EVOLUTION OF THE TGF-BETA SUPERFAMILY WITH EMPHASIS ON NODAL

Yuan Shen

2013

Degree Sought: M. Phil. Computer-based Biology

SCHOOL OF LIFE SCIENCES

UNIVERSITY OF NOTTINGHAM

Supervisor: Chris Wade

Advisor: Matt Loose

Abstract

Nodal is a ligand of the TGF-beta superfamily. It has the function of determining the left-right axis and inducing the endoderm and mesoderm. Nodal signals can also act as morphogens. Although it has been detected for 20 years, the relationships between different species within Nodal are still unclear. The purpose of this study is to investigate the evolution of the TGF-beta gene with the main focus on Nodal. That is: (1) to determine the relationships within the Nodal family; (2) to examine whether Nodal is duplicated or not during evolution. To achieve this, whether Nodal is monophyletic or not and the relationship of Nodal with other ligands in the TGF-beta superfamily will be examined first. The phylogenetic trees to examine the relationships among the ligands are built under software PhyML with the Maximum Likelihood method. As a result, Nodal is monophyletic, but its neighbour ligand or ligand group is nonetheless uncertain. This study demonstrates that the fish sequences are all in the group in which the bird Nodal is located. Duplication of Nodal has occurred when vertebrates evolved from Urochordata. In addition, deletions have occurred in birds and mammals.

Acknowledgements

I would like to express my sincere thankfulness and grateful appreciation to my supervisor Chris Wade and my advisor Matt Loose for giving me valuable advice and great support throughout the research.

I would also like to say thank you to the participants of my colleges whose thoughts and insights allow me to obtain rich information and many help for this study.

And thanks for my friends that have gone back to their countries to give me a happy life during these years.

At last, I would like to thank my family for their support throughout my study in the United Kingdom.

Table of contents

List of Figures

List of Tables

CHAPTER 1 INTRODUCTION

In a novel by Louis Cha, *The Deer and the Cauldron*, the protagonist of the story, Wei Xiaobao, was forced to kill Duolong (who was the head of the imperial Praetorians but also a friend of Wei in the Qing Dynasty) to save his trapped rebel friends by stabbing a sharp dagger into Duolong's heart. However, without knowing who attacked him, Duolong escaped the call of Death because his heart was on the right side of his body!

Duolong's condition is called dextrocardia in medical science. There are two types of dextrocardia: isolated dextrocardia and dextrocardia situs inversus (Abbott and Meakins, 1915). Those individuals who have situs inversus will have their heart on the right, while their liver is on the left. Moreover, the position of their stomach is also changed. What makes some organs be set on the left side while some are on the right side? What is the mechanism of the asymmetry? These questions of general interest have long intrigued biologists and anatomists. With the development of molecular genetics, it has been recognized that a gene called *Nodal* plays an important role. This gene is a member of the transforming growth factor-beta superfamily (TGF-beta superfamily), a family of extracellular signalling molecules.

1.1 THE TGF-BETA SUPERFAMILY

1.1.1 General background

The TGF-beta superfamily is a large family of cell regulatory proteins that have sequence similarity. TGF-betas are produced by a variety of cells and are

composed of a large number of ligands, including TGF-beta1, TGF-beta2, TGF-beta3 and bone morphogenetic protein (BMP), etc. The first TGF-beta gene was cloned in 1985 (Derynck, et al. 1985). It was found that some TGF-beta genes exist in animals such as nematodes, flies, vertebrates, etc. Members of this superfamily have the function of controlling cellular processes such as growth regulation, embryo development, and tissue and immune system homeostasis. (Herpin, 2004) The TGF-beta superfamily is named from the first member found in this superfamily. The TGF-beta is named as a transforming growth factor because it can transform normal fibroblast phenotypes; that is to say, if an epidermal growth factor (EGF) exists, it can change fibroblast cell wall growth characteristics creating the ability to grow in agar (Serra & Chang, 2003). TGF-beta signalling is mainly known for its role in morphogenesis. In addition, it also plays an important role in dorsal-ventral patterning in both deuterostomes and protostomes (Pang, 2011).

TGF-beta superfamily ligands are cytokines. A Cytokine (CK) is a type of protein or small peptide that can transmit information between cells and has immune regulation functions. It is soluble with a small molecular weight, and is actