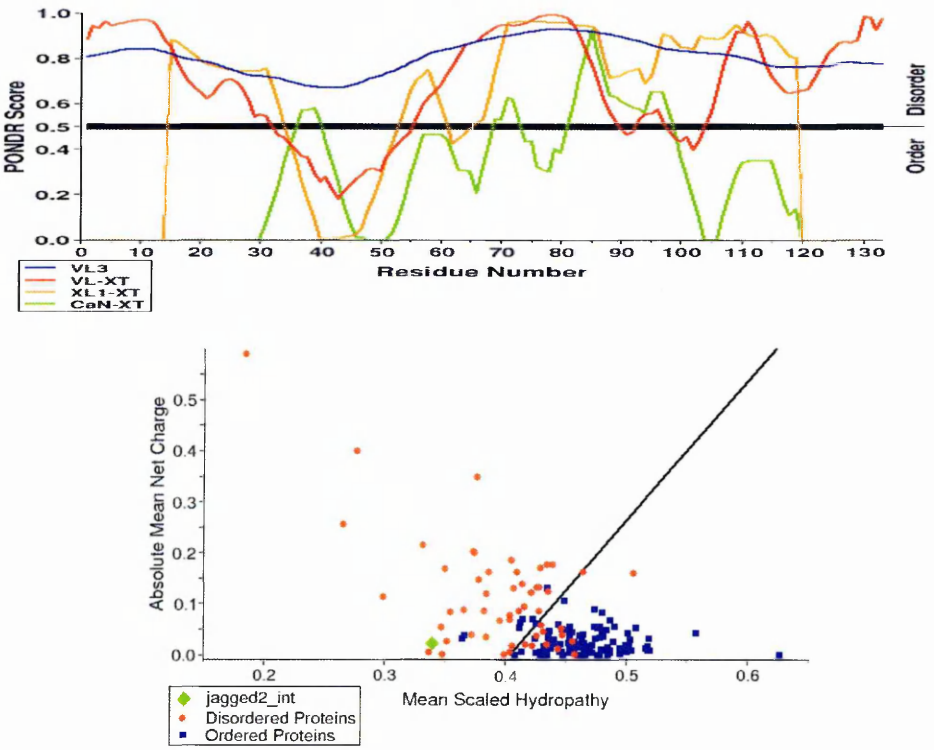
The firs plot visualized the combined prediction of all the methods combined in PONDR. The threshold is 0.5 and the regions predicted over the score are disordered.

The second plot shows two planes, the left is of the disordered proteins and the right one is of the ordered proteins. The unknown protein e.g. Jagged or Delta is visualized in green and is positioned in the plain where it belongs in the base of the prediction. Only Delta-3 is predicted to be ordered, there rest 4 proteins are disordered.

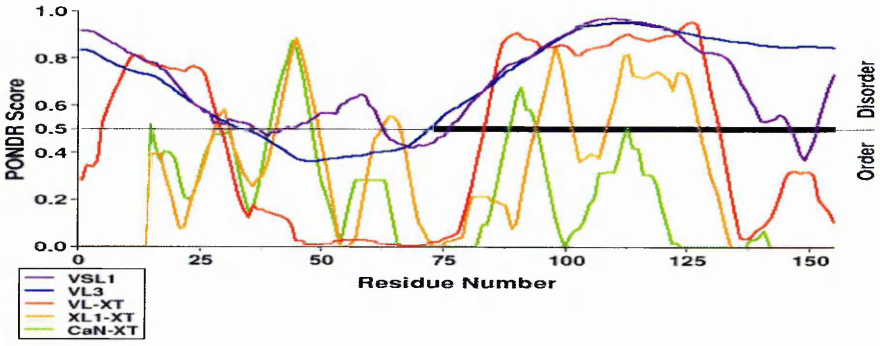
Jagged -1

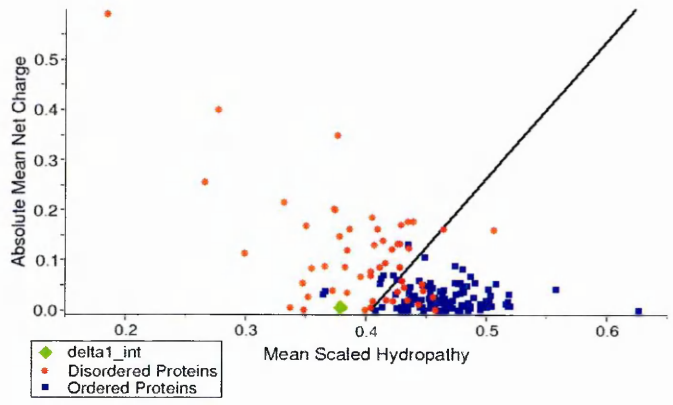


Jagged -2

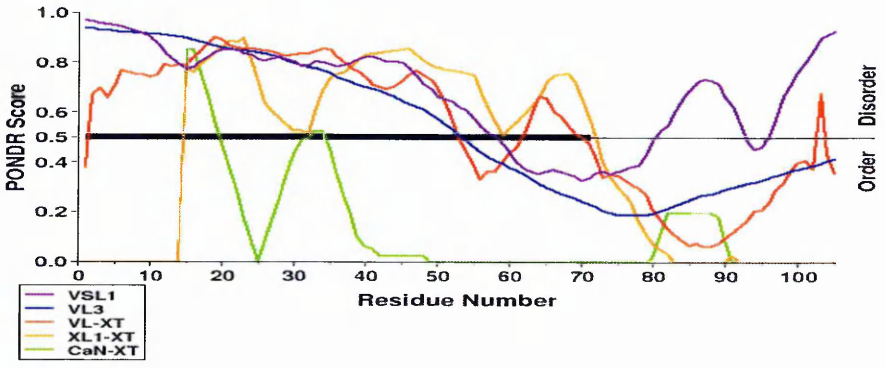


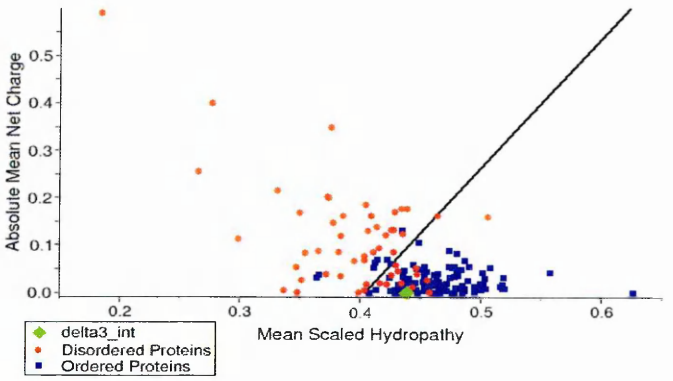
Delta -1


PONDR Score
1.0
0.8
0.6
0.5
0.4
0.2
0.0
0
25
50
75
100
125
150
Residue Number
Order
Disorder
VSL1
VL3
VL-XT
XL1-XT
CaN-XT


Absolute Mean Net Charge
0.5
0.4
0.3
0.2
0.1
0.0
0.2
0.3
0.4
0.5
0.6
Mean Scaled Hydropathy
delta1_int
Disordered Proteins
Ordered Proteins

Delta -3


PONDR Score
1.0
0.8
0.6
0.5
0.4
0.2
0.0
0
10
20
30
40
50
60
70
80
90
100
Residue Number
Order
Disorder
VSL1
VL3
VL-XT
XL1-XT
CaN-XT


Absolute Mean Net Charge
0.5
0.4
0.3
0.2
0.1
0.0
0.2
0.3
0.4
0.5
0.6
Mean Scaled Hydropathy
delta3_int
Disordered Proteins
Ordered Proteins

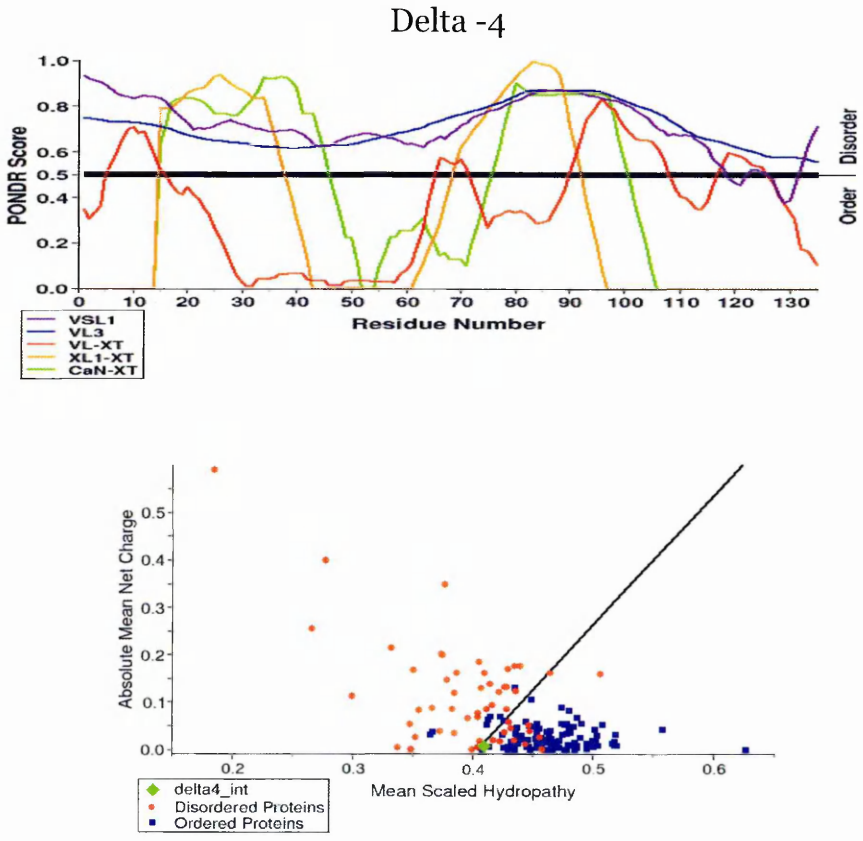


Figure 12. Protein disorder predicted by PONDR® (Predictor of Naturally Disordered Regions). PONDR® score is calculated using different predictors and plotted against the residue number. VSL1 combines two predictors optimized for long (>30 residues) and short (<=30 residues) disordered regions, respectively; VL3 is a neural network predictor trained on 152 long regions of disorder that were characterized by various methods and a set of ordered proteins consisting of 290 PDB-Select-25 chains having no disordered residues; VL-XT integrates three feedforward neural networks: VL1, the N-terminus predictor (XN), and the C-terminus predictor (XC); XL1 is a neural network predictor optimized to predict regions of disorder greater than 39 amino acids, and was trained on 7 disordered regions identified from missing electron density in X-ray structures; CaN is a neural network predictor that was trained on regions of 13 homologous *calcineurin* proteins. The charge-hydropathy plots compare the absolute, mean net charge (neglecting histidines) and the mean, scaled Kyte-Doolittle hydropathy. The dataset used in this plot include 105 completely ordered proteins, 54 completely disordered proteins, and 64 proteins with disordered regions.

SECONDARY STRUCTURE

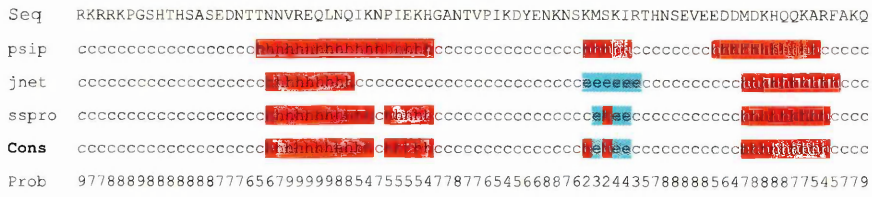
The intracellular regions of the protein families we studied do not seem to belong to any of the known domain types. Nevertheless, secondary structure prediction methods currently available can usually achieve high levels of accuracy that may allow one to note the consensus features of a family. So even if the intracellular region of a Notch ligand is predicted to be non-globular and intrinsically disordered, there is suficient motivation to perform secondary structure predictions. First, intrinsically disordered regions are known, in specific instances, to fold upon binding to their targets. Second, the interaction with the inner side of the membrane may in itself drive the formation of secondary structure elements. Secondary structure predictions can help in identifying stretches that show some intrinsic propensity to form secondary structure elements. These stretches may be the same that adopt a well defined structure upon binding to a protein target or through interaction with the membrane.

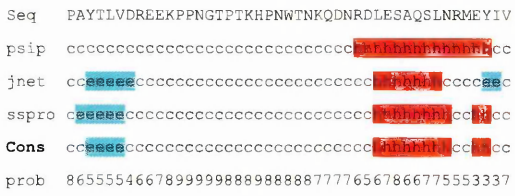
Secondary structure predictions based on different methods were found to be in a good overall agreement (Figure 13).

While Jagged-1 and -2 are characterized by three helices predicted respectively in the N-terminal region, in the central region and at the C-terminus with a relative high consensus score, predictions for Delta proteins display a different pattern.

Delta-1 and -4 are characterized by α-helix in the N-terminal region predicted with a moderate consensus score, and four segments of β-strands. Of these, two are in the central region and two at the C-terminus, the letters being predicted with a high consensus score. The pattern for Delta-3 is similar, but the consensus score is lower (Figure 13).

Jagged -1





Jagged -2

Delta -1

Delta-3

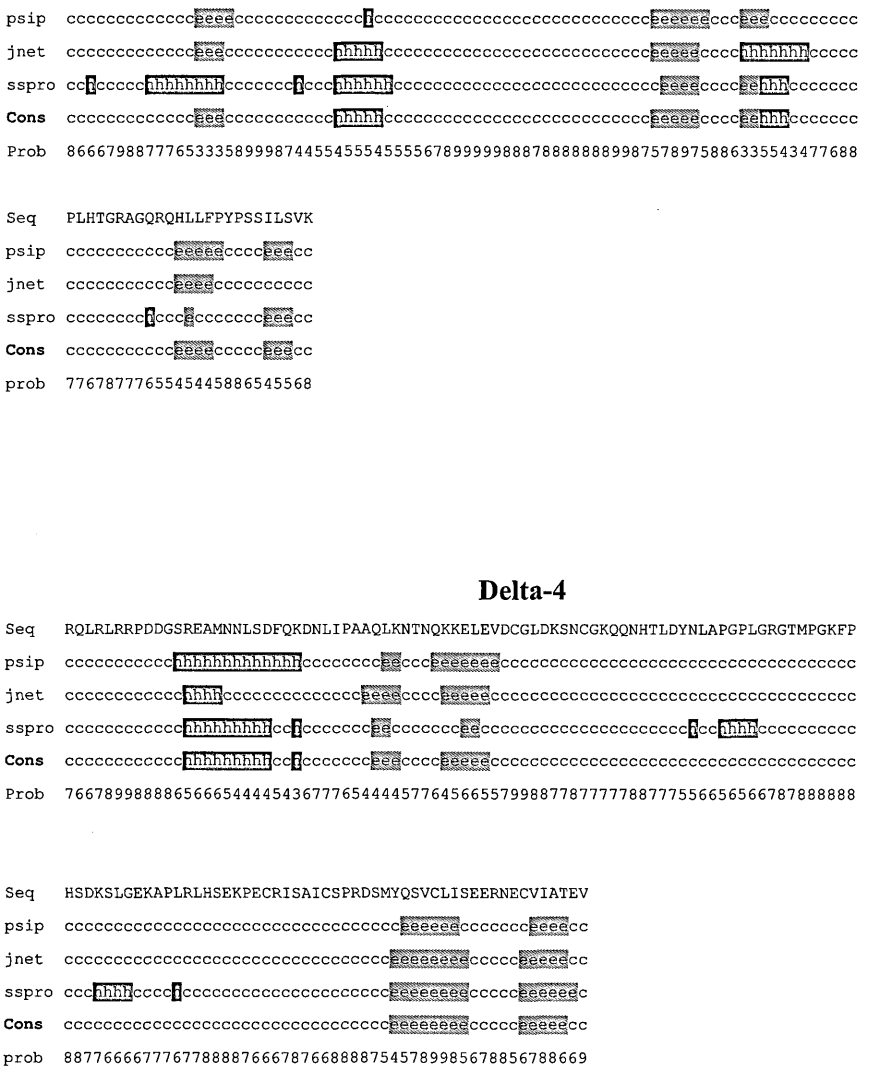


Figure 13. Secondary structure predictions for the intracellular region of human Jagged and Delta proteins. Predictions were run from the PHYRE server. Amino acid sequence, Psi-Pred, Jnet, SSpro, consensus predictions and probability score are shown. Helical segments are highlighted in red, β-strands in light blue.

PATTERN RECOGNITION

Pattern descriptions cover most aspects of post-translational modifications. The accuracy of current pattern-search methods is substantially increased as compared to the simple pattern representations used in earlier methods. Moreover, protein families are better targets of prediction since the conservation of the patterns detected can provide additional support to the prediction. Phosphorylation patterns are especially important in signal transduction so their conservation in a protein

family may indicate important biological functions. With this in mind we decided to analyze all the proteins under study with the pattern prediction servers recommended in the literature.

ELM

Pattern recognition by ELM (Appendix 4) was run assuming that the intra-cellular region of Notch ligands, although normally belonging to the plasma membrane environment, can also be localized in the cytoplasm and in the nucleus. Several potential binding sites for different domains were identified in ligands of both the Jagged and Delta family. Most of these are "signaling" domains, as classified by SMART (14-3-3, FHA, PDZ, SH2, SH3, WW). Additionally, a motif involved in endocytosis (TRG_ENDOCYTIC) and several phosphorylation sites were also identified. It is interesting to remark that, while a few motifs are shared by all ligands (for ex. LIG_PDZ_3 and LIG_WW_4), most of them are restricted to selected ligands. For example, the PDZ type I binding motif can be found in Delta-1 and Delta-4, but not in Delta-3 and in the Jagged-1 and -2 ligands. Potential binding sites for SH2 and SH3 domains display different specificities. Finally, the tyrosine-based endocytic signal (TRG_ENDOCYTIC_2) can be found in Jagged-1, Delta-1 and Delta-4, but not in Jagged-2 and Delta-3. Some of these recognition patterns (LIG_14-3-3, LIG_FHA_1, LIG_SH2, LIG_WW_4) require phosphorylation of specific Ser, Thr, or Tyr residues. It is therefore possible to combine the phosphorylation site predictions with binding motif recognition by ELM and phosphorylation sites predicted by ELM itself. In most cases, the interpretation is not straightforward, because there is no consensus between the different prediction methods. In a few cases, however, a consensus is reached and predictions are expected to be more reliable. In Jagged-1, for example, region 40-43 is predicted to contain a FHA domain binding motif, which would require

phosphorylation of T40. T40, on the other hand, is not a phosphorylation site according to either DisPhos, NetPhos, or ELM. On the contrary, S103 in Jagged-2 is predicted to be phosphorylated by all three methods, and is also predicted to be a binding site for WW type 4 domains (Figure 14).

Jagged-1	D	N	E	Jagged-2	D	N	E
LIG_14-3-3_3 HTHSAS 9-14 HSASED 11-16 [RHK][STALV].[ST].[PESRDIF]	N N	Y Y	Y Y				
LIG_FHA_1 THSA 10-13 TVPI 40-43 T..[ILA]	N N	N N	Y N				
LIG_SH2_GRB2 YENK 46-49 Y.N.	N	Y	N				
LIG_SH2_STAT5 YTLV 83-86 Y[VLTFIC]..	Y	Y	N				
LIG_WW_4 PNGTPT 93-98 ...[ST]P.	N	Y	Y	LIG_WW_4 KNFTPP 47-52 PGRSPG 100-105 ...[ST]P.	N Y	N Y	Y Y

Delta-1	D	N	E	Delta-3	D	N	E	Delta-4	D	N	E
				LIG_FHA_1 TGRA 84-87 T..[ILA]	N	N	N				
LIG_SH2_SRC YVIS 140-143 Y[QDEVAIL][DEPYHI][IPVGAHS]	N	Y	N	LIG_SH2_SRC YVIS 62-65 Y[QDEVAIL][DENPYHI][IPVGAHS]	N	Y	N				
LIG_SH_STAT5 YVS 140-143 YVLTFIC]..	N	Y	N	LIG_SH2_STAT5 YVIS 62-65 Y[VLTFIC]..	N	Y	N				
LIG_WW_4 EKGTPT 106-111 ...[ST]P.	N	N	Y	LIG_WW_4 LAGTPE 16-21 EVATPL 73-78 ...[ST]P.	N N	Y N	Y Y	LIG_WW_4 AICSPR 107-112 [ST]P.	N	Y	Y
MOD_TYR_ITIM VDYNLV 71-76 [ILV].(Y)..[ILV]	N	N									
MOD_TYR_ITSM KDTKYQSV 132-139 ..T.(Y)..[IV]	N	Y									

Figure 14. Combined predictions of binding motifs and phosphorylation sites. Binding motifs thatrequire Ser/Thr or Tyr phosphorylation are extracted from the ELM predictions. Potential phosphorylation sites are predicted by DisPhos (D), NetPhos (N), and ELM(E). Legend: **LIG_14-3-3_3**, 14-3-3 proteins interacting motif (Ser/Thr phosphorylation required); **LIG_FHA_1**, forkhead-associated domain interaction motif 1, (Thr phosphorylation required); **LIG_SH2_GRB2**, Src Homology 2 (SH2) domains interaction motif (tyrosine phosphorylation required); **LIG_SH2_STAT5**, STAT5 Src Homology 2 (SH2) domain binding motif (tyrosine phosphorylation

required); **LIG_WW_4**, class IV WW domains interaction motif (phosphorylation-dependent interaction); **LIG_SH3_2** class II SH3 domains binding motif; **MOD_TYR_ITIM**, immunoreceptor tyrosine-based inhibitory motif (tyrosine phosphorylation required); **MOD_TYR_ITSM**, immunoreceptor tyrosine-based switch motif (tyrosine phosphorylation required).

PHOSPHORYLATION SITES

Several potential phosphorylation sites are predicted by NetPhos in both Jagged and Delta ligands. The number of sites is however drastically reduced assuming that phosphorylation is occurring preferably in disordered regions. DISPHOS , NetPhos Yin-O-Yan and SignalP predictions could be found in Appendix 4.

The DISPHOS predictor is based on a set of over 2000 experimentally determined, non redundant phosphorylation sites, and assumes that phosphorylation occurs mainly in regions of intrinsic disorder, as predicted by PONDR®.

Table 2 is combine all serine, threonine, and tyrosine residues is in the ligands. **S**erines are in red, **T**hreonines in blue, t**Y**rosines in green.

	S	**T**	**Y**
Jagged-1	S8, S61		Y83
Jagged-2	S11, S18, S85, S103	T1	
Delta-1	S102, S103, S119, S126		
Delta-3	S8		
Delta-4	S14, S98		

Table 2. **S**erines are in red, **T**hreonines in blue, t**Y**rosines in green for all ligands.

Several potential phosphorylation sites are predicted by NetPhos in both Jagged and Delta ligands. The number of sites is however drastically reduced assuming that phosphorylation is occurring preferably in disordered regions (DisPhos) . Sites that are candidates both for Ser/Thr phosphorylation and for O-glycosylation by β-N-acetylglucosamine (Yin-Yang) can also be identified. Simple monosaccharide modification by β-N-acetylglucosamine of Ser and Thr hydroxyls is reversible and inducible, and thus fulfils the requirements for a signal transduction modification.

METAL BINDING POTENTIAL

Histidines and Cysteines, which are in their reduced form in the intracellular environment, are the usual ligands of structural Zn^{2+} ions in zinc proteins, including zinc fingers and several transcription factors. Although no specific pattern corresponding to known zinc binding motifs could be identified in the sequence of human Notch ligands (**Figure 15**), their amino acid composition is peculiar in respect to potential zinc binding capacities. Delta-1 and Delta-4 contain respectively a His_4Cys_5 and a His_3Cys_6 array of His and Cys residues, with a total of nine potentially zinc binding residues and a percentage of cysteine which is much higher then the average observed in human proteins (His = 2.64%; Cys = 2.31%). Jagged-1 and Delta-3, on the contrary, contain respectively a His_6 and a His_5 array of Histidines and no Cysteines, with a histidine content higher than what statistically expected. Also Jagged-2 contains a His_4Cys array, although in this case the composition is not significantly different from the average. Preliminary experimental results confirm indeed that recombinant proteins corresponding to the intracellular region of Delta-4 and Jagged-1 can bind to columns containing immobilized Ni^{2+} ions, and experiments are underway to confirm if they can bind Zn^{2+} ions (**Table 3**) .

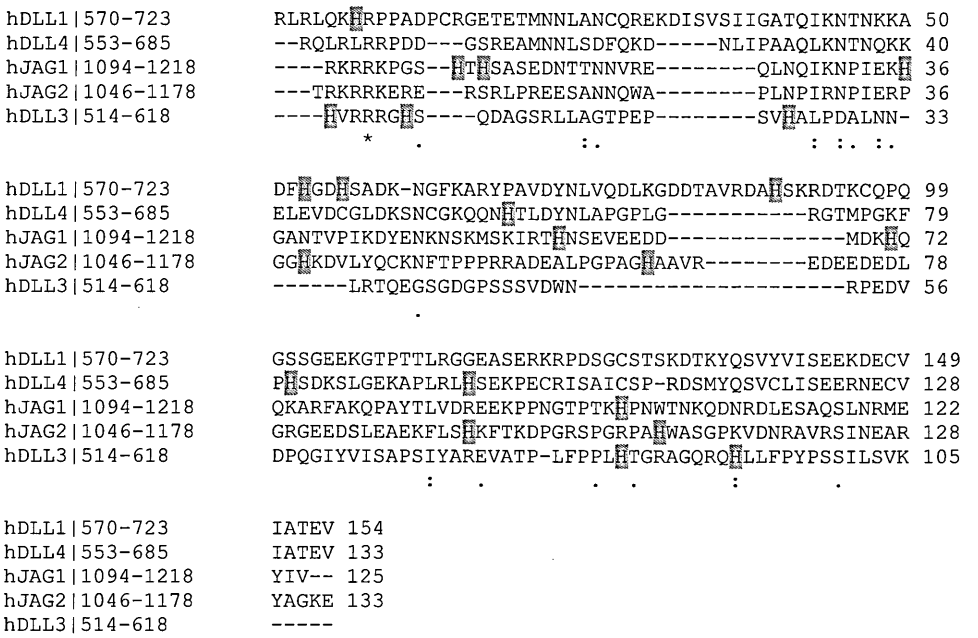


Figure 15. Human Notch ligands intracellular region. A ClustalW alignment of human Notch ligands cytoplasmic tail. Histidines are highlighted in light blue, Cysteines in yellow.

	His (%)	Cys (%)
hDLL1	4 (2.6)	5 (3.2)
hDLL4	3 (2.3)	6 (4.5)
hJAG1	6 (4.8)	0 (0.0)
hJAG2	4 (3.0)	1 (0.8)
hDLL3	5 (4.8)	0 (0.0)

Table 3. Histidine and cysteine content in human Notch ligand cytoplasmic tail. The number of His and Cys is shown; the percentage is given in parenthesis; values above the average are in green, below the average in red. The average values calculated for human proteins (His = 2.64%; Cys = 2.31%) can be found at www.pasteur.fr/~tekaia/aafreq.html and do not distinguish between intra- and extra-cellular proteins. No standard deviation is given.

DISCUSSION*

DIFFERENT TAILS FOR THE SAME DOG?

From the sequence analysis of the intracellular region of Jagged and Delta proteins, two features emerge.

The first is a relatively evident clustering of Notch ligands in distinct groups, when ligands are compared basing upon the sequence of their intracellular region only. These groups include orthologues of human Jagged-1 (group J1), of human Jagged-2 (group J2), of human Delta-1 (D1) and Delta-4 (D4). Two additional, more heterogenous groups include orthologues of human Delta-3 (D3) and other more distantly related ligands (DX). It is remarkable that *Drosophila* Serrate, which is usually considered to be the orthologue of human Jagged, rather belongs to the group including mammalian Delta-3 proteins, as well as *Drosophila* Delta. Given the recent experimental reports on the importance of the intracellular region of Notch ligands in bidirectional signaling, we propose that the sequence of the intracellular region can provide an effective ground for the classification of Notch ligands. This new classification has several advantages: (i) the intracellular region is relatively short (100-150 residues) compared to the full length ligand (600-1000 residues) as well as to the extracellular region; sequence alignments are thus easier and phylogenetic analysis more sensitive; (ii) the extra-cellular region, with its relatively well conserved architecture, is likely to provide the structural scaffold required for binding to the receptor, but might be rather tolerant to changes in regions that are not directly implicated in receptor binding (for example in the multiple tandem EGF repeats); changes in these regions would mask differences that are functionally more relevant; (iii) the intracellular region couples Notch ligands both to receptor

binding-dependent and receptor binding-independent signaling networks, through post-translational dynamical modifications of the cytoplasmic tail and networks of protein-protein interactions; it is thus expected to be most informative about evolutionary conservation or differentiation of function.

* Summary - figures of consensus predictions are in Appendix 2

The second feature emerging from the systematic sequence analysis is the striking conservation throughout species of the intracellular region for each selected ligand. The sequence conservation is not limited to the C-terminus, which is known to interact with PDZ containing proteins through a short well defined tetrapeptide motif, but extends well beyond the C-terminal residues. On the other hand, structural predictions supported by preliminary experimental results (see below) point towards a mainly disordered nature for Notch ligands cytoplasmic tail. Intrinsic disorder suggests that the cytoplasmic tail might act as a flexible linker between the inner face of the plasma membrane and the C-terminal protein interacting motif. If the role as a linker is true, one might expect a relative high variability in the amino acid sequence. This variability, on the contrary, is not observed. Because of the importance of Notch signaling in tissue patterning and morphogenesis, it is plausible to speculate that the intracellular region of Notch ligands is kept under a strong selective pressure because subtle changes in its amino acid sequence can have drastic consequences on the coupling of the Notch transduction pathway to different networks of protein-protein interactions. Precise sequence characteristics might be required for specific patterns of post-translational modifications to take place, for specific protein-protein interactions to occur, and possibly for the modulation of the conformational properties at the interface with the membrane environment. Unfortunately, very little is known at present about these events, and additional experimental work is needed.

NO STRUCTURE, NO FUNCTION?

Predictions on the intracellular region of human Jagged and Delta ligands are consistently pointing to a lack of globularity, thus assigning these regions to the group of "natively unfolded" or "intrinsically unstructured" proteins (**Figure 15**). Indeed, experimental results obtained in our laboratory confirm that the recombinant proteins corresponding to the intracellular region of human Jagged-1 and Delta-4, expressed in *E. coli* and purified, are mainly disordered in solution. Although in the past decades structural biology has been dominated by the dogma that "structure determines function", recent evidence is suggesting that this might not always be true. The availability of entire genome sequences, more sophisticated prediction tools, and experimental evidence show that intrinsically unstructured proteins and disordered regions in proteins are quite common, and should be considered as a rule, rather then as an exception (Dunker et al., 2002; Dunker et al., 2001; Dyson and Wright, 2005; Romero et al., 2004; Tompa, 2002; Tompa, 2005; Tompa et al., 2005; Uversky, 2002a; Uversky, 2002b).

In the four eukaryotic genomes surveyed, more than 30% of sequences are predicted to have disordered regions longer then 50 residues and, in *Drosophila,* a staggering 17% of proteins are predicted to be wholly disordered. IUPs and regions of disorder are more frequently found in proteins involved in signaling networks (Iakoucheva et al., 2002).

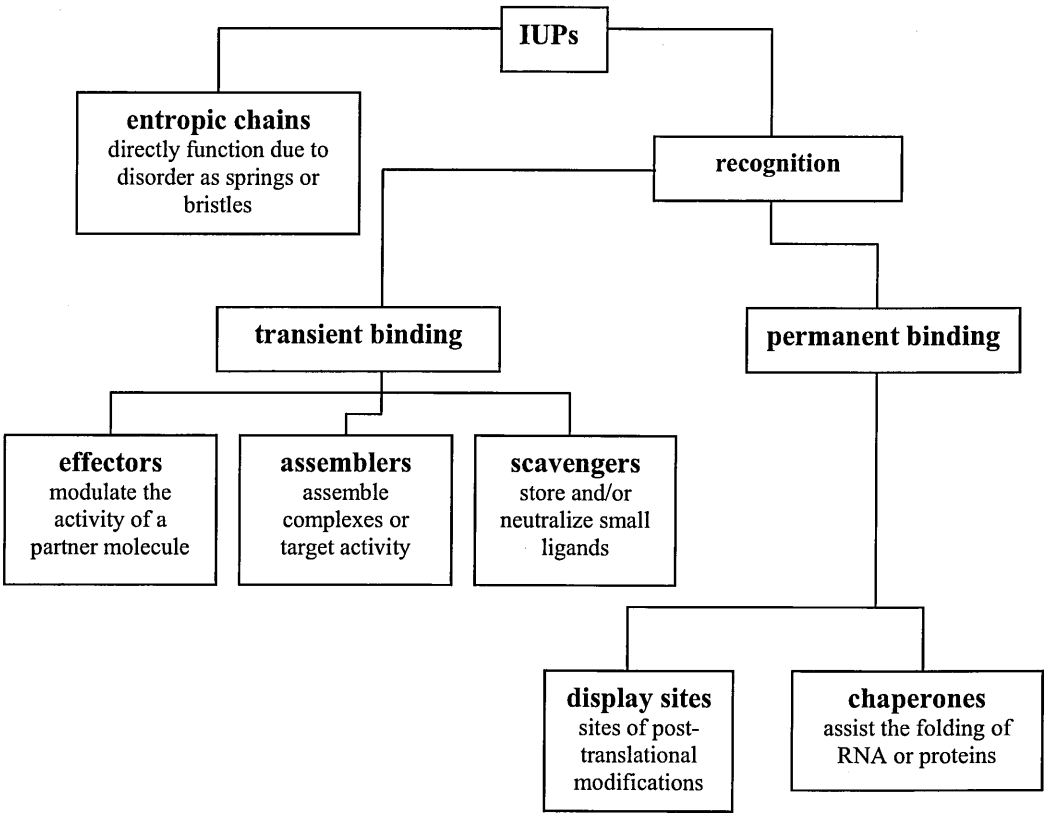


Figure 15. Functional classification scheme of IUPs. The function of IUPs stems either directly from their capacity to fluctuate freely in a large conformational space (entropic chain functions) or the ability to transiently or permanently bind partner molecule(s).

IUPs are usually characterized by a high number of charged residues compared to the number of hydrophobic residues, which results in the lack of a hydrophobic core, little or no secondary structure elements, high hydrodynamic radius, and often a high net charge at physiological pH, calculations for Delta and Jagged proteins are shown in **Table 4**. From the biophysical point of view, IUPs can be considered as polypeptide chains that in physiological conditions are sampling a much wider conformational space with respect to globular proteins. It has been proposed that this extended sampling can indeed have several advantages. IUPs have a much larger interaction surface/volume ratio compared to globular proteins, which allows for the accommodation of a relatively high number of docking sites on a relatively short polypeptide chain, at the same time reducing the protein volume, therefore the

molecular crowding. The extended conformational sampling has interesting thermodynamic consequences. It enables IUPs to couple folding to binding maintaining high specificity and low affinity due to the balance between the enthalpic contribution to binding and the opposite entropic effect. Indeed, weak although specific interactions are most important in molecular recognition. It has also been proposed that the high capture radius of IUPs is the ground for the so called "fly-casting" mechanism, whereby the unfolded polypeptide binds weakly at relatively long distances and then folds as it 'reels in' its target. The fly-casting mechanism predicts an increased rate of binding, which may well be important when the cellular concentrations of a regulatory protein and its target are low, as is the case for many signaling and transcriptional processes.

	DLL1	DLL3	DLL4	JAG1	JAG2	*globular*
Order Promoting AA: I,L,W,V,F, Y,C	C 5 3.2% I 6 3.9% L 6 3.9% F 2 1.3% W 0 0.0% Y 4 2.6% V 9 5.8% **20.7%**	C 0 0.0% I 4 3.8% L 10 9.5% F 2 1.9% W 1 1.0% Y 3 2.9% V 7 6.7% **25.8%**	C 6 4.4% I 5 3.7% L 14 10.4% F 2 1.5% W 0 0.0% Y 2 1.5% V 5 3.7% **25.2%**	C 0 0.0% I 5 4.0% L 4 3.2% F 1 0.8% W 1 0.8% Y 3 2.4% V 5 4.0% **15.2%**	C 1 0.8% I 3 2.3% L 7 5.3% F 3 2.3% W 2 1.5% Y 2 1.5% V 4 3.0% **16.7%**	*C 1.6%* *I 5.4%* *L 8.4%* *F 4.0%* *W 1.6%* *Y 3.6%* *V 7.0%* *31.6%*
Disorder Promoting AA: E,K,R,G,Q,S, P,A	A 11 7.1% R 11 7.1% Q 7 4.5% E 11 7.1% G 11 7.1% K 15 9.7% P 7 4.5% S 12 7.7% **54.8%**	A 8 7.6% R 9 8.6% Q 5 4.8% E 4 3.8% G 9 8.6% K 1 1.0% P 12 11.4% S 12 11.4% **57.2%**	A 8 5.9% R 10 7.4% Q 7 5.2% E 10 7.4% G 8 5.9% K 10 7.4% P 9 6.7% S 11 8.1% **54.0%**	A 6 4.8% R 9 7.2% Q 7 5.6% E 11 8.8% G 3 2.4% K 14 11.2% P 8 6.4% S 8 6.4% **52.8%**	A 13 9.8% R 18 13.5% Q 2 1.5% E 16 12.0% G 10 7.5% K 9 6.8% P 14 10.5% S 7 5.3% **66.9%**	*A 8.2%* *R 4.6%* *Q 3.7%* *E 6.0%* *G 8.0%* *K 6.1%* *P 4.6%* *S 6.3%* *47.5%*
Net charge	+1	0	+1	+4	+3	
Mean hydrophob.	0.371	0.497	0.416	0.313	0.362	
AA	155	105	135	125	133	

Table 4. Amino acid composition. Order-promoting amino acids (W, C, F, I, Y, V, L and N), disorder-promoting amino acids (A, R, G, Q, S, P, E and K) and content (%) in the cytoplasmic region of human Delta and Jagged proteins. In the last column (globular, in italics), the amino acid composition of globular proteins is also shown for comparison.

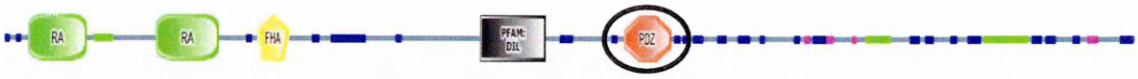
On the other hand, secondary structure predictions are consistently pointing to the presence of several stretches of α-helix and β-strand in the cytoplasmic tail of human Notch ligands. These results can appear at odds with the lack of globularity predicted by other tools, but might reflect the propensity of there regions to adopt a defined secondary structure in well determined circumstances. These circumstances are still unknown, but might be represented by the binding to the target protein, by some yet unidentified post-translational modification, or by the peculiar environment represented by the interface between the plasma membrane and the cytoplasm. These conditions are currently under investigation in our laboratory.

DOES THE TAIL MAKE THE DIFFERENCE?

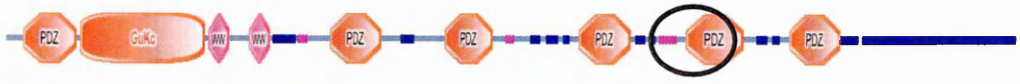
Despite the high number of potential binding sites present on the intracellular regions of human Jagged and Delta ligands, there are relatively few experimental reports on the identification of interacting proteins *in vitro* and *in vivo* (**Figure 16**). Peptide affinity chromatography with a 14 residue peptide corresponding to the C-terminus of human Delta-1 lead to the identification of Dlg1 as a binding partner from HeLa cells extracts. In living cells, Dlg1 was found to bind to Delta-1 and -4, but not to Jagged-1 (Six et al., 2004). In a similar study, a 27 residue peptide corresponding to the C-terminal region of human Delta-1 was used to identify binding partners from either mouse brain lysate or human neuroblastoma cells (Wright et al., 2004). This study leads to the identification of MAGI (membrane associated inactive guanylate kinase) proteins as binding partners for Delta-1. The interaction was confirmed by *in vitro* and *in vivo* experiments. The entire intracellular region of mouse Delta-1 was used in GST pull-down experiments *in vitro* and in a mammalian two-hybrid system *in vivo* to identify a member of the MAGI family, activin receptor interacting protein 1, as a binding partner (Pfister et al., 2003). Finally, AF6 was identified as a binding partner for Jagged-1 intracellular

region, as determined by GST pull-down experiments (Ascano et al., 2003). In these experiments, transiently transfected 293T cells expressing AF6 were lysated and cell lysates incubated with the intracellular region of Jagged-1 fused to glutathione S-transferase (GST) and immobilized on glutathione-agarose beads. Detection was carried out by Western blot analysis using appropriate antibodies. No interaction could be detected using a Jagged-1 contruct lacking the PDZ recognition motif (five C-terminal amino acids) or an AF6 construct lacking the PDZ domain, showing that this interaction id PDZ-mediated.

Dlg1 (the human homologue of *Drosophila* Discs Large protein), MAGI (membrane associated guanylate kinase) proteins and AF6 (afadin) are membrane associated proteins found at cell junctions and involved in the organization of the cytoskeleton.

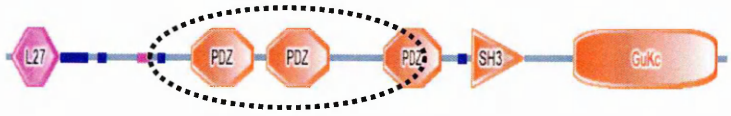


AFAD_HUMAN (AF6/afadin)



MAGI1_HUMAN (membrane associated guanylate kinase inverted 1)

AIP1_MOUSE (activin receptor interacting protein 1)



DLG1_HUMAN (*Drosophila* discs large)

Figure 16. Domain architecture of PDZ-containing proteins interacting with Jagged/Delta. RA, Ras association domain; FHA, forkhead associated domain; PDZ, domain present in PSD-95, Dlg, and

ZO-1/2; GuKc, guanylate kinase homologue; WW, domain with two conserved Trp; L27, domain present in receptor targeting proteins Lin-2 and Lin-7; SH3, src homology 3 domain. The target PDZ domain, where identified, is enclosed in a circle. Regions of intrinsic disorder are in blue, low complexity regions in magenta, coiled coils in green.

These proteins all contain PDZ domains. As anticipated by predictions, Delta-1 and -4, but not Delta-3, contain a C-terminal PDZ Class I binding motif. Jagged-1, but not Jagged-2, contains a C-terminal PDZ Class II binding motif. These studies, while confirming experimentally a link between the Notch signaling network and scaffolding proteins involved in cell remodeling, also raise several questions. The C-terminal tetrapeptide of Delta-1, -4, and Jagged-1 is required for recognition by the target PDZ domain. Although this motif is necessary, it is still under debate if it is also sufficient. In human cells, there are over 400 proteins containing at least one PDZ domain and several of these proteins contain more than one PDZ domain. It is thus difficult to envisage a situation where a specific interaction is occurring only through the C-terminal tetrapeptide. It can be speculated that two possible, non mutually exclusive alternatives exist. In the first hypothesis, the interaction surface with the PDZ domain extends to regions upstream of the C-terminus. In other words, whereas the gross energy of binding would come from the interaction between the tetrapeptide and the PDZ domain, the specificity of the interaction might reside in a larger region. Alternatively, a multiple lock-key mechanism might be effective to achieve the wanted specificity. For example, Dlg1 contains three PDZ and one SH3 domains. While both Delta-1 and -4 are predicted to contain PDZ binding motifs, only Delta-1 is also predicted to possess a potential SH3 binding site. The multiple lock-key mechanisms, in other words the possibility of accommodating several binding sites on the same flexible polypeptide chain, together with the modular architecture of globular proteins would provide a very simple mean to achieve specificity at the molecular level. It is also conceivable, however, that time is controlling the multiple lock-key mechanism. The different binding motifs would be

used at different times during the cell cycle, in order to recruit specific proteins. In this case, binding motifs would be switched on and off by specific post-translational modifications like phosphorylation. Indeed, several potential phosphorylation sites have been identified on the cytoplasmic tail of Notch ligands. Unfortunately, very little is known about the dynamics of phosphorylation from experimental studies. Furthermore, potential Yin-Yang can also be predicted. These are serine or threonine residues that are candidates for both phosphorylation and glycosylation through the attachment of either a phosphate or a β-N-acetylglucosamine moiety (O-GlcNAc) to the hydroxyl oxygen of the amino acid. O-GlcNAc modification is reversible, inducible by specific signals, and may therefore be involved in signal transduction mechanisms. It has been found in several cytoplasmic and nuclear proteins, among them the estrogen receptor β, the C-terminal domain of RNA polymerase II, and c-Myc. The current hypothesis is that Yin-Yang sites would have access to three states (phosphorylation on/phosphorylation off/glycosylation) rather then only two (phosphorylation on/phosphorylation off) in signaling networks.

To address all these issues, in our laboratory we have expressed, purified, and immobilized on a matrix recombinant proteins corresponding to the full length cytoplasmic regions of human Jagged and Delta, and we are using these baits to identify binding partners and possibly post-translational modifications.

WHEN THE DOG LOSES ITS TAIL.

Recent studies have shown that not only Notch receptors, but also Notch ligands undergo a two-step proteolytic processing, the first cleavage occurring on the external side of the membrane, the second occurring at a yet unidentified site within the trans-membrane region. The result is the release of the cytoplasmic tail of the Notch ligand into the cytoplasm, and, according to some reports, a partial

localization also in the nucleus. The role of this proteolytic processing is not clear yet. It is possible that the cytoplasmic tail released from the membrane is simply acting as a cargo for the proteins docked to it. The partial localization in the nucleus however suggests an additional role. While the intracellular regions of human Jagged-1 and Delta-3 are rich in histidines, the same regions in Delta-1 and -4 are particularly rich in cysteines. These variations might be random, but it is interesting to remark that histidines and cysteines are the physiological ligands for zinc ions in zinc binding proteins, including several transcription factors. It is thus tempting to speculate that zinc ions might bind to the cytoplasmic tail of Notch ligands, mediating homo- or hetero-dimerization of the ligand itself, and perhaps playing a role in determining the conformation of the intra-cellular region. Although the histidine and cysteine motifs found do not correspond to any known zinc-binding pattern, in our laboratory we are investigating the effects of zinc ions on the conformational properties of the recombinant forms of human Jagged-1 and Delta-4 cytoplasmic tails.

APPENDIX 1

BLAST and CLUSTAL description

BLAST

The BLAST program is perhaps the most frequently used scientific software today. It is designed to compare biological sequences in terms of an alignment score, and is perhaps the most essential tool for the comparison of protein sequences. BLAST's heart is a heuristic algorithm, optimized for very fast sequence similarity searches. The BLAST algorithm is based on the observation that related sequences share regions of high similarity characterized by a relatively high density of aligned residues. Like other programs of sequence comparison, BLAST uses scoring matrices, such as the PAM or BLOSUM matrices for proteins. The version of the BLAST program optimized for detecting protein sequence similarities is called BLASTP. The program first locates the regions of high similarity (called HSPs, high scoring segments) by finding matching words of n residues (n is usually 3 for proteins and 11 for DNA), then splicing these into contiguous segments using a heuristic rule. In the next step, the contiguous segments are elongated into both directions, continuing for as long as the score (composed of the respective elements of the scoring matrix) increases. Once this elongation process is finished, the next step is to determine the statistical significance of the resulting HSP. The statistics of BLAST scores is based on a simple idea: it is possible to model distribution of similarity scores that randomly occur between sequences. Using this distribution, it is possible to express, using the tools of statistics, the probability that a certain score S appears by chance. A high score, such as those occurring between evolutionarily related proteins, will have a very small probability to occur by chance.

The distribution of local sequence alignment scores follows the so-called extreme value (or Gumbell) distribution. In the limit of sufficiently large sequence lengths m and n, the statistics of HSP scores are characterized by two parameters, K and lambda. Most simply, the expected number of HSPs with score at least S is given by the formula

$$E = Kmnc^{-\lambda S}$$

where m and n are the length of the two sequences compared. The parameters K and *lambda* can be thought of simply as natural scales for the search space (size $m \times n$) and the scoring system, respectively. If S is high, it is expected to occur extremely rarely by pure chance, so E is a very low number for biologically significant similarities. It has to be mentioned, that E values are database dependent, while the S scores are not. E is also related to the probability of S occurring by chance. This probability P can be mathematically expressed since the number of HSPs with score >=S is described by a Poisson distribution. It can be shown that the probability to find at least one HSP with a score at least equal to S will be

$$P = 1 - e^{-E}$$

This is the P-value associated with the score S. For very low values (e.g. E<0.01), P values and E values are nearly identical.

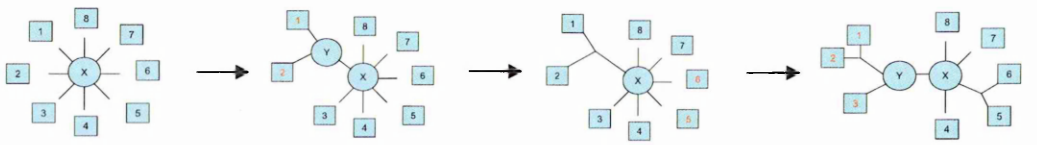
The use of this simple statistics can be extended in two important ways: 1) it can be applied not only for the comparison of two sequences, but also for the comparison of a sequence with a database. 2) Even though it was originally deduced for ungapped alignments, it was shown by computational experiments as well as by analytical results that it can be used for the practically important case of gapped alignments.

The practical interpretation of *E* (or *P*) values is not straightforward. While it is considered certain that very low E values, such as E <.0001 are biologically important (e.g. they occur only between evolutionarily related sequences), some biologically significant similarities have much higher *E* values. This is because the known protein universe is characterized by protein groups vastly different in the number of members, in average protein size, in the similarity within the group, etc. For this reason, there are no "universal threshold values" above which *E* values would surely correspond to biologically important similarities.

CLUSTAL

The principle of dynamic programming used for pair-wise alignments can be extended to multiple alignments as well. Since the task is very time consuming, practical applications use heuristic approaches, which are hierarchical and progressive in nature. CLUSTAL is a family of programs developed by Des Higgins to perform multiple alignments of biological sequences. CLUSTALV was developed by Higgins and Sharp in 1988. CLUSTALW (1994) is a significant improvement. It uses a three-step algorithm that starts with a pair-wise alignment of all sequence pairs in order to determine sequence similarity. Then an order of addition of sequences to alignments is determined based on pair-wise similarity, using a hierarchical approach, based on a neighbor-joining tree-building algorithm (NJ) which provides a "guide tree". Finally, a multiple alignment based on the order defined by the guide tree, in which the most similar pairs of sequences are assembled into pair-wise alignments, and then new sequences are added and/or alignments are combined in a progressive manner.

The key step of CLUSTAL is to determine the order in which sequences/partial alignments are to be joined. This joining hierarchy can be best pictured as a tree, and CLUSTAL uses a fast, distance-based algorithm to build a guide tree-hierarchy. First versions of CLUSTAL used the so-called UPGMA algorithm, but recent version use neighbor-joining (NJ) method (Saitou and Nei, 1987). Conceptually (see sketch below), the NJ algorithm starts with a star-like tree in which there is central node (root, denoted by "x" in the sketch) and all sequences are the leaves. Then the closest sequences are combined in a hierarchical, greedy fashion so as to yield a final, binary tree.



The first computational step of the NJ algorithm (and of CLUSTAL) is i) the determination of the pair-wise similarities/distances between the objects (sequences). Then ii) the distance matrix is modified so that the separation between each pair of nodes is adjusted based on their average divergence from all other nodes. Subsequently iii) the nearest pair of nodes in this modified matrix is linked and replaced by their common ancestor ("pruning"). Steps ii) and iii) (matrix modification and neighbor joining, respectively) are repeated until two nodes remain, separated by a single branch. NJ is widely used for generating evolutionary trees because it is fast and thus suited for large datasets and for generating a large number of trees. It permits lineages with largely different branch lengths and correction for multiple substitutions. However, as distances are used instead of sequences information is reduced. Another disadvantage is that it gives only one

possible tree which is strongly dependent on the model of evolution used. Nevertheless, CLUSTAL is known to give high quality multiple alignments for most practical applications, so we decided to use this algorithm.

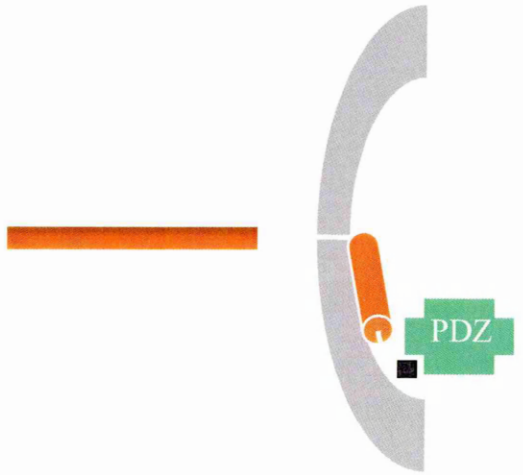
The qualitative interpretation of phylogenetic trees is based on the intuitive expectation that the branch length of a leaf (sequence) should be proportional to the number of substitutions after the last differentiation event. This is only an approximation which is fulfilled only in the sense that adjacent pairs of more similar sequences are usually separated by shorter branches whereas more distant (divergent) sequence-pairs have longer branches. As the method of our choice, the neighbor-joining algorithm, uses an additive tree, it can assign a negative length to the branch. In this case, the interpretation of branch lengths as an estimated number of substitutions gets into difficulties. This problem can be corrected without changing the overall topology of the tree.

Rooting of trees is a separate problem. Distance methods such as NJ can construct a root, but since this question is not relevant to our analysis, we did not use rooted trees.

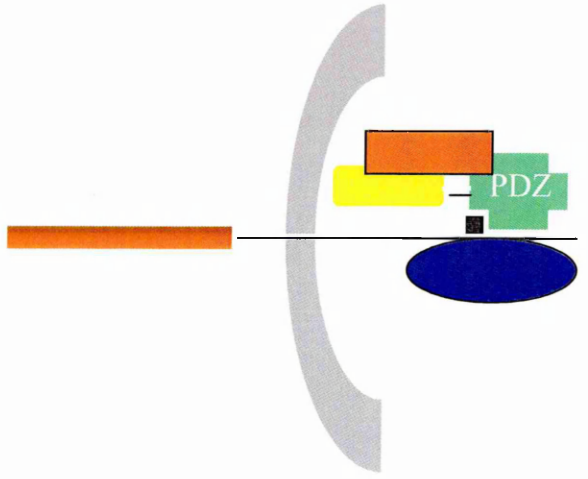
APPENDIX 2

Summary of consensus predictions

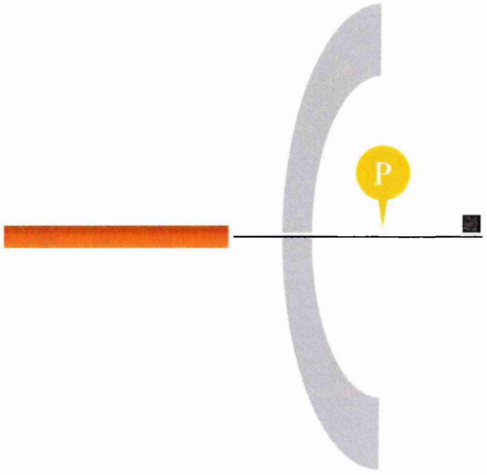
In general, the intracellular regions of the analyzed proteins seem to be disordered (our lab experiment confirm it), have relatively few phosphorylation sites, but seem to have PDZ-binding motifs predicted on their C-termini, the latter is in good agreement with experiment (**Table** 5).



If these predictions are real, then all 5 proteins could play a role of a flexible linker and/or scaffold to support many proteins in their protein-protein interaction fulfilling their function in the cell.



There are enough evidences that all tails could be glycosylated or phosphorylated, thus way the proteins could be switched "on" or "off" by phosphorylation, it can actually turn a nonpolar hydrophobic protein into a polar and extremely hydrophilic molecule. Receiving signal for the outside part of the ligand, the protein could becomes phosphorylated or dephosphorylated and again to begin or stops working. This is the mechanism in many forms of signal transduction and transmembrane ligands.

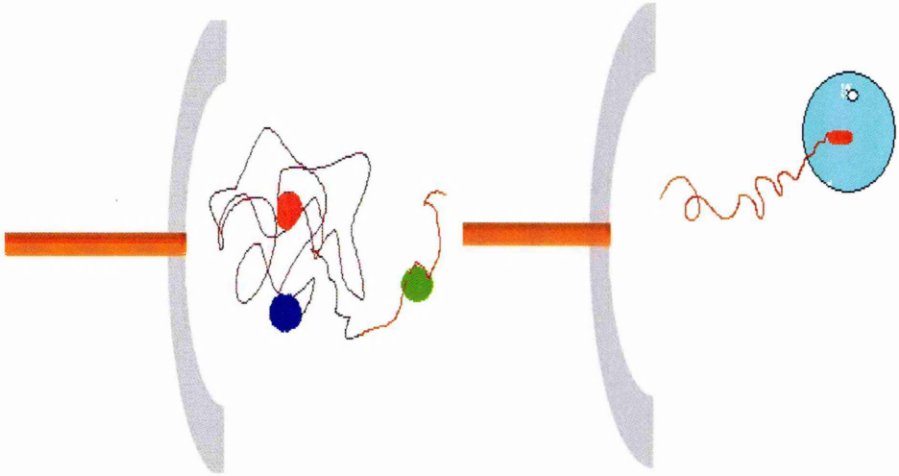


Even though the amino acid composition (hydrophobic/hydrophilic balance) is grossly reminiscent of signal peptides, there are no signal-like sequences predicted (**Table 5**).

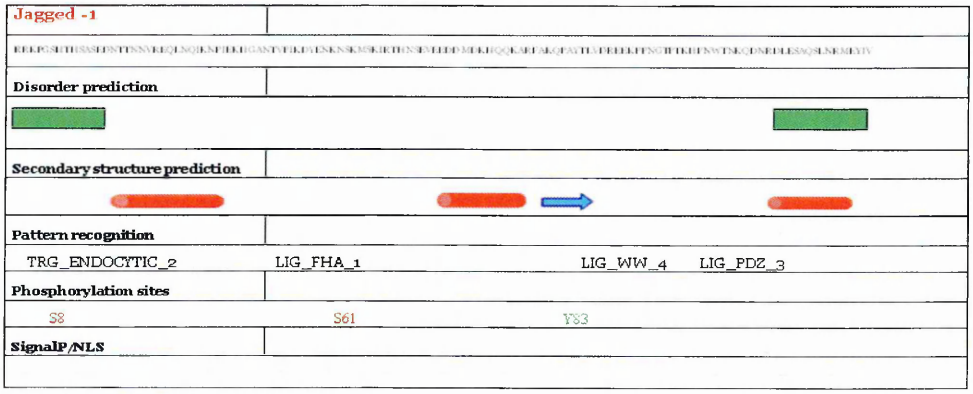
Even though there are predictions of secondary structure prediction. This structure could be form as the ligand is link with the membrane and more reliable if it is cleaved form the membrane. After releasing the intercellular part in the cell, the protein trying to protect itself and prepares for its function:

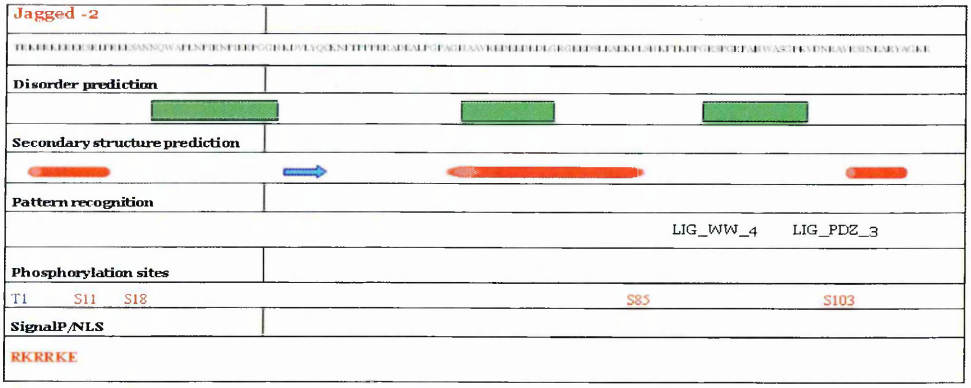
- forms structure with is only chain or around other protein(s)
- Could bind to different proteins and have function in the cytoplasm,

- Could be transported in the nucleus, if his target is the DNA or specific processes in the nucleus.

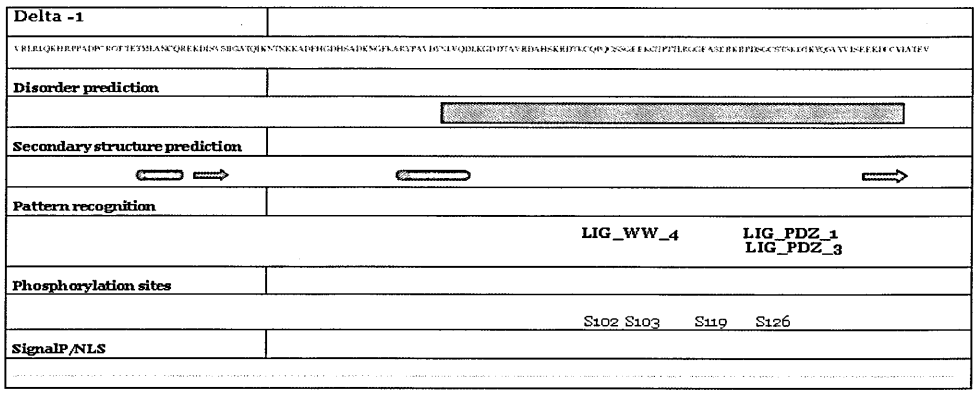


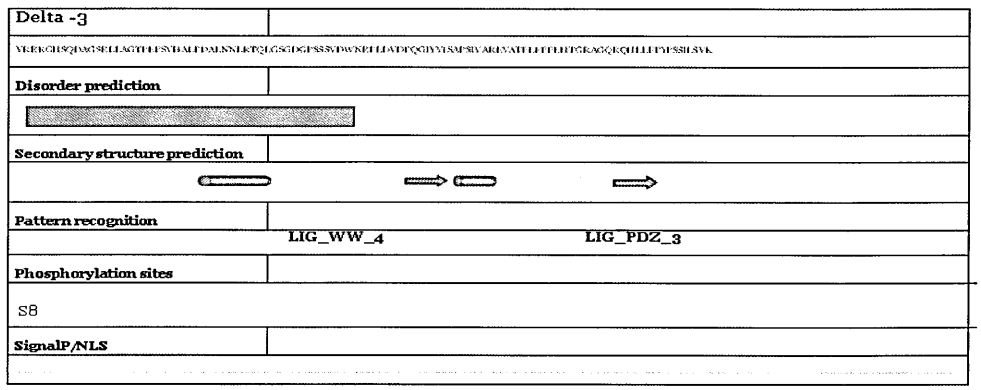
Jagged -1 and -2





Delta -1, -3 and -4





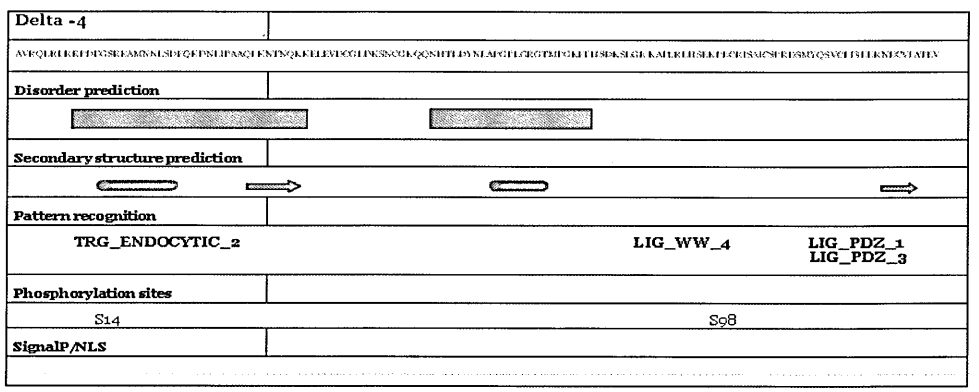


Table 5. Summary of consensus predictions

APPENDIX 3

Collection of Notch ligands. Proteins were identified from a combination of BLAST similarity searches using the intracellular region as query, domain database searches using the DSL and MNNL domains, and genome sequence databases (ENSEMBLE). Only ligands containing the full length putative cytoplasmic region are reported. Proteins are named using the Swiss-Prot/trEMBL entry name when available, the NCBI accession number, or the ENSEMBLE name.

Mammals					
Homo sapiens (human)	JAG1_HUMAN	JAG2_HUMAN	DLL1_HUMAN	DLL3_HUMAN	DLL4_HUMAN
Macaca fascicularis / Macaca mulatta (macaque)	ENSMMUP00000023809	ENSMMUG00000001276	ENSMMUP00000027839	ENSMMUP00000025105	ENSMMUP00000026826 ENSMMUP00000019092
Pan troglodytes (chimp)	ENSPTRP00000022729	ENSPTRP00000011587	ENSPTRP00000032142		ENSPTRP00000044880
Mus musculus (mouse)	JAG1_MOUSE	JAG2_MOUSE	DLL1_MOUSE	DLL3_MOUSE	DLL4_MOUSE
Rattus norvegicus (rat)	JAG1_RAT	JAG2_RAT	DLL1_RAT	DLL3_RAT	XM_230472.3
Bos taurus (cow)	ENSBTAP00000017029 [NO IC]	XP_595574.2	XP_877844.1	ENSBTAP00000013852	ENSBTAP00000013680
Canis familiaris (dog)	XP_858823.1	XP_548004.2	ENSCAFP00000006075		XM_852991
Monodelphis domestica (opossum)	ENSMODP00000006044	ENSMODP00000018384	ENSMODP00000006944	ENSMODP00000017099	ENSMODP00000000234
Birds					
Gallus gallus (chicken)	Q90819_CHICK	O42347_CHICK	Q90656_CHICK		ENSGALP00000013851
Amphibians					
Xenopus laevis/ Xenopus tropicalis(frog)	Q90YD2_XENLA P79941	ENSXETG00000006790	ENSXETP00000048762 Q91902_XENLA		ENSXETP00000046649 ENSXETP00000006289
Cynops pyrrhogaster(newt)			Q8AW87_CYNPY		

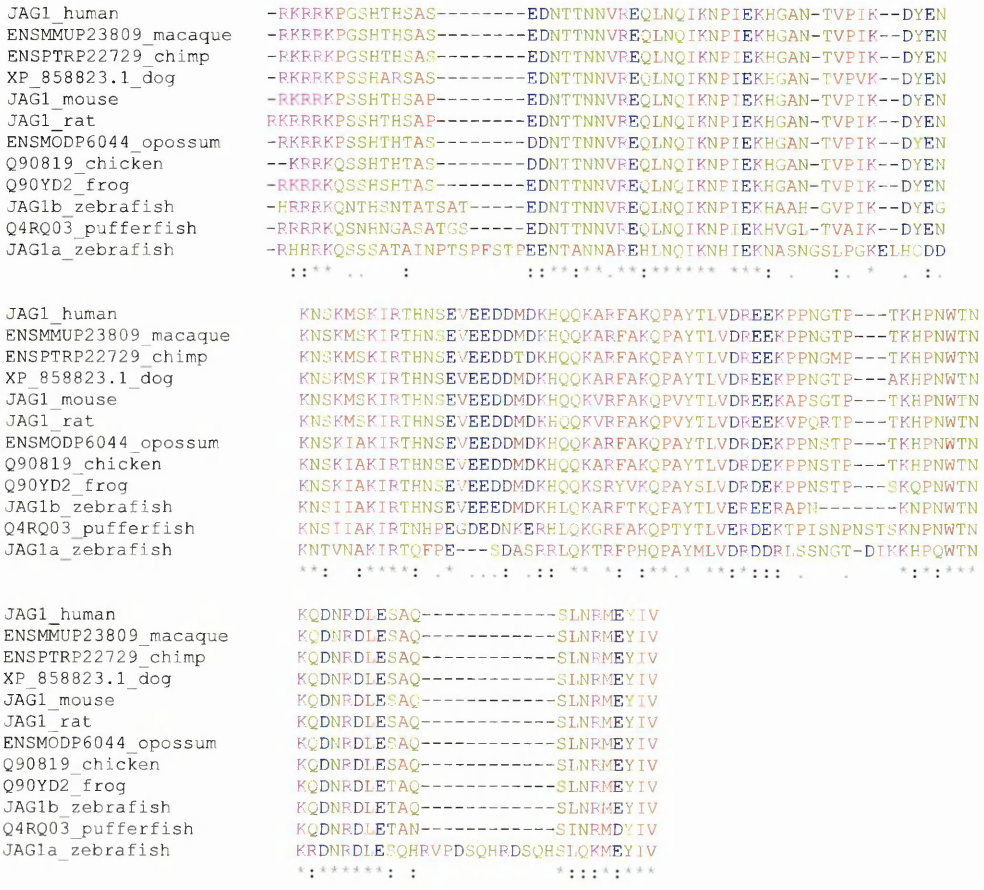
Fishes					
Brachydanio rerio (zebrafish)	JAG1A_BRARE JAG1B_BRARE	Q90Y55_BRARE	DLLA_BRARE DLLD_BRARE	DLLB_BRARE DLLC_BRARE	Q5RGG6_BRARE
Tetraodon nigroviridis (pufferfish)	Q4RQ03_TETNG	GSTENT00023297001	Q4SZZ8_TETNG Q4T963_TETNG	Q4RLS7_TETNG	Q4SC13_TETNG Q4RLS7_TETNG GSTENT00020707001
Fugu rubripes (fugu)	NEWSINFRUP00000173504 [NO IC]	NEWSINFRUP00000149799	NEWSINFRUP00000158918		NEWSINFRUP00000135910
Insects					
Drosophila Melanogaster (fruitfly)	SERR_DROME		DL_DROME		
Apis mellifera (bee)	XP_394560		XP_393831		
Echinodermata					
Lytechinus variegatus (greenurchin)			Q8T4P0_LYTVA		
Chordata					
Ciona savingny/ Ciona intestinalis (seavase)	ENSCINP00000008182 [NO IC]	ENSCING00000003969 [NO IC]	Q95YG0_CIOSA	ENSCINP00000008682	
Halocynthia roretzi (sea squirt)			Q8MP01_HALRO		
Metazoa					
Caenorhabditis elegans (worm)	LAG2_CAEEL		APX1_CAEEL		

APPENDIX 4

Multiple sequence alignment

Multiple sequence alignment of the Notch ligands intracellular region corresponding to group J1. Residues are colored as follows: A,V,F,P,M,I,L,W (hydrophobic) in red; D,E (acidic) in blue; R,K (basic) in magenta; S,T,Y,H,C,N,G,Q in green.

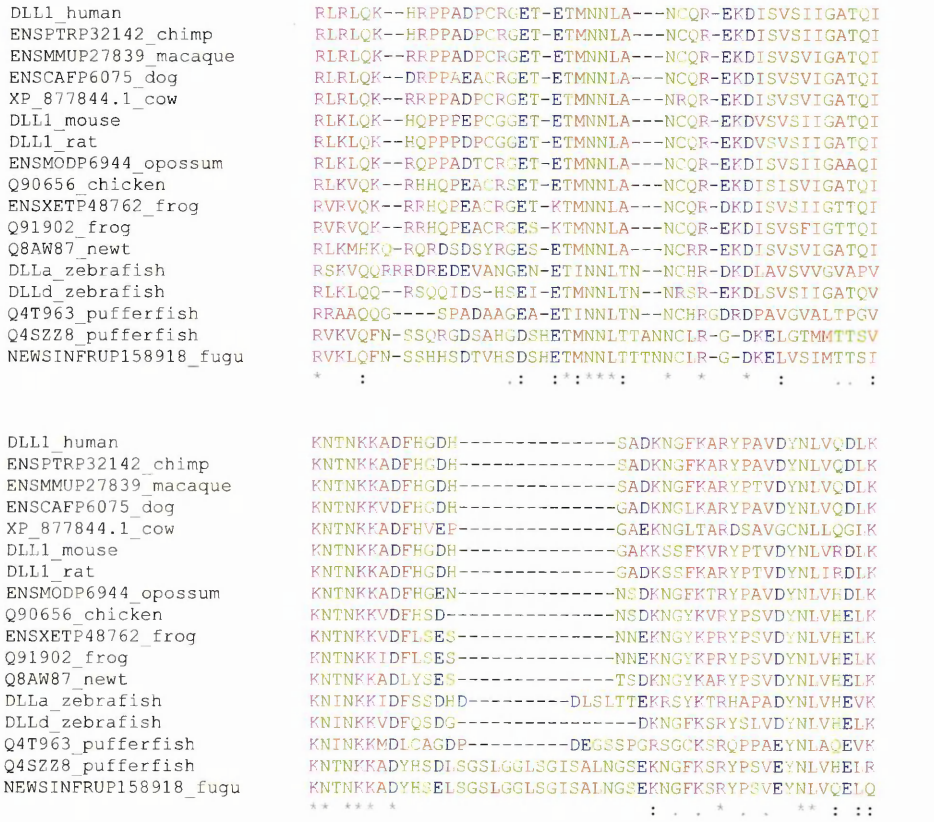
Jagged -1

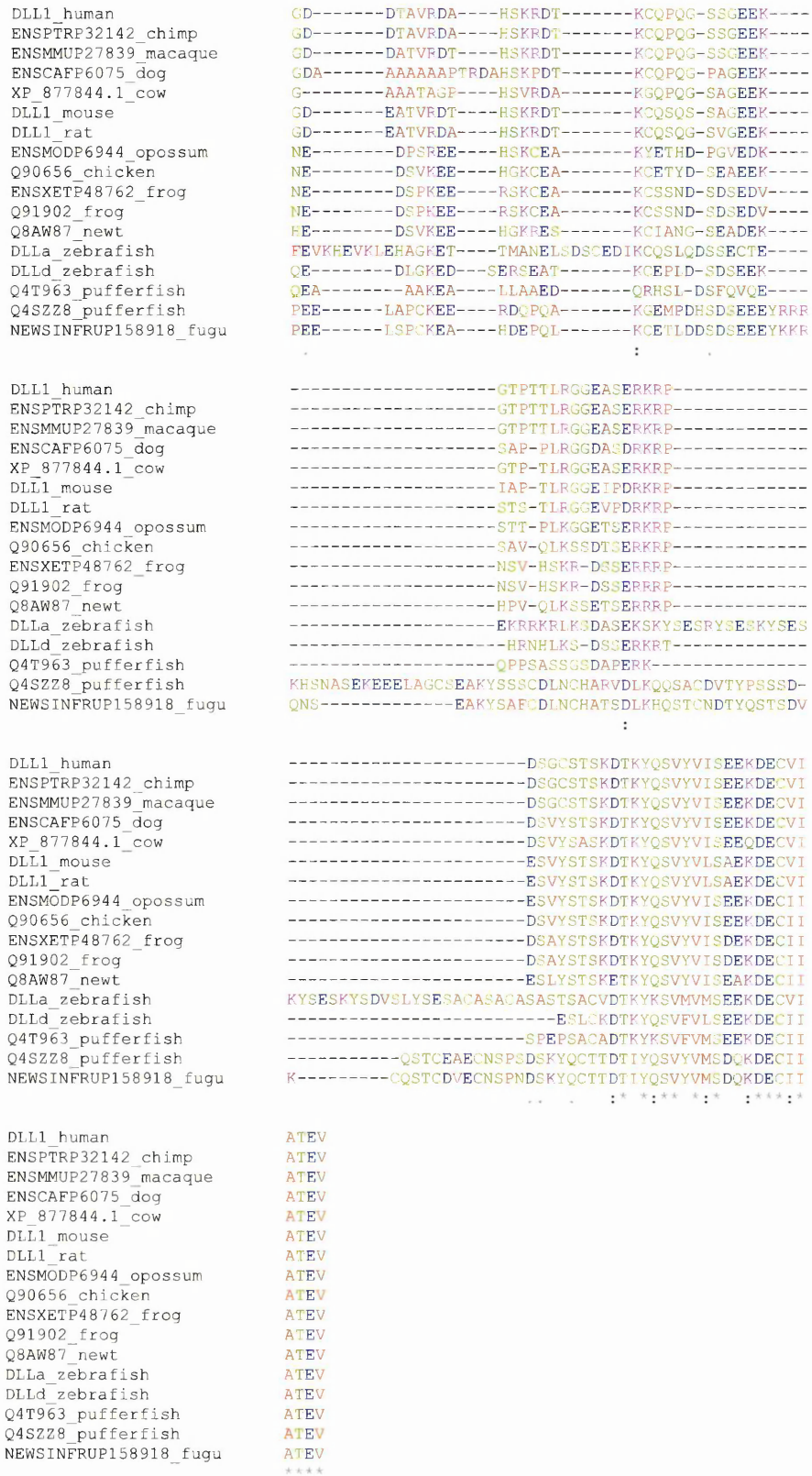


Jagged -2

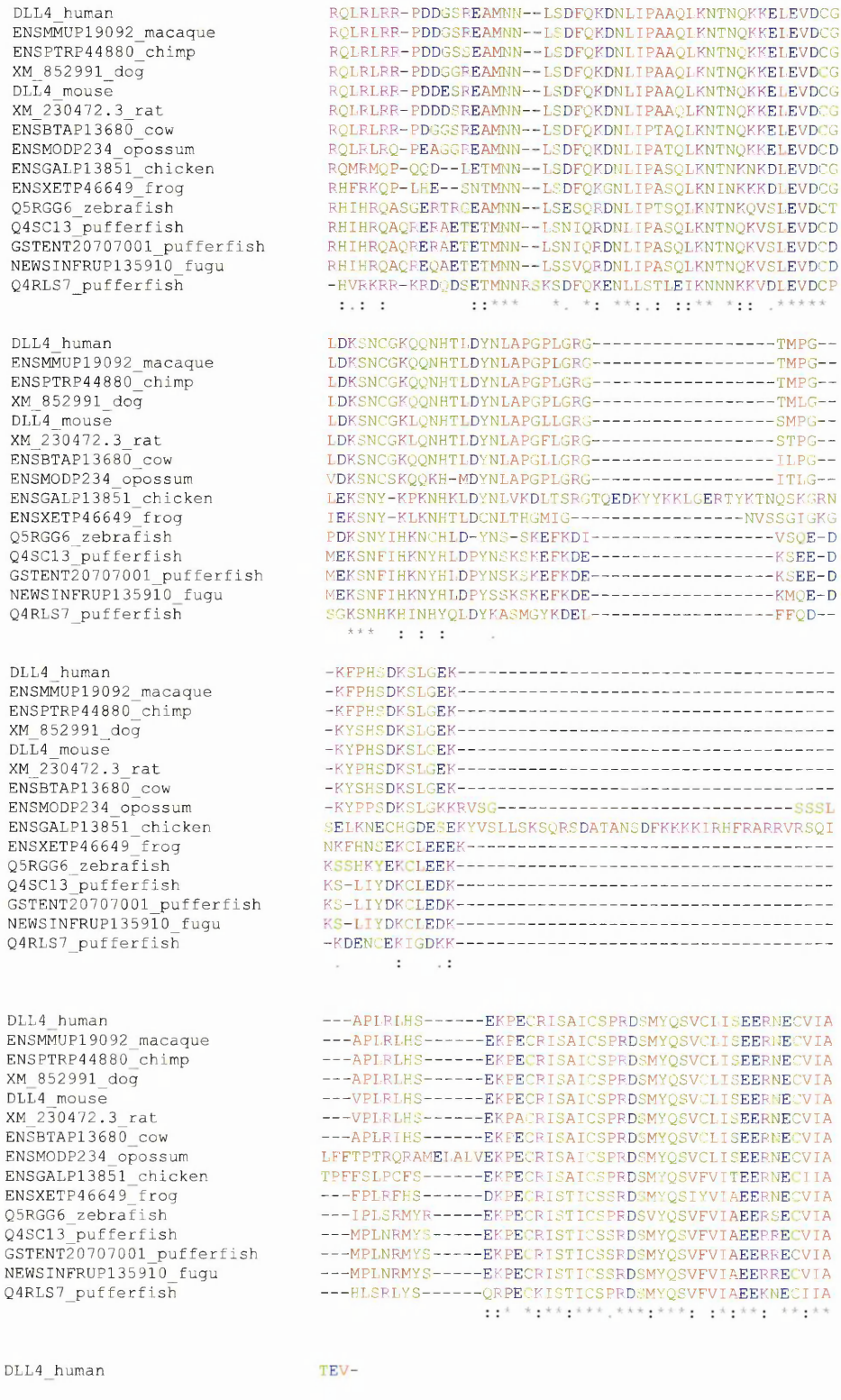

JAG2_human YQCKNFTPPPRRADEALPGPAGHAAVREDEEDE---------DLGRGEED
ENSMMUG1276_macaque YQCKNFTPPPRRADEALPGPAGHAGVREDEEDE---------DLGRGEED
ENSPTRP11587_chimp YQCKNFTPPPRRADEALPGPAGHAAVREDEEDE---------DLGRGEED
XP_548004.2_dog YPCKNFTPPPRRVGEALPGPAGRGEGGEEEEEE---------EPGRGEGG
JAG2_mouse YQCKNFTPPPRRAGEALPGPAGHGAGGEDEEDE---------ELSRGDGD
XP_595574.2_cow FPCKNFTPPPRRVGEALPGPAG---AGEDEEDE---------EPGRGEDE
JAG2_rat YQCKNFTPPPRRAGEALPGPASHGAGGEDEEDE---------ELSRGDGR
ENSMODP18384_opossum YECKNFISPPKRTHDEVEEYVERERFV----------IEVEKFL
O42347_chicken YECKNFISPQKRTCDAVEEYVEYEEEEDEEEER---------DEEMDKFL
ENSXETG6790_frog YECKNFLSHQKRTYDTEEE--EKEEEVGEEVT----------EVDKCL
Q90Y55_zebrafish FERTKLMGSPDRTCNSGDDEEDMEEDELELVEEWCGTEGGKHPVQKYSKS
NEWSINFRUP149799_fugu ----------QPYIK-
GSTENT23297001_pufferfish ----------HRDVQY
JAG2_human SLEAEKFLSHKFTKDPGRS-PGRPAHWASGPKVDNRAVRSINEARYAGKE-----
ENSMMUG1276_macaque SLEAEKFLSHKFTKDPGRS-PGRPAHWASGPKVDNRAVRSINEARYTGKE-----
ENSPTRP11587_chimp FLEAEKFLSHKFTKDPGRS-PGRPGHWASVPKVDNRAVRSINEARYAGKE-----
XP_548004.2_dog CLEAEKFLSHKFTKDPSGS-PGRPACWASGPKVDNRAVRSVNDSRHAGKE-----
JAG2_mouse SPEAEKFISHKFTKDPSCS-LGRPACWAPGPKVDNRAVRSTKDVRRAGRE-----
XP_595574.2_cow SPEAEKFLAHKFTKDPGRS-PGRPARWASGPKVDNRTLGGVSAARRAGRE-----
JAG2_rat LSRSREVPLTQIHQRPQLL-PGKASLLAPGPKVDNRAVRSTKDVRCAGRE-----
ENSMODP18384_opossum SQKLQKPPLTKGPGDLKESPLGKWAHRGASHKVDNRSLKNVNDY--EDGKD----
O42347_chicken SHKLTKPLPTKAS-DASESSPVKKSLQIG--KMDNRSVKNVNASNFEGSRD----
ENSXETG6790_frog SQTCPKTLTSKGDVDCSESSPVKKPHRTSDYKMDNRCVKNVNTS----------
Q90Y55_zebrafish AARTKNGLICTTRSTSGSSSPSLKAAYWTGFSPKDNNCKNVNNATAGQEHKEHCV
NEWSINFRUP149799_fugu ----------
GSTENT23297001_pufferfish ECR---KLVAAADRKCDGAEGVEEAE--PGEELEEDDERGMG----------

Delta -1


DLL1_human RLRLQK--HRPPADPCRGET-ETMNNLA---NCQR-EKDISVSIIGATQI
ENSPTRP32142_chimp RLRLQK--HRPPADPCRGET-ETMNNLA---NCQR-EKDISVSIIGATQI
ENSMMUP27839_macaque RLRLQK--RRPPADPCRGET-ETMNNLA---NCQR-EKDISVSVIGATQI
ENSCAFP6075_dog RLRLQK--DRPPAEACRGET-ETMNNLA---NCQR-EKDISVSVIGATQI
XP_877844.1_cow RLRLQK--RRPPADPCRGET-ETMNNLA---NRQR-EKDISVSVTGATQI
DLL1_mouse RLKLQK--HQPPPEPCGGET-ETMNNLA---NCQR-EKDVSVSIIGATQI
DLL1_rat RLKLQK--HQPPPDPCGGET-ETMNNLA---NCQR-EKDVSVSIIGATQI
ENSMODP6944_opossum RLKLQK--RQPPADTCRGET-ETMNNLA---NCQR-EKDISVSIIGAAQI
Q90656_chicken RLKVQK--RHHQPEACRSET-ETMNNLA---NCQR-EKDISISVIGATQI
ENSXETP48762_frog RVRVQK--RRHQPEACRGET-KTMNNLA---NCQR-DKDISVSIIGTTQI
Q91902_frog RVRVQK--RRHQPEACRGES-KTMNNLA---NCQR-EKDISVSFIGTTQI
Q8AW87_newt RLKMHKQ-RQRDSDSYRGES-ETMNNLA---NCRR-EKDISVSVIGATQI
DLLa_zebrafish RSKVQQRRRDREDEVANGEN-ETINNLTN--NCHR-DKDLAVSVVGVAPV
DLLd_zebrafish RLKLQQ--RSQQIDS-HSEI-ETMNNLTN--NRSR-EKDLSVSIIGATQV
Q4T963_pufferfish RRAAQQG----SPADAAGEA-ETINNLTN--NCHRGDRDPAVGVALTPGV
Q4SZZ8_pufferfish RVKVQFN-SSQRGDSAHGDSHETMNNLTTANNCLR-G-DKELGTMMTTSV
NEWSINFRUP158918_fugu RVKLQFN-SSHHSDTVHSDSHETMNNLTTTNNCLR-G-DKELVSIMTTSI
DLL1_human KNTNKKADFHGDH----------SADKNGFKARYPAVDYNLVQDLK
ENSPTRP32142_chimp KNTNKKADFHGDH----------SADKNGFKARYPAVDYNLVQDLK
ENSMMUP27839_macaque KNTNKKADFHGDH----------SADKNGFKARYPTVDYNLVQDLK
ENSCAFP6075_dog KNTNKKVDFHGDH----------GADKNGLKARYPAVDYNLVQDLK
XP_877844.1_cow KNTNKKADFHVEP----------GAEKNGLTARDSAVGCNLLQGIK
DLL1_mouse KNTNKKADFHGDH----------GAKKSSFKVRYPTVDYNLVRDLK
DLL1_rat KNTNKKADFHGDH----------GADKSSFKARYPTVDYNLIRDLK
ENSMODP6944_opossum KNTNKKADFHGEN----------NSDKNGFKTRYPAVDYNLVHDLK
Q90656_chicken KNTNKKVDFHSD----------NSDKNGYKVRYPSVDYNLVHELK
ENSXETP48762_frog KNTNKKVDFLSES----------NNEKNGYKPRYPSVDYNLVHELK
Q91902_frog KNTNKKIDFLSES----------NNEKNGYKPRYPSVDYNLVHELK
Q8AW87_newt KNTNKKADLYSES----------TSDKNGYKARYPSVDYNLVHELK
DLLa_zebrafish KNINKKIDFSSDHD---------DLSLTTEKRSYKTRHAPADYNLVHEVK
DLLd_zebrafish KNINKKVDFQSDG----------DKNGFKSRYSLVDYNLVHELK
Q4T963_pufferfish KNINKKMDLCAGDP---------DEGSSPGRSGCKSRQPPAEYNLAQEVK
Q4SZZ8_pufferfish KNTNKKADYHSDLSGSLGGLSGISALNGSEKNGFKSRYPSVEYNLVHELR
NEWSINFRUP158918_fugu KNTNKKADYHSELSGSLGGLSGISALNGSEKNGFKSRYPSVEYNLVQELQ


DLL1_human
ENSPTRP32142_chimp
ENSMMUP27839_macaque
ENSCAFP6075_dog
XP_877844.1_cow
DLL1_mouse
DLL1_rat
ENSMODP6944_opossum
Q90656_chicken
ENSXETP48762_frog
Q91902_frog
Q8AW87_newt
DLLa_zebrafish
DLLd_zebrafish
Q4T963_pufferfish
Q4SZZ8_pufferfish
NEWSINFRUP158918_fugu

Delta -4


DLL4_human
ENSMMUP19092_macaque
ENSPTRP44880_chimp
XM_852991_dog
DLL4_mouse
XM_230472.3_rat
ENSBTAP13680_cow
ENSMODP234_opossum
ENSGALP13851_chicken
ENSXETP46649_frog
Q5RGG6_zebrafish
Q4SC13_pufferfish
GSTENT20707001_pufferfish
NEWSINFRUP135910_fugu
Q4RLS7_pufferfish

```
ENSMMUP19092_macaque          TEV-
ENSPTRP44880_chimp            TEV-
XM_852991_dog                 ----
DLL4_mouse                    TEV-
XM_230472.3_rat               TEV-
ENSBTAP13680_cow              TEV-
ENSMODP234_opossum            TEV-
ENSGALP13851_chicken          TEV-
ENSXETP46649_frog             TEV-
Q5RGG6_zebrafish              TEV-
Q4SC13_pufferfish             TEV-
GSTENT20707001_pufferfish     TEV-
NEWSINFRUP135910_fugu         TEV-
Q4RLS7_pufferfish             TEVR
                              :.
```

Delta -3

```
DLL3_human                    --HVRRR-GHSQDAGSRLLAGTPEPSVHA----------
ENSMMUPP25105_macaque         --HVRRR-GHAQDAGSRLLAGTPEPSVHAL---------
DLL3_mouse                    --HVRRR-GPGQDTGTRLLSGTREPSVHTL---------
DLL3_rat                      --HVRRR-GPGRDTGTRLLSGTREPSVHTL---------
ENSBTAP13852_cow              --HVRRR-GPSRDTGPRLLAGTPEPSVHAL---------
ENSMODP17099_opossum          --RARRR-SPG--ARPLPPSADPAPPTPPP----------
SERR_fruitfly                 --RLAYRTSSGMNLTPSLDALRHE---EEK----------
XP_394560_bee                 -RTVRQRSSLTATTSSETSLHRHRSDLDEK----------
APX1_worm                     -HSFSKWKHPSSQQAGGSTILPTTTSIPMS----------
DLL_fruitfly                  -KRKRKRAQEKDDAEARKQNEQNAVATMHHNGSGVGVALASASLGGKTGS
XP_393831_bee                 -KRRQKREQAKADEEARLQNERNAVHSSMSKRGGGMGGGAGVGTGGSQGV
Q95YG0_seavase                RNSRKAVKSSSETSESPMESVQTWDAGQSA----------
ENSCINP8682_seavase            --RTNRNRSTKPDTSTPTDTTPTTQEVDTP----------
```

```
DLL3_human                    ------------------------------PDALNN---LR--TQE-G
ENSMMUPP25105_macaque         ------------------------------PDALNN---LR--TQE-G
DLL3_mouse                    ------------------------------PDALNN---LR--LQD-G
DLL3_rat                      ------------------------------PDALNN---LR--LQD-G
ENSBTAP13852_cow              ------------------------------PDALNN---MR--TQE-G
ENSMODP17099_opossum          ------------------------------ADALNN---LR--AHERG
SERR_fruitfly                 ----------------------------SNNLQNEENLRRYTNPLK
XP_394560_bee                 ----------------------------SNNLQNEENLRRYANPLK
APX1_worm                     ------------------------------------------TTSSG
DLL_fruitfly                  ----------------------NSGLTFDGGNPNIIKNTWDKSVNNICAS
XP_393831_bee                 GIVGNVGLLGGGGSGMISAGGGGSVCTLGTGDAHMIKNTWTANKSVNNVA
Q95YG0_seavase                ----------------------------ADVSKPSVVKAEVSIDAE
ENSCINP8682_seavase           ------------------------------PTLATK---VEVNFTD-E
```

```
DLL3_human                    SGDGPSSSVDWNR------------------------PEDVDPQGIYVIS
ENSMMUPP25105_macaque         PGDVPSSSVDWNR------------------------PEDVDSRGIYVIS
DLL3_mouse                    AGDGPSSSADWNH------------------------PEDGDSRSIYVIP
DLL3_rat                      AGDGPTSSADWNH------------------------PEDGDSRSIYVIP
ENSBTAP13852_cow              PGDGPSPSSDWN-------------------------PEDGDARSIYVIS
ENSMODP17099_opossum          PGHLKAPKHERTQRLL---------------------EPQLGRSPISFSIR
SERR_fruitfly                 GSTSSLRAATGMELSLNP-------------------APELAASAASSSAL
XP_394560_bee                 EQDQGEPRVSVVR------------------------PLSGTSLGALGAT
APX1_worm                     TGSPVYKVCIIDS------------------------EHRGNAPGSSSDS
DLL_fruitfly                  AAAAAAAAAAADECLMYGGYVASVADNNNANSDFCVAPLQRAKSQKQLNT
XP_393831_bee                 SARQDDLDSSFQTDVTLDSSCSG--------YKPEPVLADGRTRTTKQLN
Q95YG0_seavase                ATKLLDPEPIQTR------------------------VALPCAKCPCSHT
ENSCINP8682_seavase           VNHLLEPEARVEL------------------------PCSNCSHHKNQTT
```

```
DLL3_human                    APS--IYAR-EVATPLPPPLHTGRAGQRQH--LLFPYPSS--ILSVK---
ENSMMUPP25105_macaque         APS--IYAR-EVATLLSPPLHTGHTGQRQN--LLFPYPSS--ILSVK---
DLL3_mouse                    APS--IYAR-EDWLIQV----------------LF---------------
DLL3_rat                      APS--IYAR-EA-------------------------------------
ENSBTAP13852_cow              APS--VYAR-EVVNPPLPTLRTLGTMDRGC--LLFPFPAS--ILPFS---
ENSMODP17099_opossum          ADD--WCLP-DDSDPRT----------------IFIIPDS--SLYGRE--
SERR_fruitfly                 HRSQPLFPPCDFERELDSSTGLKQAHKRSSQILLHKTQNSDMRKNTVGSL
XP_394560_bee                 EES--LEMVSDESRHRLPPLYKAPSAEARNNTASFTYEEGPHKPYSKPRL
```

```
APX1_worm                EPD--HHCPPPHRHSPPPAYSS----------LVLYKKVPMAADDESSF
DLL_fruitfly             DPTLMHRGSPAGSSAKGASGGGPGAAEGKRISVLGEGSYCSQRWPSLAAA
XP_393831_bee            TEAAAHRASHLFQKEKDCLGLGLGIGVGVGVIESAKRSSVFAGNATTDSC
Q95YG0_seavase           VVTVEMAKVENHQVDKGP--------------CPTYEEACETSPCL---
ENSCINP8682_seavase      STT-KMGDPPTHEGARCP----------------TYEEACEDSPCLP--

DLL3_human               --------------------------------------------------
ENSMMUPP25105_macaque    --------------------------------------------------
DLL3_mouse               --------------------------------------------------
DLL3_rat                 --------------------------------------------------
ENSBTAP13852_cow         --------------------------------------------------
ENSMODP17099_opossum     --------------------------------------------------
SERR_fruitfly            DSPRKDFGKRSINCKSMPPSSGDEGSDVLATTVMV---------------
XP_394560_bee            QEP--TYSQQASSSQTSGP------HQVLTVHV-----------------
APX1_worm                RV-------------------------------------------------
DLL_fruitfly             GVAGACSSQLMAAASVAGSGAGTAQQQRSVVCGTPHM-------------
XP_393831_bee            CAAEAALLKRPTNITEGGSGPPGSGGGGGGETGCGVYVIDDHYRHDTSLA
Q95YG0_seavase           --------------------------------------------------
ENSCINP8682_seavase      --------------------------------------------------

DLL3_human               -------
ENSMMUPP25105_macaque    -------
DLL3_mouse               -------
DLL3_rat                 -------
ENSBTAP13852_cow         -------
ENSMODP17099_opossum     -------
SERR_fruitfly            -------
XP_394560_bee            -------
APX1_worm                -------
DLL_fruitfly             -------
XP_393831_bee            ATLATEV
Q95YG0_seavase           -------
ENSCINP8682_seavase      -------
```

Cellular Localization

PredictNLS

This program allows only one letter symbols for amino acid residues. If the reported motif can be traced to an Experimental NLS, the experimental NLS will is reported. If the reported NLS cannot be traced to any experimental NLS the prediction accuracy can be assessed by the number of nuclear proteins, in which this motif is found. All motifs in the NLS database are found in 3 or more families. All NLSs found in the query sequence are highlighted in red in the output report. The DNA binding NLS used to predict DNA binding is reported.

The prediction accuracy is estimated from the fraction of proteins which bind DNA. The probability of the NLS being found within the DNA binding domain is estimated.

There are no results for Jagged -1, Delta -1,-3,-4.

Jagged -2

Input Sequence (NLS's in Red)	T**RKRRKE**RERSRLPREESANNQWAPLNPIRNPIERPGGHKDVLYQCKNFTPPPRRADEAL PGPAGHAAVREDEEDEDLGRGEEDSLEAEKFLSHKFTKDPGRSPGRPAHWASGPKVDNRA VRSINEARYAGKE
Sequence Length	*133*
NLS's found. *No gives position of Motif*	• *RKRRKE 1*

Statistical data for Nuclear Localization Signals present in the Input Sequence

Generalized NLS (notation)	Type	No with NLS	%Nuc Proteins	%Non Nuc Proteins	Protein Swiss Id	Protein Localizations (Swiss anno.)
R[KR]{3,4}K[DE]	Potential	30	100	0	h2b astru	nuc
					creb bovin	nuc
					creb chlvr	nuc
					crem canfa	nuc
					bbf2 drome	nuc
					sus drome	nuc
					atf1 human	nuc
					atf6 human	nuc
					crea human	nuc
					creb human	nuc
					if16 human	nuc
					zep2 human	nuc
					atf1 mouse	nuc
					crea mouse	nuc
					creb mouse	nuc
					crem mouse	nuc
					h2b margl	nuc
					h2b3 psami	nuc
					h2b4 psami	nuc
					h2b patgr	nuc

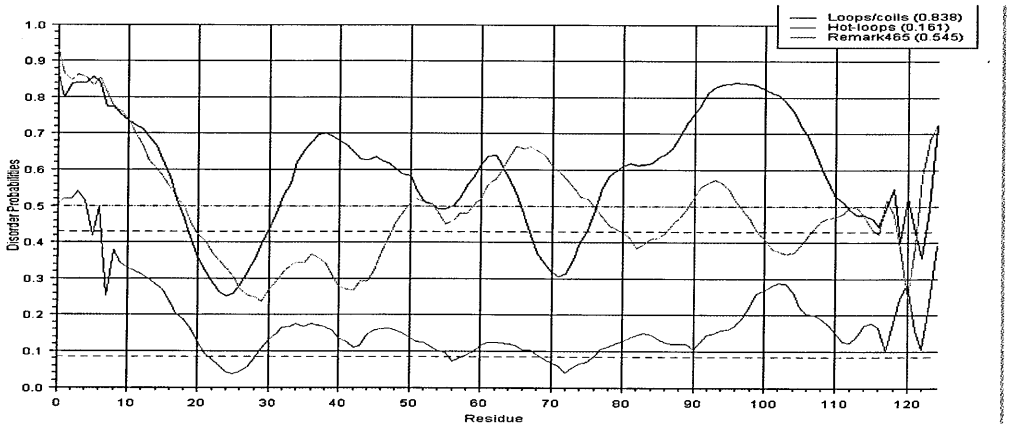
					creb rat	nuc
					crem rat	nuc
					h2b sipnu	nuc
					h2b1 strpu	nuc
					h2bn strpu	nuc
					h2bo strpu	nuc
					hmgh strpu	nuc
					orc5 yeast	nuc
					sko1 yeast	nuc

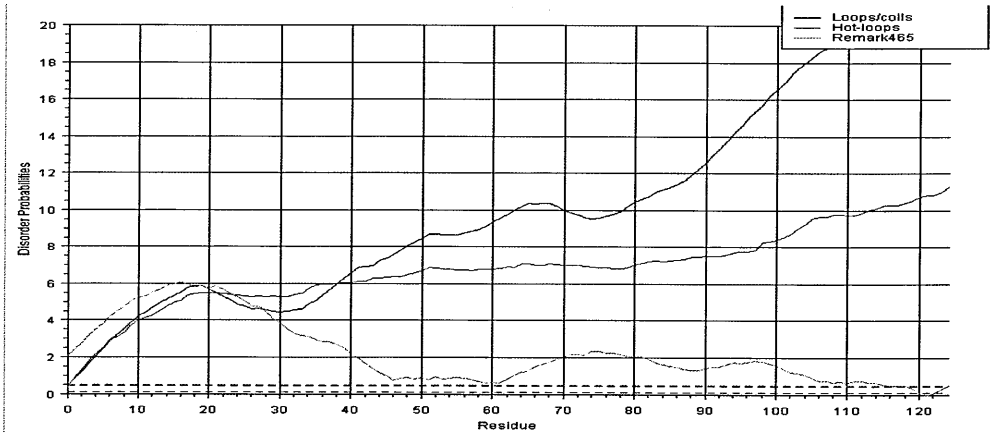
Globularity/disorder prediction

DISEMBL

Protein disorder predicted by DISEMBL. Disorder score is calculated using different predictors and plotted against the residue number. The Loops/Coils definition is based on the assignment of a secondary structure state other than helix or strand as disordered; the Hot Loop definition is based on Loops/Coils residues that display a high crystallographic B factor; the Remark-465 definition (missing coordinates in the PDB file) is based on residues that show no electron density in X-ray structures. Residues found to be disordered according to the above definitions are in bold capitals in the amino acid sequence.

Jagged -1





>jag1_LOOPS 1-19, 31-67, 76-125

RKRRKPGSHT HSASEDNTTn nvreqlnqik **NPIEKHGANT VPIKDYENKN SKMSKIRTHN SEVEEDD**mdk hqqka**RFAKQ PAYTLVDREE KPPNGTPTKH PNWTNKQDNR DLESAQSLNR MEYIV**

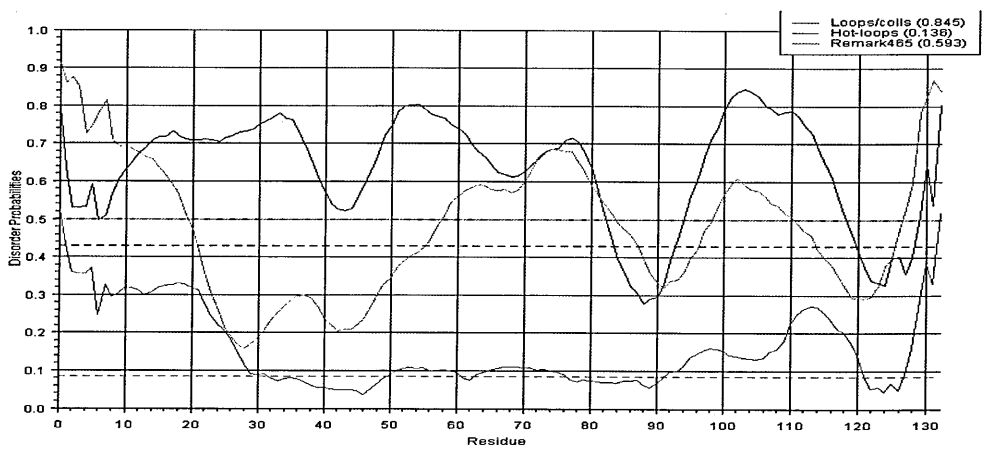
>jag1_HOTLOOPS 1-22, 30-69, 77-125

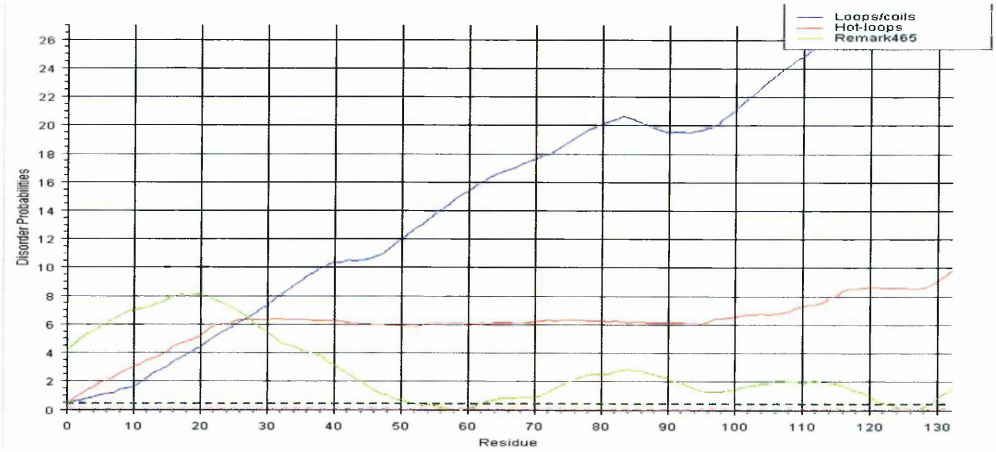
RKRRKPGSHT HSASEDNTTN NVreqlnqi**K NPIEKHGANT VPIKDYENKN SKMSKIRTHN SEVEEDDMD**k hqqkar**FAKQ PAYTLVDREE KPPNGTPTKH PNWTNKQDNR DLESAQSLNR MEYIV**

>jag1_REM465 1-18, 60-76

RKRRKPGSHT HSASEDNTtn nvreqlnqik npiekhgant vpikdyenkn skmskirth**N SEVEEDDMDK HQQKAR**fakq paytlvdree kppngtptkh pnwtnkqdnr dlesaqslnr meyiv

Jagged -2





>jag2_LOOPS 1-84, 94-120

TRKRRKERER SRLPREESAN NQWAPLNPIR NPIERPGGHK DVLYQCKNFT PPPRRADEAL PGPAGHAAVR EDEEDEDLGR GEEDsleaek fls**HKFTKDP GRSPGRPAHW ASGPKVDNRA** vrsinearya gke

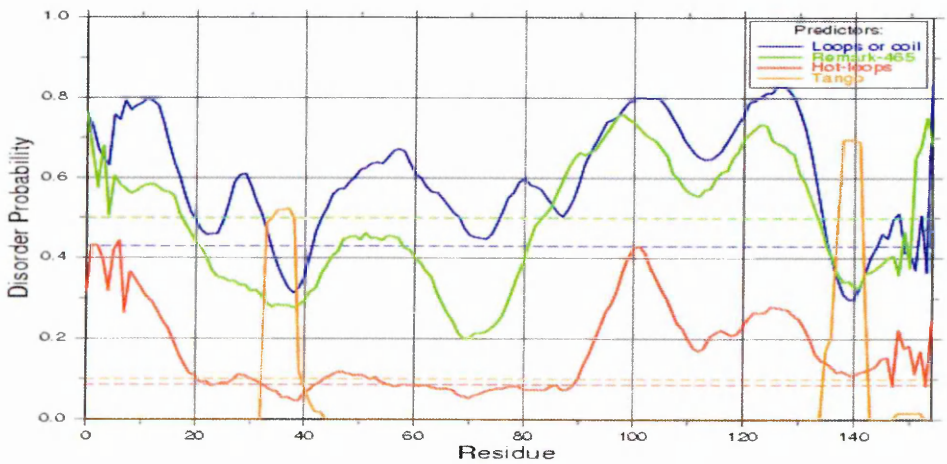
>jag2_HOTLOOPS 1-32, 93-122

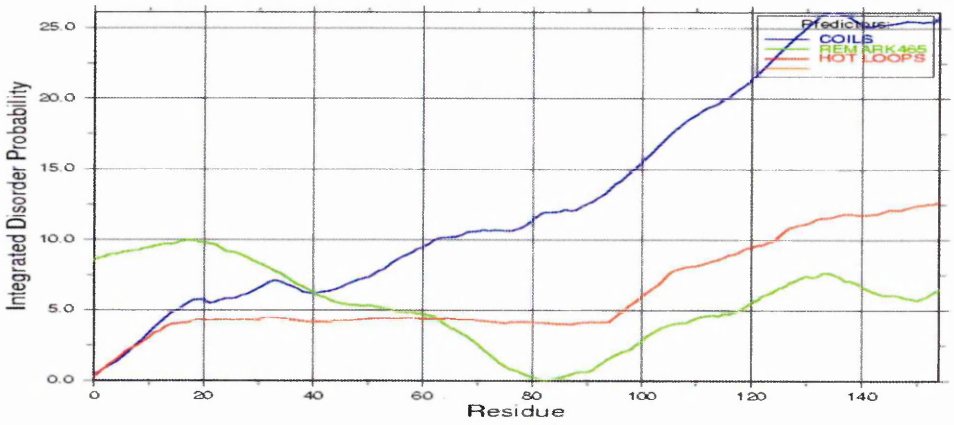
TRKRRKERER SRLPREESAN NQWAPLNPIR NPierpgghk dvlyqcknft ppprradeal pgpaghaavr edeededlgr geedsleaek fl**SHKFTKDP GRSPGRPAHW ASGPKVDNRA VR**sinearya gke

>jag2_REM465 1-20, 59-84, 100-111

TRKRRKERER SRLPREESAN nqwaplnpir npierpgghk dvlyqcknft ppprrade**AL PGPAGHAAVR EDEEDEDLGR GEED**sleaek flshkftkd**P GRSPGRPAHW A**sgpkvdnra vrsinearya gke

Delta -1





>dll1_LOOPS 1-34, 43-136

VRLRLQKHRP PADPCRGETE TMNNLANCQR EKDIsvsiig at**QIKNTNKK ADFHGDHSAD KNGFKARYPA VDYNLVQDLK GDDTAVRDAH SKRDTKCQPQ GSSGEEKGTP TTLRGGEASE RKRPDSGCST SKDTKY**qsvy viseekdecv iatev

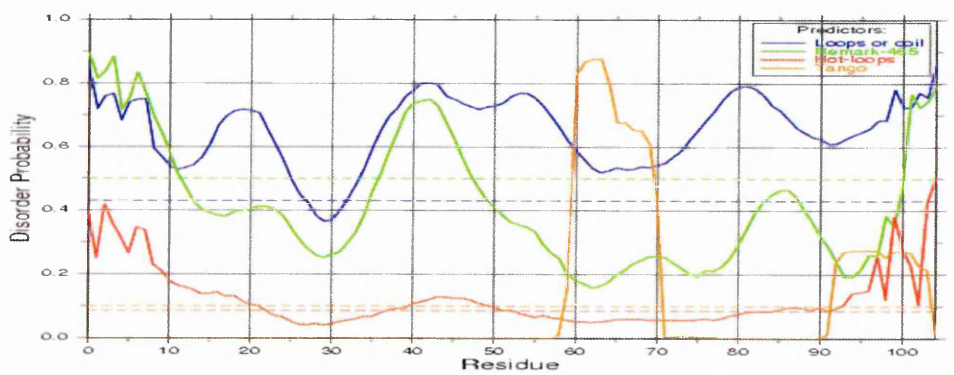
>dll1_HOTLOOPS 1-23, 91-155

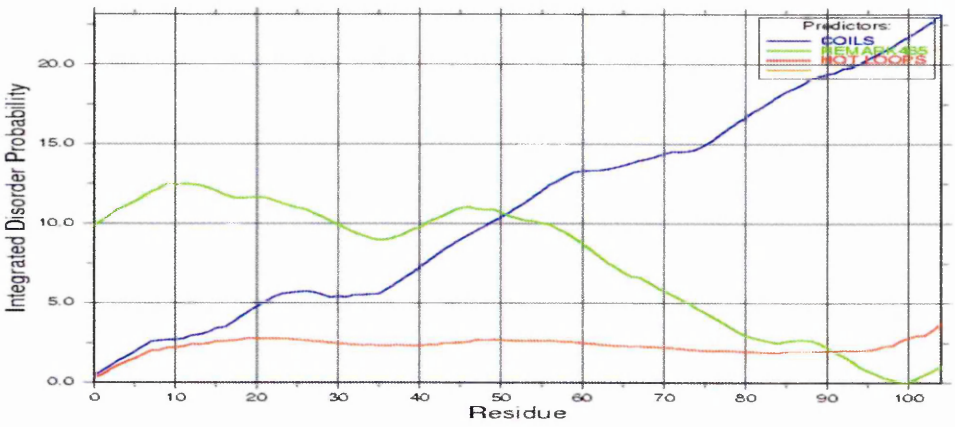
VRLRLQKHRP PADPCRGETE TMNnlancqr ekdisvsiig atqikntnkk adfhgdhsad kngfkarypa vdynlvqdlk gddtavrdah **SKRDTKCQPQ GSSGEEKGTP TTLRGGEASE RKRPDSGCST SKDTKYQSVY VISEEKDECV IATEV**

>dll1_REM465 1-18, 85-134

VRLRLQKHRP PADPCRGEte tmnnlancqr ekdisvsiig atqikntnkk adfhgdhsad kngfkarypa vdynlvqdlk gddt**AVRDAH SKRDTKCQPQ GSSGEEKGTP TTLRGGEASE RKRPDSGCST SKDT**kyqsvy viseekdecv iatev

Delta -3





>dll3_LOOPS 1-27, 33-105

HVRRRGHSQD AGSRLLAGTP EPSVHALpda ln**NLRTQEGS GDGPSSSVDW NRPEDVDPQG IYVISAPSIY AREVATPLFP PLHTGRAGQR QHLLFPYPSS ILSVK**

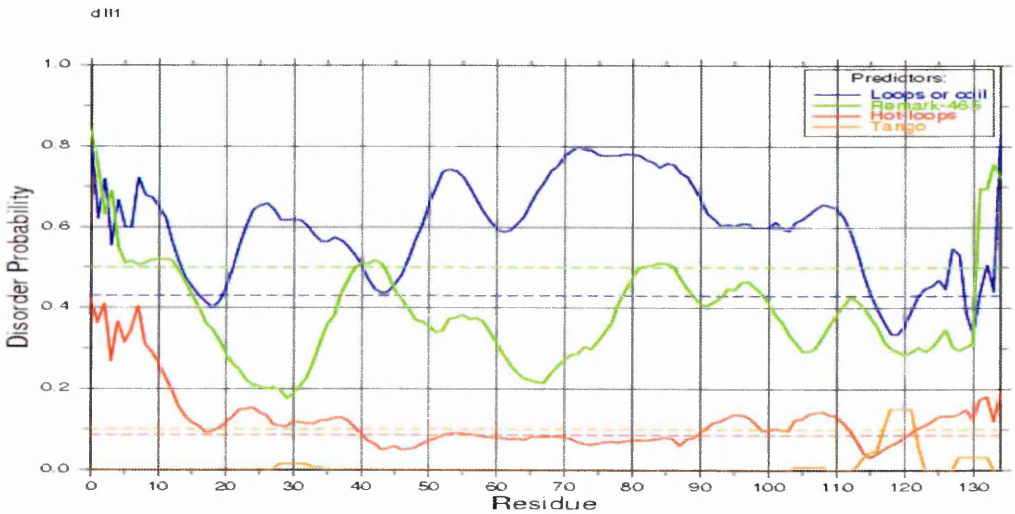
>dll3_HOTLOOPS 1-22, 38-51, 86-105

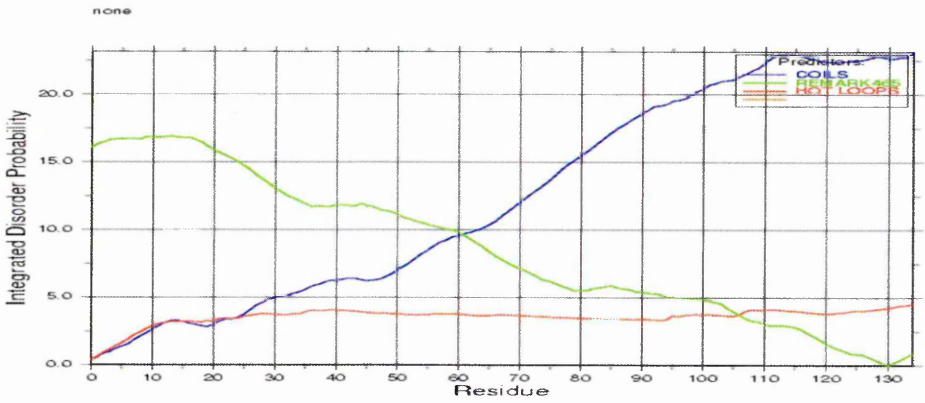
HVRRRGHSQD AGSRLLAGTP EPsvhalpda lnnlrtq**EGS GDGPSSSVDW N**rpedvdpqg iyvisapsiy arevatplfp plhtg**RAGQR QHLLFPYPSS ILSVK**

>dll3_REM465 1-12, 37-48

HVRRRGHSQD AGsrllagtp epsvhalpda lnnlrt**QEGS GDGPSSSV**dw nrpedvdpqg iyvisapsiy arevatplfp plhtgragqr qhllfpypss ilsvk

Delta -4





>dll4_LOOPS 1-16, 21-116, 124-135

AVRQLRLRRP DDGSREamnn **LSDFQKDNLI PAAQLKNTNQ KKELEVDCGL DKSNCGKQQN HTLDYNLAPG PLGRGTMPGK FPHSDKSLGE KAPLRLHSEK PECRISAICS PRDSMY**qsvc lis**EERNECV IATEV**

>dll4_HOTLOOPS 1-41, 91-113, 122-135

AVRQLRLRRP DDGSREAMNN LSDFQKDNLI PAAQLKNTNQ Kkelevdcgl dksncgkqqn htldynlapg plgrgtmpgk fphsdkslge **KAPLRLHSEK PECRISAICS PRD**smyqsvc l**ISEERNECV IATEV**

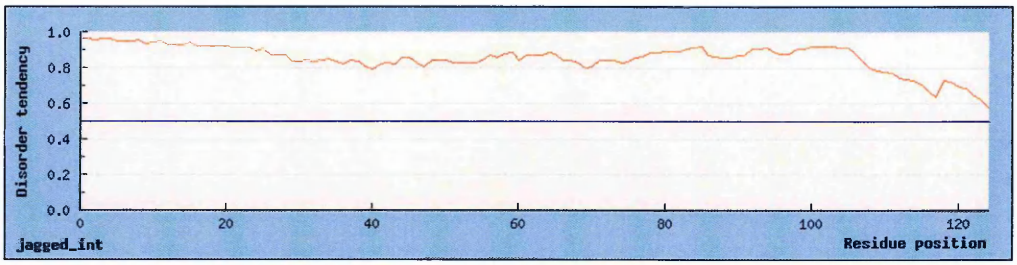
>dll4_REM465 1-13

AVRQLRLRRP DDGsreamnn lsdfqkdnli paaqlkntnq kkelevdcgl dksncgkqqn htldynlapg plgrgtmpgk fphsdkslge kaplrlhsek pecrisaics prdsmyqsvc liseernecv iatev

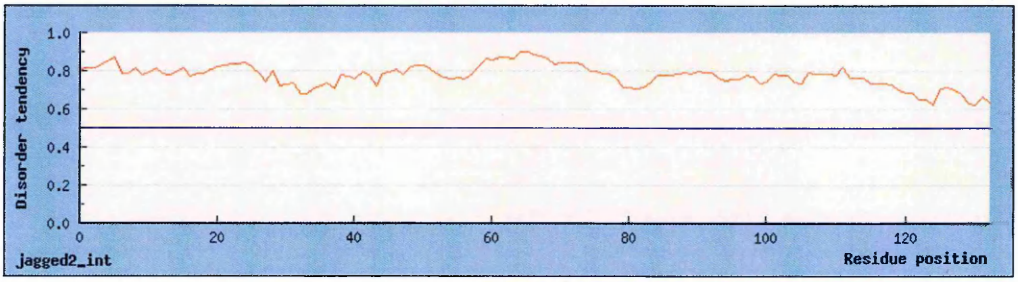
IUPRED

Protein disorder tendency predicted by IUPRED is predicted from the pairwise energy content estimated from the amino acid composition and averaged over a window of 21 residues. A value of 0.5 is considered as the disorder threshold.

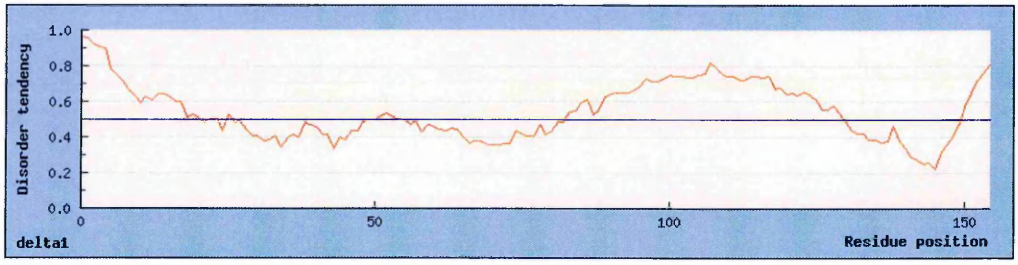
Jagged -1



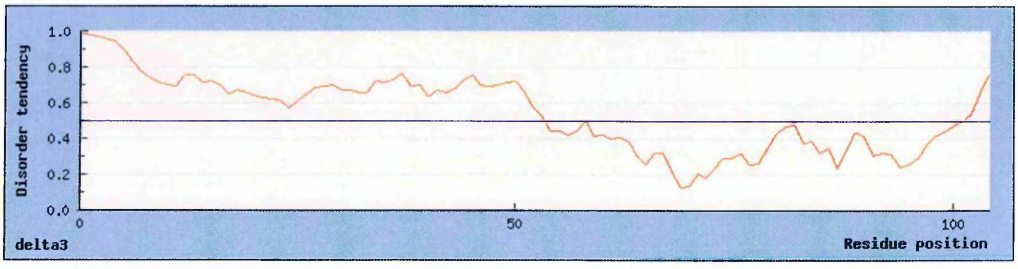
Jagged -2



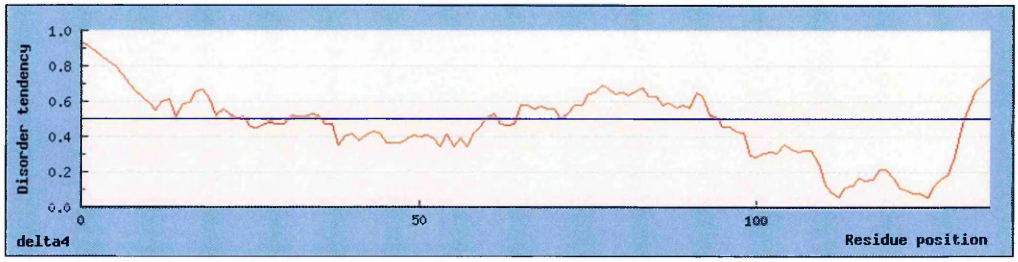
Delta -1



Delta -3



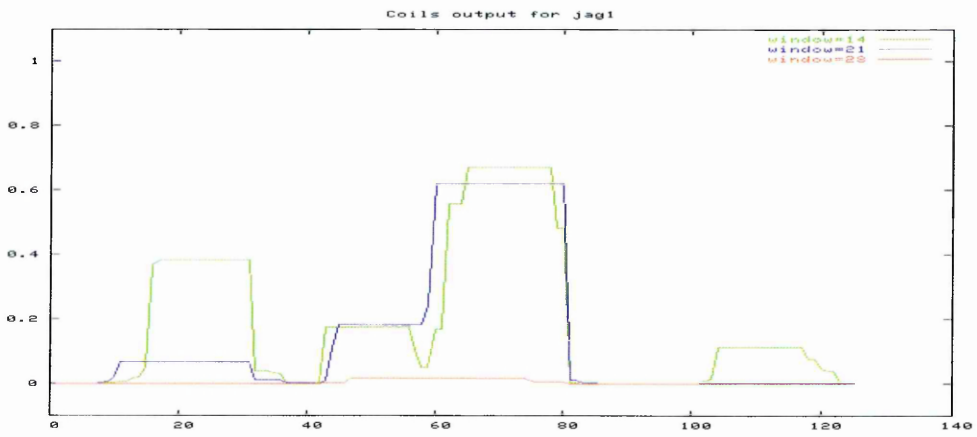
Delta -4



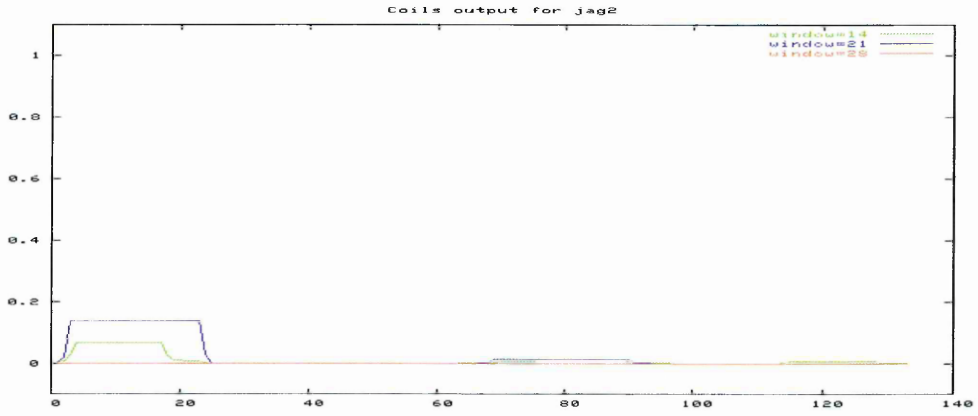
Coils

Prediction of coiled coil regions. The coil probability is plotted against the amino acid sequence number using different windows. For comparison prediction for coil-coil regions of human keratin are shown. One could see the big difference between keratin's coil regions and the predictions for the Jagged and Delta families. This visualization once more proofs the speculation for the disorderness of the intercellular part of both protein families.

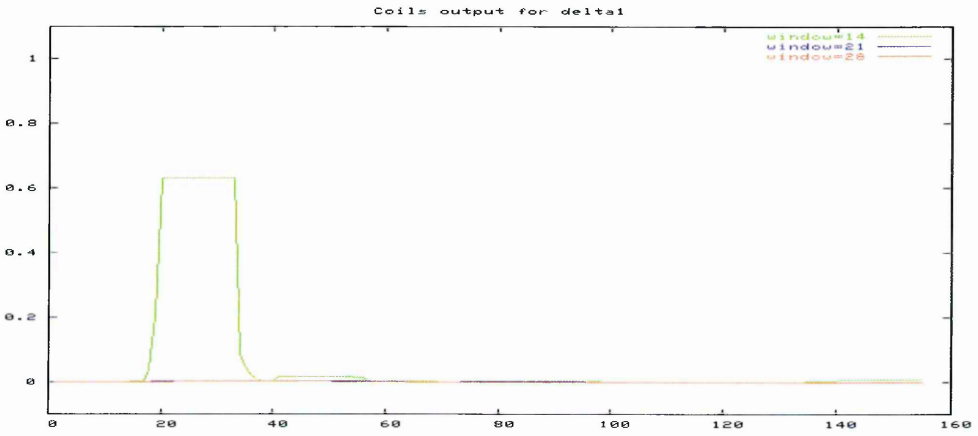
Jagged -1



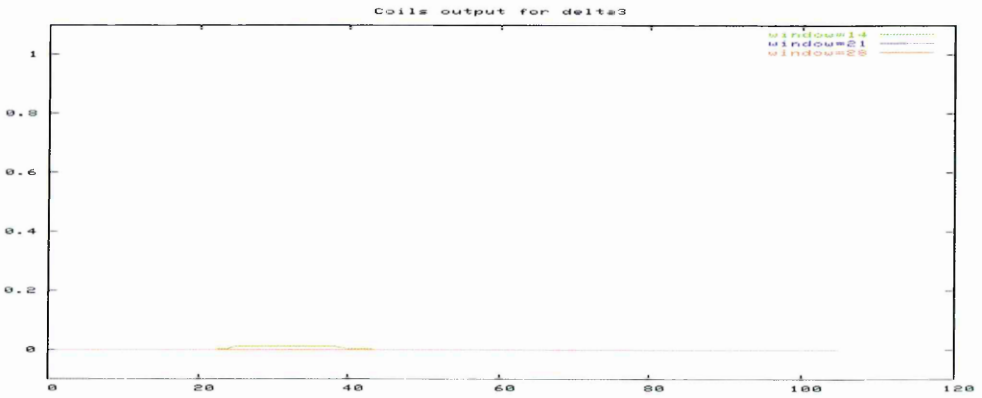
Jagged -2



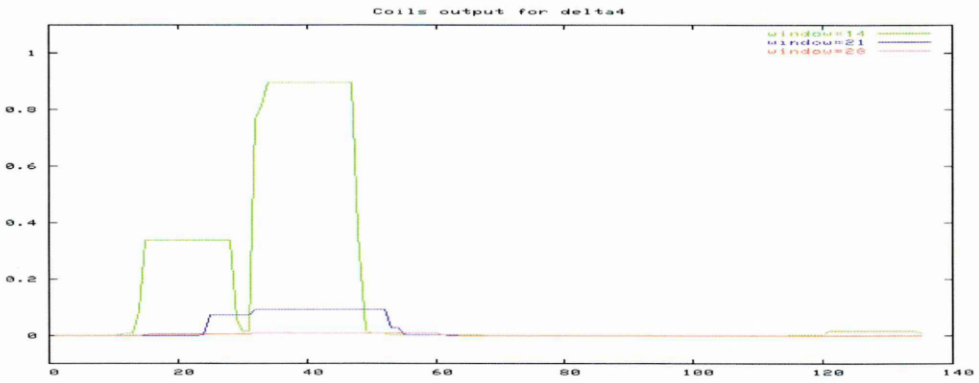
Delta -1



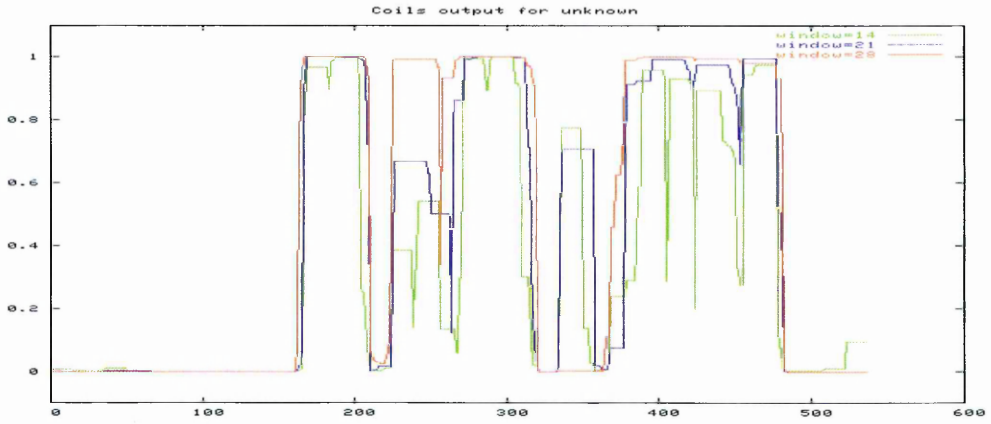
Delta -3



Delta -4



Keratin human as control



Key	From	To	Length	Description
CHAIN	1	494	494	Keratin, type I cytoskeletal 12.
REGION	1	124	124	Head.
REGION	125	435	311	Rod.
REGION	125	160	36	Coil 1A.
REGION	164	182	19	Linker 1.
REGION	183	274	92	Coil 1B.
REGION	275	297	23	Linker 12.
REGION	298	435	138	Coil 2.
REGION	436	494	59	Tail.

According t Swiss-Prot these are the predicted regions in the sequence of the Q99456|K1C12_HUMAN Keratin, type I cytoskeletal 12 - Homo sapiens (Human).

Pattern recognition

ELM

Comparative pattern recognition of functional sites in Jagged-1 and -2 intra-cellular region. Motifs found in the "plasma membrane" cellular compartment are shown in a green background, additional motifs found in the "cytoplasm" are shown in a light blue background; no additional motifs were found in the "nuclear" compartment (red background). Legend: **LIG_14-3-3_3**, 14-3-3 proteins interacting motif (Ser/Thr phosphorylation required); **LIG_FHA_1**, forkhead-associated domain interaction motif 1, (Thr phosphorylation required); **LIG_PDZ_2**, class II PDZ domains interacting motif; **LIG_PDZ_3**, class III PDZ domains binding motif; **LIG_SH2_GRB2**, Src Homology 2 (SH2) domains interaction motif (tyrosine phosphorylation required); **LIG_SH2_STAT5**, STAT5 Src Homology 2 (SH2) domain binding motif (tyrosine phosphorylation required); **LIG_SH3_2** class II SH3 domains binding motif; **LIG_SH3_3**, non-canonical class I SH3 domains binding motif; **LIG_SH3_5**, PXXDY motif recognized by some SH3 domains; **LIG_TRAF2_1**, tumor necrosis factor receptor associated protein binding motif; **LIG_WW_3**, group III WW domain binding motif; **LIG_WW_4**, class IV WW domains interaction motif (phosphorylation-dependent interaction); **TRG_ENDOCYTIC_2**, sorting signal responsible for the interaction with mu subunit of AP (Adaptor Protein) complex; **MOD_CDK**, Ser/Thr cyclin dependent kinase (CDK) phosphorylation site; **MOD_CK1_1**, casein kinase 1 (CK2) Ser/Thr phosphorylation motif; **MOD_CK2_1**, casein kinase 2 (CK2) Ser/Thr phosphorylation motif; **MOD_GSK3_1**, glycogen synthase kinase 3 Ser/Thr phosphorylation site; **MOD_PLK**, Polo-like-kinase Ser/Thr phosphorylation

site; **MOD_ProDKin_1**, Proline-Directed Kinase Ser/Thr phosphorylation site;

Jagged-1	Jagged-2
Plasma membrane	
LIG_14-3-3_3 HTHSAS HSASED 9-14 11-16 [RHK][STALV].[ST].[PESRDIF]	
LIG_FHA_1 THSA TVPI 10-13 40-43 T..[ILA]	
LIG_PDZ_2 EYIV 122-125 .[VYF].[VIL]	
LIG_PDZ_3 REQL MEYI 23-26 121-124 .[DE].[IVL]	LIG_PDZ_3 KDVL DEAL DEDL EDSL 40-43 57-60 75-78 83-86 .[DE].[IVL]
LIG_SH2_GRB2 YENK 46-49 Y.N.	
LIG_SH3_5 PIKDY 42-46 P..DY	
LIG_SH2_STAT5 YTLV 83-86 Y[VLTFIC]..	
	LIG_SH3_2 PLNPIR 25-30 P..P.[KR]
	LIG_SH3_3 QWAPLNP GRSPGRP 22-28 101-107 ...[PV]..P
TRG_ENDOCYTIC_2 YTLV 83-86 Y..[LMVIF]	
Cytoplasm	
	LIG_TRAF2_1 PREE 14-17 [PSAT].[QE]E
	LIG_WW_3 PPPRR 51-55 .PPR.
LIG_WW_4 PNGTPT 93-98 ...[ST]P.	LIG_WW_4 KNFTPP PGRSPG 47-52 100-105 ...[ST]P.
MOD_CDK	MOD_CDK

PNGTPTK 93-99 ...([ST])P.[KR]	PGRSPGR 100-106 ...([ST])P.[KR]
MOD_CK1_1 SKMSKIR SAQSLNR 51-57 114-120 S..([ST])...	
MOD_CK2_1 HTHSASE THNSEVE 9-15 58-64 ...([ST])..E	MOD_CK2_1 AVRSINE 120-126 ...([ST])..E
MOD_GSK3_1 KPGSHTHS GSHTHSAS HSASEDNT SKMSKIRT 5-12 7-14 11-18 51-58 ...([ST])...[ST]	MOD_GSK3_1 KFLSHKFT 90-97 ...([ST])...[ST]
	MOD_PLK REESANN EEDSLEA 15-21 82-88 .[DE].[ST][ILFWMVA]..
MOD_ProDKin_1 PNGTPTK 93-99 ...([ST])P..	MOD_ProDKin_1 KNFTPPP PGRSPGR 47-53 100-106 ...([ST])P..
Nucleus	
-	-

Comparative pattern recognition of functional sites in Delta intracellular region. Motifs found in the "plasma membrane" cellular compartment are shown in a green background, additional motifs found in the "cytoplasm" are shown in a light blue background; no additional motifs were found in the "nuclear" compartment (red background). Where motif recognition requires phosphorylation, the phosphorylated Ser/Thr/Tyr residue is in red. Legend: **LIG_CYCLIN_1**, cyclin recognition site; **LIG_FHA_1**, forkhead-associated domain interaction motif 1, (Thr phosphorylation required); **LIG_PDZ_1**, class I PDZ domains interacting motif; **LIG_PDZ_3**, class III PDZ domains binding motif; **LIG_SH2_SRC**, Src Homology 2 (SH2) domains interaction motif (tyrosine phosphorylation required); **LIG_SH2_STAT5**, STAT5 Src Homology 2 (SH2) domain binding motif; **LIG_SH3_2** class II SH3 domains binding motif; **LIG_SH3_3**, non-canonical class I SH3 domains binding motif; **LIG_SH3_5**, PXXDY motif recognized by some SH3 domains;

LIG_TRAF2_1, tumor necrosis factor receptor associated protein binding motif; **LIG_WW_4**, class IV WW domains interaction motif (phosphorylation-dependent interaction); **TRG_ENDOCYTIC_2**, sorting signal responsible for the interaction with mu subunit of AP (Adaptor Protein) complex; **MOD_CK1_1**, casein kinase 1 (CK2) Ser/Thr phosphorylation motif; **MOD_CK2_1**, casein kinase 2 (CK2) Ser/Thr phosphorylation motif; **MOD_GSK3_1**, glycogen synthase kinase 3 Ser/Thr phosphorylation site; **MOD_PK_1**, phosphorylase kinase Ser/Thr phosphorylation site; **MOD_PKA_1**, cAMP-dependent protein kinase A (PKA) Ser/Thr phosphorylation site; **MOD_PKA_2**, cAMP-dependent protein kinase A (PKA) Ser/Thr phosphorylation site; **MOD_PKB_1**, Protein kinase B Ser/Thr phosphorylation site; **MOD_PLK**, Polo-like-kinase Ser/Thr phosphorylation site; **MOD_ProDKin_1**, Proline-Directed Kinase Ser/Thr phosphorylation site; **MOD_TYR_ITIM**, immunoreceptor tyrosine-based inhibitory motif (tyrosine phosphorylation required); **MOD_TYR_ITSM**, immunoreceptor tyrosine-based switch motif (tyrosine phosphorylation required).

Delta-1	Delta-3	Delta-4
Plasma membrane		
	LIG_FHA_1 TGRA 84-87 T..[ILA]	
LIG_PDZ_1 ATEV 152-155 .[ST].[VIL]		LIG_PDZ_1 ATEV 132-135 .[ST].[VIL]
LIG_PDZ_3 DECV 147-150 .[DE].[IVL]	LIG_PDZ_3 PDAL PEDV 28-31 53-56 .[DE].[IVL]	LIG_PDZ_3 KDNL NECV 26-29 127-10 .[DE].[IVL]

LIG_SH2_SRC YVIS 140-143 Y[QDEVAIL][DENPYHI][IPVGAHS]	LIG_SH2_SRC YVIS 62-65 Y[QDEVAIL][DENPYHI][IPVGAHS]	
LIG_SH2_STAT5 YVIS 140-143 Y[VLTFIC]..	LIG_SH2_STAT5 YVIS 62-65 Y[VLTFIC]..	
LIG_SH3_2 PADPCR 11-16 P..P.[KR]		
LIG_SH3_3 HRPPADP 8-14 ...[PV]..P	LIG_SH3_3 AREVATP VATPLFP 71-77 74-80 ...[PV]..P	
LIG_SH3_5 PAVDY 69-73 P..DY		
TRG_ENDOCYTIC_2 YPAV YNLV YQSV 68-71 73-76 136-139 Y..[LMVIF]		TRG_ENDOCYTIC_2 YQSV 116-119 Y..[LMVIF
Cytoplasm		
		LIG_CYCLIN_1 RQLRL KELEV 3-7 42-46
LIG_TRAF2_1 SGEE 103-106 [PSAT].[QE]E		
LIG_WW_4 EKGTPT 106-111 ...[ST]P	LIG_WW_4 LAGTPE EVATPL 16-21 73-78 ...[ST]P.	LIG_WW_4 AICSPR 107-112 ...[ST]P.
MODCK1_1 SGCTSK SKDTKYQ 126-132 131-137 S..([ST])...	MOD_CK1_1 SAPSIYA 65-71 S..([ST])...	MOD_CK1_1 SDKSLGE 84-90 S..([ST])...
MOD_CK2_1 PQGSSGE QGSSGEE 99-105 100-106 ...([ST])..E		MOD_CK2_1 SDKSLGE 84-90 ...([ST])..E
MOD_GSK3_1 DAHSKRDT RPDSGCST 88-95 123-130 GCSTSKDT SKDTKYQS 127-134 131-138	MOD_GSK3_1 LAGTPEPS NLRTQEGS PYPSSILS 16-23 33-40 96-103 ...([ST])...[ST]	MOD_GSK3_1 CRISAICS AICSPRDS PRDSMYQS 103-110 107-114 111-118

...([ST])...[ST]		...([ST])...[ST]
MOD_PK_1 KDISVSI KYQSVYV 32-38 135-141 [RK]..(S)[VI]..		
MOD_PKA_1 KRDTKCQ 92-98 [RK][RK].[ST]...		
MOD_PKA_2 KRDTKCQ 92-98 .R.([ST])...		MOD_PKA_2 GRGTMPG CRISAIC PRDSMYQ 73-79 103-109 111-117 .R.([ST])...
MOD_PKB_1 RKRPDSGCS 121-129 R.R..([ST])...	MOD_PKB_1 RRRGHSQDA 3-11 R.R..([ST])...	
MOD_PLK KDISVSI GDHSADK GDDTAVR 32-38 55-61 81-87 .[DE].[ST][ILFWMVA]..	MOD_PLK PEPSVHA 20-26 .[DE].[ST][ILFWMVA]..	MOD_PLK SDKSLGE 84-90 .[DE].[ST][ILFWMVA]..
MOD_ProDKin_1 EKGTPTT 106-112 ...([ST])P..	MOD_ProDKin_1 LAGTPEP EVATPLF 16-22 73-79 ...([ST])P..	MOD_ProDKin_1 AICSPRD 107-113 ...([ST])P..
MOD_TYR_ITIM VDYNLV 71-76 [ILV].(Y)..[ILV]		
MOD_TYR_ITSM KDTKYQSV 132-139 ..T.(Y)..[IV]		
Nucleus		
	-	-

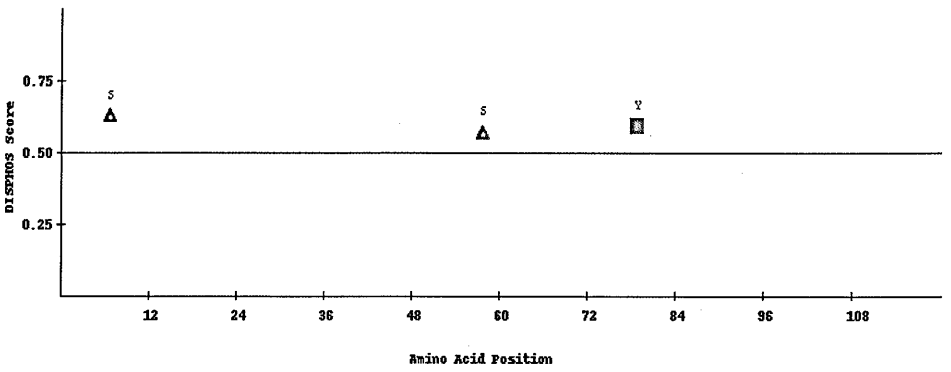
Phosphorylation

DISPHOS

Serine, threonine and tyrosine phosphorylation sites predicted by DISPHOS (Disorder-enhanced phosphorylation sites predictor). A plot of DISPHOS score is shown against the residue number, and residues that are above the threshold of 0.5 are marked.

Jagged -1

RKRRKPGSHTHSASEDNTTNNVREQLNQIKNPIEKHGANTVPIKDYENKNSKMSKIRTHNSEVEEDDMDKHQQKARF
AKQPAYTLVDREEKPPNGTPTKHPNWTNKQDNRDLESAQSLNRMEYIV

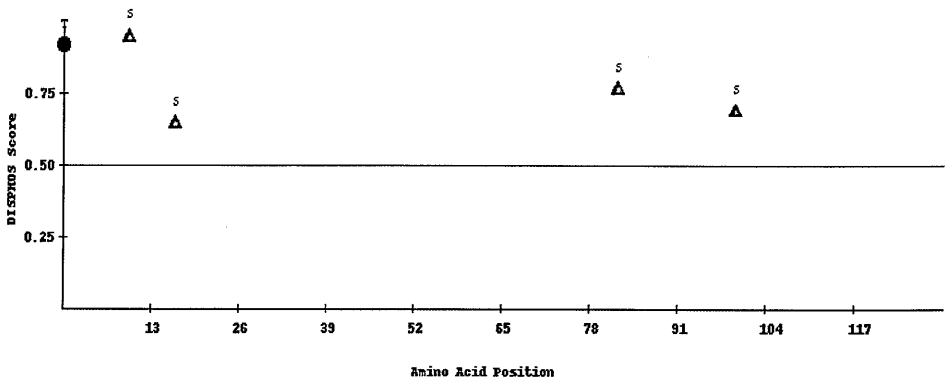


Position	Residue	Score	Sequence	Yes/No
8	S	0.629	RKPGSHTHS	YES
10	T	0.047	PGSHTHSAS	
12	S	0.330	SHTHSASED	
14	S	0.270	THSASEDNT	

18	T	0.019	SEDN**T**TNNV	
19	T	0.008	EDNT**T**NNVR	
40	T	0.013	HGAN**T**VPIK	
46	Y	0.085	PIKD**Y**ENKN	
51	S	0.124	ENKN**S**KMSK	
54	S	0.389	NSKM**S**KIRT	
58	T	0.106	SKIR**T**HNSE	
61	S	0.571	RTHN**S**EVEE	YES
83	Y	0.593	KQPA**Y**TLVD	YES
84	T	0.138	QPAY**T**LVDR	
96	T	0.467	PPNG**T**PTKH	
98	T	0.147	NGTP**T**KHPN	
104	T	0.038	HPNW**T**NKQD	
114	S	0.066	RDLE**S**AQSL	
117	S	0.223	ESAQ**S**LNRM	
123	Y	0.012	NRME**Y**IV**	

Jagged -2

TRKRRKERERSRLPREESANNQWAPLNPIRNPIERPGGHKDVL**Y**QCKNF**T**PPPRRADEALPGPAGHAAVREDEEDEDL
GRGEEDSLEAEKFLSHKF**T**KDPGRSPGRPAHWASGPKVDNRAVRSINEAR**Y**AGKE

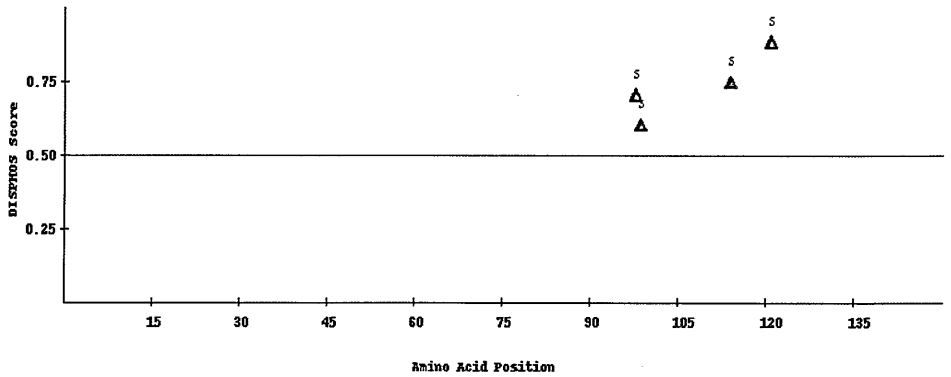


Position	Residue	Score	Sequence	Yes/No
1	T	0.920	****TRKRR	YES
11	S	0.950	ERERSRLPR	YES
18	S	0.649	PREESANNQ	YES
44	Y	0.190	KDVLYQCKN	

50	T	0.138	CKNF**T**PPPR	
85	S	0.769	GEED**S**LEAE	YES
93	S	0.412	EKFL**S**HKFT	
97	T	0.407	SHKF**T**KDPG	
103	S	0.692	DPGR**S**PGRP	YES
112	S	0.362	AHWA**S**GPKV	
123	S	0.208	RAVR**S**INEA	
129	Y	0.070	NEAR**Y**AGKE	

Delta -1

VRLRLQKHRPPADPCRGETETMNNLANCQREKDISVSIIGATQIKNTNKKADFHGDHSADKNGFKARYPAVDYNLVQ
DLKGDDTAVRDAHSKRDTKCQPQGSSGEEKGTPTTLRGGEASERKRPDSGCSTSKDTKYQSVYVISEEKDECVIATEV

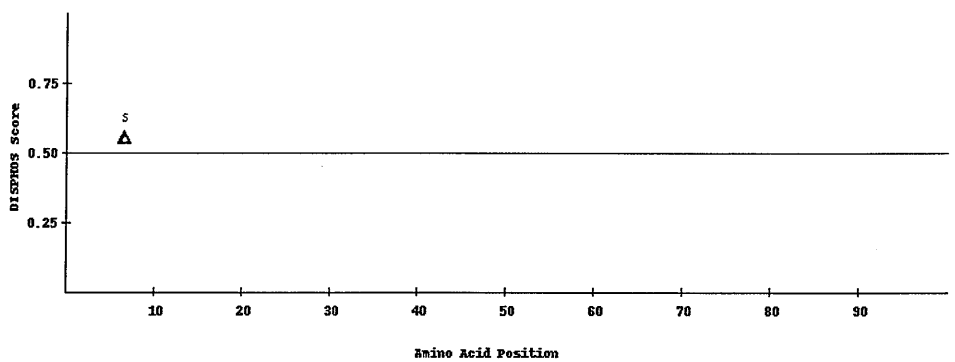


Position	Residue	Score	Sequence	Yes/No
19	T	0.148	CRGE**T**ETMN	
21	T	0.029	GE**T**ETMNNL	
35	S	0.078	EKDI**S**VSII	
37	S	0.099	DIS**V**SIIGA	
42	T	0.019	IIGA**T**QIKN	
47	T	0.025	QIKN**T**NKKA	
58	S	0.117	HGDH**S**ADKN	
68	Y	0.017	FKAR**Y**PAVD	
73	Y	0.236	PAVD**Y**NLVQ	

84	T	0.299	KGDD**T**AVRD	
91	S	0.380	RDAH**S**KRDT	
95	T	0.433	SKRD**T**KCQP	
102	S	0.704	QPQG**S**SGEE	YES
103	S	0.602	PQGS**S**GEEK	YES
109	T	0.260	EEKG**T**PTTL	
111	T	0.174	KGTP**T**TLRG	
112	T	0.297	GTPT**T**LRGG	
119	S	0.748	GGEA**S**ERKR	YES
126	S	0.885	KRPD**S**GCST	YES
129	S	0.460	DSGC**S**TSKD	
130	T	0.226	SGCS**T**SKDT	
131	S	0.249	GCST**S**KDTK	
134	T	0.202	TSKD**T**KYQS	
136	Y	0.313	KDTK**Y**QSVY	
138	S	0.430	TKYQ**S**VYVI	
140	Y	0.167	YQSV**Y**VISE	
143	S	0.243	VYVI**S**EEKD	
153	T	0.032	CVIA**T**EV**	

Delta -3

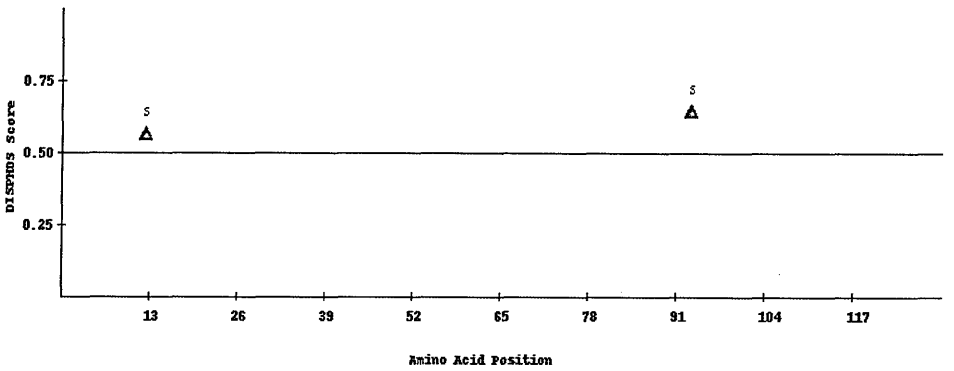
HVRRRGHSQDAGSRLLAGTPEPSVHALPDALNNLRTQEGSGDGPSSSVDWNRPEDVDPQGIYVISAPSIYAREVATPLF
PPLHTGRAGQRQHLLFPYPSSILSVK



Position	Residue	Score	Sequence	Yes/No
8	S	0.555	RRGH**S**QDAG	YES
13	S	0.500	QDAG**S**RLLA	
19	T	0.090	LLAG**T**PEPS	
23	S	0.329	TPEP**S**VHAL	
36	T	0.065	NNLR**T**QEGS	
40	S	0.332	TQEG**S**GDGP	
45	S	0.473	GDGP**S**SSVD	
46	S	0.360	DGPS**S**SVDW	
47	S	0.372	GPSS**S**VDWN	
62	Y	0.186	PQGI**Y**VISA	
65	S	0.067	IYVI**S**APSI	
68	S	0.077	ISAP**S**IYAR	
70	Y	0.078	APSI**Y**AREV	
76	T	0.064	REVA**T**PLFP	
84	T	0.013	PPLH**T**GRAG	
97	Y	0.031	LLFP**Y**PSSI	
99	S	0.109	FPYP**S**SILS	
100	S	0.102	PYPS**S**ILSV	
103	S	0.225	SSIL**S**VK**	

Delta -4

AVRQLRLRRPDDGSREAMNNLSDFQKDNLIPAAQLKNTNQKKELEVDCGLDKSNCGKQQNHTLDYNLAPGPLGRGT
MPGKFPHSDKSLGEKAPLRLHSEKPECRISAICSPRDSMYQSVCLISEERNECVIATEV



Position	Residue	Score	Sequence	Yes/No
14	S	0.566	PDDG**S**REAM	YES
22	S	0.107	MNNL**S**DFQK	
38	T	0.015	QLKN**T**NQKK	
53	S	0.050	GLDK**S**NCGK	
62	T	0.065	QQNH**T**LDYN	
65	Y	0.109	HTLD**Y**NLAP	
76	T	0.237	LGRG**T**MPGK	
84	S	0.300	KFPH**S**DKSL	
87	S	0.409	HSDK**S**LGEK	
98	S	0.646	LRLH**S**EKPE	YES
106	S	0.448	ECRI**S**AICS	
110	S	0.262	SAIC**S**PRDS	
114	S	0.412	SPRD**S**MYQS	
116	Y	0.130	RDSM**Y**QSVC	
118	S	0.236	SMYQ**S**VCLI	
123	S	0.165	VCLI**S**EERN	
133	T	0.026	CVIA**T**EV**	

NetPhos

Serine, threonine, and tyrosine phosphorylation sites predicted by NetPhos (Neural network-based phosphorylation sites predictor).

The amino acid sequence and the scores for all serines, threonines, and tyrosines are shown. Residues predicted to be phosphorylated (score > 0.5) are shown in red in the amino acid sequence.

Jagged -1

RKRRKPGSHTHSASEDNTTNNVREQLNQIKNPIEKHGANTVPIKDYENKNSKMSKIRTHNSEVEEDDMDKHQQKARFAKQ 80

PAYTLVDREEKPPNGTPTKHPNWTNKQDNRDLESAQSLNRMEYIV 160

Serine predictions

Name	Pos	Context	Score	Pred
Sequence	8	RKPGSHTHS	0.979	*S*
Sequence	12	SHTHSASED	0.977	*S*
Sequence	14	THSASEDNT	0.787	*S*
Sequence	51	ENKNSKMSK	0.053	.
Sequence	54	NSKMSKIRT	0.996	*S*
Sequence	61	RTHNSEVEE	0.983	*S*
Sequence	114	RDLESAQSL	0.006	.
Sequence	117	ESAQSLNRM	0.412	.

Threonine predictions

Name	Pos	Context	Score	Pred
Sequence	10	PGSHTHSAS	0.054	.
Sequence	18	SEDNTTNNV	0.227	.
Sequence	19	EDNTTNNVR	0.021	.
Sequence	40	HGANTVPIK	0.253	.
Sequence	58	SKIRTHNSE	0.093	.
Sequence	84	QPAYTLVDR	0.516	*T*
Sequence	96	PPNGTPTKH	0.527	*T*
Sequence	98	NGTPTKHPN	0.504	*T*
Sequence	104	HPNWTNKQD	0.365	.

Tyrosine predictions

Name	Pos	Context	Score	Pred
Sequence	46	PIKDYENKN	0.821	*Y*
Sequence	83	KQPAYTLVD	0.754	*Y*
Sequence	123	NRMEYIV--	0.050	.

Jagged -2

TRKRRKERERSRLPREESANNQWAPLNPIRNPIERPGGHKDVLYQCKNFTPPPRRADEALPGPAGHAAVREDEEDEDLGR 80
GEEDSLEAEKFLSHKFTKDPGRSPGRPAHWASGPKVDNRAVRSINEARYAGKE 160

Serine predictions

Name	Pos	Context	Score	Pred

Sequence	11	ERERSRLPR	0.944	*S*
Sequence	18	PREESANNQ	0.141	.
Sequence	85	GEEDSLEAE	0.987	*S*
Sequence	93	EKFLSHKFT	0.992	*S*
Sequence	103	DPGRSPGRP	0.994	*S*
Sequence	112	AHWASGPKV	0.008	.
Sequence	123	RAVRSINEA	0.850	*S*

Threonine predictions

Name	Pos	Context	Score	Pred
Sequence	1	----TRKRR	0.583	*T*
Sequence	50	CKNFTPPPR	0.084	.
Sequence	97	SHKFTKDPG	0.230	.

Tyrosine predictions

Name	Pos	Context	Score	Pred
Sequence	44	KDVLYQCKN	0.775	*Y*
Sequence	129	NEARYAGKE	0.729	*Y*

Delta -1

VRLRLQKHRPPADPCRGETETMNNLANCQREKDISVSIIGATQIKNTNKKADFHGDHSADKNGFKARYPAVDYNLVQDLK 80

GDDTAVRDAHSKRDTKCQPQGSSGEEKGTPTTLRGGEASERKRPDSGCSTSKDTKYQSVYVISEEKDECVIATEV 160

Serine predictions

Name	Pos	Context	Score	Pred
delta1	35	EKDISVSII	0.766	*S*
delta1	37	DISVSIIGA	0.684	*S*
delta1	58	HGDHSADKN	0.669	*S*
delta1	91	RDAHSKRDT	0.998	*S*
delta1	102	QPQGSSGEE	0.957	*S*
delta1	103	PQGSSGEEK	0.955	*S*
delta1	119	GGEASERKR	0.860	*S*
delta1	126	KRPDSGCST	0.997	*S*
delta1	129	DSGCSTSKD	0.447	.
delta1	131	GCSTSKDTK	0.041	.
delta1	138	TKYQSVYVI	0.668	*S*

delta1	143	VYVISEEKD	0.807	*S*

Threonine predictions

Name	Pos	Context	Score	Pred
delta1	19	CRGETETMN	0.249	.
delta1	21	GETETMNNL	0.054	.
delta1	42	IIGATQIKN	0.032	.
delta1	47	QIKNTNKKA	0.424	.
delta1	84	KGDDTAVRD	0.753	*T*
delta1	95	SKRDTKCQP	0.972	*T*
delta1	109	EEKGTPTTL	0.350	.
delta1	111	KGTPTTLRG	0.086	.
delta1	112	GTPTTLRGG	0.950	*T*
delta1	130	SGCSTSKDT	0.980	*T*
delta1	134	TSKDTKYQS	0.474	.
delta1	153	CVIATEV--	0.050	.

Tyrosine predictions

Name	Pos	Context	Score	Pred
delta1	68	FKARYPAVD	0.517	*Y*
delta1	73	PAVDYNLVQ	0.254	.
delta1	136	KDTKYQSVY	0.981	*Y*
delta1	140	YQSVYVISE	0.685	*Y*

Delta -3

```
HVRRRGHSQDAGSRLLAGTPEPSVHALPDALNNLRTQEGSGDGPSSSVDWNRPEDVDPQGIYVISAPSIYAREVATPLFP    80
PLHTGRAGQRQHLLFPYPSSILSVK                                                          160
```

Serine predictions

Name	Pos	Context	Score	Pred
delta3	8	RRGHSQDAG	0.996	*S*
delta3	13	QDAGSRLLA	0.100	.
delta3	23	TPEPSVHAL	0.677	*S*
delta3	40	TQEGSGDGP	0.619	*S*

delta3	45	GDGPSSSVD	0.044	.
delta3	46	DGPSSSVDW	0.961	*S*
delta3	47	GPSSSVDWN	0.852	*S*
delta3	65	IYVISAPSI	0.004	.
delta3	68	ISAPSIYAR	0.620	*S*
delta3	99	FPYPSSILS	0.010	.
delta3	100	PYPSSILSV	0.006	.
delta3	103	SSILSVK--	0.913	*S*

Threonine predictions

Name	Pos	Context	Score	Pred
delta3	19	LLAGTPEPS	0.631	*T*
delta3	36	NNLRTQEGS	0.043	.
delta3	76	REVATPLFP	0.256	.
delta3	84	PPLHTGRAG	0.399	.

Tyrosine predictions

Name	Pos	Context	Score	Pred
delta3	62	PQGIYVISA	0.914	*Y*
delta3	70	APSIYAREV	0.154	.
delta3	97	LLFPYPSSI	0.009	.

Delta -4

AVRQLRLRRPDDGSREAMNNLSDFQKDNLIPAAQLKNTNQKKELEVDCGLDKSNCGKQQNHTLDYNLAPGPLGRGTMPGK 80

FPHSDKSLGEKAPLRLHSEKPECRISAICSPRDSMYQSVCLISEERNECVIATEV 160

Serine predictions

Name	Pos	Context	Score	Pred
delta4	14	PDDGSREAM	0.825	*S*
delta4	22	MNNLSDFQK	0.027	.
delta4	53	GLDKSNCGK	0.005	.
delta4	84	KFPHSDKSL	0.994	*S*
delta4	87	HSDKSLGEK	0.989	*S*
delta4	98	LRLHSEKPE	0.996	*S*
delta4	106	ECRISAICS	0.964	*S*
delta4	110	SAICSPRDS	0.995	*S*
delta4	114	SPRDSMYQS	0.991	*S*

delta4	118	SMYQSVCLI	0.022	.
delta4	123	VCLISEERN	0.984	*S*

Threonine predictions

Name	Pos	Context	Score	Pred
delta4	38	QLKNTNQKK	0.014	.
delta4	62	QQNHTLDYN	0.004	.
delta4	76	LGRGTMPGK	0.361	.
delta4	133	CVIATEV--	0.050	.

Tyrosine predictions

Name	Pos	Context	Score	Pred
delta4	65	HTLDYNLAP	0.180	.
delta4	116	RDSMYQSVC	0.894	*Y*

YIN-O-YANG: PHOSPHORYLATION VS. GLYCOSYLATION

Neural network-based predictions of O-ß-GlcNAc attachment sites in eukaryotic proteins are combined with predictions of Ser/Thr phosphorylation sites (NetPhos) to identify potential Yin-Yang sites. Scores for glycosylation and phosphorylation for all serines and threonines are shown, and Ying-Yang sites are shown in red and underlined in the amino acid sequence.

Jagged -1

RKRRKPGSHTHSASEDNTTNNVREQLNQIKNPIEKHGANTVPIKDYENKNSKMSKIRTHNSEVEEDDMDKHQQKARFAKQ
PAYTLVDREEKPPNGTPTKHPNWTNKQDNRDLESAQSLNRMEYIV

SeqName	Residue	O-GlcNAc result	Potential (o-glcnac)	Thresh. (1)	Thresh. (2)	NetPhos potential (Thresh=0.5)	YinOYang?
Sequence	8	S*	+++ 0.6057	0.3358	0.4030	0.979	*

Jagged -2

RKRRKERERSRLPREESANNQWAPLNPIRNPIERPGGHKDVLYQCKNFTPPPRRADEALPGPAGHAAVREDEEDEDLGRG
EEDSLEAEKFLSHKFTKDPGRSPGRPAHWASGPKVDNRAVRSINEARYAGKE

SeqName	Residue	O-GlcNAc result	Potential (o-glcnac)	Thresh. (1)	Thresh. (2)	NetPhos potential (Thresh=0.5)	YinOYang?
Sequence	102	S*	++ 0.4014	0.3176	0.3785	0.994	*
Sequence	111	S	+ 0.3997	0.3699	0.4490		

Delta -1

RLRLQKHRPPADPCRGETETMNNLANCQREKDISVSIIGATQIKNTNKKADFHGDHSADKNGFKARYPAVDYNLVQDLKG
DDTAVRDAHSKRDTKCQPQGSSGEEKGTPTTLRGGEASERKRPDSGCSTSKDTKYQSVYVISEEKDECVIATEV

SeqName	Residue	O-GlcNAc result	Potential (o-glcnac)	Thresh. (1)	Thresh. (2)	NetPhos potential (Thresh=0.5)	YinOYang?
Sequence	83	T*	+ 0.3597	0.3584	0.4335	0.753	*
Sequence	101	S*	++ 0.3921	0.3143	0.3740	0.957	*
Sequence	102	S*	+ 0.3567	0.3172	0.3780	0.955	*
Sequence	125	S*	+ 0.3914	0.3446	0.4149	0.997	*
Sequence	130	S	+ 0.3739	0.3498	0.4219		
Sequence	133	T	+ 0.3434	0.3373	0.4050		
Sequence	152	T	+ 0.4400	0.3711	0.4506		

Delta -3

RRRGHSQDAGSRLLAGTPEPSVHALPDALNNLRTQEGSGDGPSSSVDWNRPEDVDPQGIYVISAPSIYAREVATPLFPPL
HTGRAGQRQHLLFPYPSSILSVK

SeqName	Residue	O-GlcNAc result	Potential (o-glcnac)	Thresh. (1)	Thresh. (2)	NetPhos potential (Thresh=0.5)	YinOYang?
Sequence	6	S*	++	0.4593	0.3262	0.3901	0.996	*
Sequence	17	T*	+	0.4126	0.3521	0.4250	0.631	*
Sequence	38	S*	+	0.3334	0.3151	0.3751	0.619	*
Sequence	44	S*	++	0.5264	0.3611	0.4372	0.961	*
Sequence	82	T	++	0.4694	0.3768	0.4583		
Sequence	97	S	+++	0.6411	0.4107	0.5040		
Sequence	98	S	+++	0.6050	0.4066	0.4985		

Delta -4

RQLRLRRPDDGSREAMNNLSDFQKDNLIPAAQLKNTNQKKELEVDCGLDKSNCGKQQNHTLDYNLAPGPLGRGTMPGKFP
HSDKSLGEKAPLRLHSEKPECRISAICSPRDSMYQSVCLISEERNECVIATEV

SeqName	Residue	O-GlcNAc result	Potential (o-glcnac)	Thresh. (1)	Thresh. (2)	NetPhos potential (Thresh=0.5)	YinOYang?
Sequence	112	S*	++	0.4971	0.3854	0.4699	0.991	*
Sequence	131	T	+	0.4372	0.3767	0.4582		

ACKNOWLEDGMENTS

This thesis was prepared at the Protein Structure and Bioinformatics Group at the International Centre for Genetic Engineering and Biotechnology (ICGEB) in Trieste, Italy, and was supported by an ICGEB master fellowship.

I would like to express my sincere thanks to my supervisor, Sándor Pongor and to my external supervisor, Martin Bishop, for their continuous support, help and advice throughout the time of my master

I am particularly grateful to Alessandro Pintar, who introduced me to the subject of Jagged and Delta families as well as for his daily guidance and help with my thesis.

The members of the Protein Structure and Bioinformatics group have created a friendly and dynamic atmosphere.

My dear friend Severina Tzankova and all other Italian and foreign friends here in Trieste, for being with me every single day, that helped me through the difficult periods.

Last but not least I wish to thank my parents for giving me their unlimited love and support.

REFERENCES

Artavanis-Tsakonas, S., Rand, M.D. and Lake, R.J. (1999) Notch signaling: cell fate control and signal integration in development. *Science*, **284**, 770-776.

Ascano, J.M., Beverly, L.J. and Capobianco, A.J. (2003) The C-terminal PDZ-ligand of JAGGED1 is essential for cellular transformation. *J. Biol. Chem.*, **278**, 8771-8779.

Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. and Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633-1649.

Choi, J.H., Jung, H.Y., Kim, H.S. and Cho, H.G. (2000) PhyloDraw: a phylogenetic tree drawing system. *Bioinformatics*, **16**, 1056-1058.

Dosztanyi, Z., Csizmok, V., Tompa, P. and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433-3434.

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M. and Obradovic, Z. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573-6582.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C. and Obradovic, Z. (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.*, **19**, 26-59.

Dyson, H.J. and Wright, P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.*, **6**, 197-208.

Ehebauer, M.T., Chirgadze, D.Y., Hayward, P., Martinez Arias, A. and Blundell, T.L. (2005) High-resolution crystal structure of the human Notch 1 ankyrin domain. *Biochem. J.*, **392**, 13-20.

Gridley, T. (2003) Notch signaling and inherited disease syndromes. *Hum. Mol. Genet.*, **12**, R9-13.

Haines, N. and Irvine, K.D. (2003) Glycosylation regulates Notch signalling. *Nat. Rev. Mol. Cell. Biol.*, **4**, 786-797.

Hambleton, S., Valeyev, N.V., Muranyi, A., Knott, V., Werner, J.M., McMichael, A.J., Handford, P.A. and Downing, A.K. (2004) Structural and functional properties of the human notch-1 ligand binding region. *Structure*, **12**, 2173-2183.

Harper, J.A., Yuan, J.S., Tan, J.B., Visan, I. and Guidos, C.J. (2003) Notch signaling in development and disease. *Clin. Genet.*, **64**, 461-472.

Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z. and Dunker, A.K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323**, 573-584.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z. and Dunker, A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037-1049.

Ikeuchi, T. and Sisodia, S.S. (2003) The Notch ligands, Delta1 and Jagged2, are substrates for presenilin-dependent "gamma-secretase" cleavage. *J. Biol. Chem.*, **278**, 7751-7754.

Julenius, K., Molgaard, A., Gupta, R. and Brunak, S. (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology*, **15**, 153-164.

Kadesch, T. (2004) Notch signaling: the demise of elegant simplicity. *Curr. Opin. Genet. Dev.*, **14**, 506-512.

Kanwar, R. and Fortini, M.E. (2004) Notch signaling: a different sort makes the cut. *Curr. Biol.*, **14**, R1043-1045.

Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499-520.

Kovall, R.A. and Hendrickson, W.A. (2004) Crystal structure of the nuclear effector of Notch signaling, CSL, bound to DNA. *Embo J.*, **23**, 3441-3451.

Lai, E.C. (2002) Protein degradation: four E3s for the notch pathway. *Curr. Biol.*, **12**, R74-78.

LaVoie, M.J. and Selkoe, D.J. (2003) The Notch ligands, Jagged and Delta, are sequentially processed by alpha-secretase and presenilin/gamma-secretase and release signaling fragments. *J. Biol. Chem.*, **278**, 34427-34437.

Le Borgne, R., Bardin, A. and Schweisguth, F. (2005) The roles of receptor and ligand endocytosis in regulating Notch signaling. *Development*, **132**, 1751-1762.

Le Borgne, R. and Schweisguth, F. (2003) Notch signaling: endocytosis makes delta signal better. *Curr. Biol.*, **13**, R273-275.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. and Bork, P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142-144.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003a) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453-1459.

Linding, R., Russell, R.B., Neduva, V. and Gibson, T.J. (2003b) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701-3708.

Lubman, O.Y., Kopan, R., Waksman, G. and Korolev, S. (2005) The crystal structure of a partial mouse Notch-1 ankyrin domain: repeats 4 through 7 preserve an ankyrin fold. *Protein Sci.*, **14**, 1274-1281.

Nair, R. and Rost, B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.*, **348**, 85-100.

Pear, W.S. and Aster, J.C. (2004) T cell acute lymphoblastic leukemia/lymphoma: a human cancer commonly associated with aberrant NOTCH1 signaling. *Curr. Opin. Hematol.*, **11**, 426-433.

Pfister, S., Przemeck, G.K., Gerber, J.K., Beckers, J., Adamski, J. and Hrabe de Angelis, M. (2003) Interaction of the MAGUK family member Acvrinp1 and the cytoplasmic domain of the Notch ligand Delta1. *J. Mol. Biol.*, **333**, 229-235.

Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M., Ausiello, G., Brannetti, B., Costantini, A., Ferre, F., Maselli, V., Via, A., Cesareni, G., Diella, F., Superti-Furga, G., Wyrwicz, L., Ramu, C., McGuigan, C., Gudavalli, R., Letunic, I., Bork, P., Rychlewski, L., Kuster, B., Helmer-Citterich, M., Hunter, W.N., Aasland, R. and Gibson, T.J. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625-3630.

Radtke, F. and Raj, K. (2003) The role of Notch in tumorigenesis: oncogene or tumour suppressor? *Nat. Rev. Cancer*, **3**, 756-767.

Rebay, I., Fleming, R.J., Fehon, R.G., Cherbas, L., Cherbas, P. and Artavanis-Tsakonas, S. (1991) Specific EGF repeats of Notch mediate interactions with Delta and Serrate: implications for Notch as a multifunctional receptor. *Cell*, **67**, 687-699.

Romero, P., Obradovic, Z. and Dunker, A.K. (2004) Natively disordered proteins: functions and predictions. *Appl Bioinformatics*, **3**, 105-113.

Screpanti, I., Bellavia, D., Campese, A.F., Frati, L. and Gulino, A. (2003) Notch, a unifying target in T-cell acute lymphoblastic leukemia? *Trends Mol Med*, **9**, 30-35.

Shimizu, K., Chiba, S., Kumano, K., Hosoya, N., Takahashi, T., Kanda, Y., Hamada, Y., Yazaki, Y. and Hirai, H. (1999) Mouse jagged1 physically interacts with notch2 and other notch receptors. Assessment by quantitative methods. *J. Biol. Chem.*, **274**, 32961-32969.

Six, E., Ndiaye, D., Laabi, Y., Brou, C., Gupta-Rossi, N., Israel, A. and Logeat, F. (2003) The Notch ligand Delta1 is sequentially cleaved by an ADAM protease and gamma-secretase. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 7638-7643.

Six, E.M., Ndiaye, D., Sauer, G., Laabi, Y., Athman, R., Cumano, A., Brou, C., Israel, A. and Logeat, F. (2004) The notch ligand Delta1 recruits Dlg1 at cell-cell contacts and regulates cell migration. *J. Biol. Chem.*, **279**, 55818-55826.

Tompa, P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527-533.

Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, **579**, 3346-3354.

Tompa, P., Szasz, C. and Buday, L. (2005) Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.*, **30**, 484-489.

Uversky, V.N. (2002a) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **11**, 739-756.

Uversky, V.N. (2002b) What does it mean to be natively unfolded? *Eur. J. Biochem.*, **269**, 2-12.

Vardar, D., North, C.L., Sanchez-Irizarry, C., Aster, J.C. and Blacklow, S.C. (2003) Nuclear magnetic resonance structure of a prototype Lin12-Notch repeat module from human Notch1. *Biochemistry*, **42**, 7061-7067.

Weinmaster, G. (2000) Notch signal transduction: a real rip and more. *Curr. Opin. Genet. Dev.*, **10**, 363-369.

Weng, A.P. and Aster, J.C. (2004) Multiple niches for Notch in cancer: context is everything. *Curr. Opin. Genet. Dev.*, **14**, 48-54.

Weng, A.P., Ferrando, A.A., Lee, W., Morris, J.P.t., Silverman, L.B., Sanchez-Irizarry, C., Blacklow, S.C., Look, A.T. and Aster, J.C. (2004) Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science*, **306**, 269-271.

Wright, G.J., Leslie, J.D., Ariza-McNaughton, L. and Lewis, J. (2004) Delta proteins and MAGI proteins: an interaction of Notch ligands with intracellular scaffolding molecules and its significance for zebrafish development. *Development*, **131**, 5659-5669.

Saitou, N. and M. Nei (**1987**) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406-425.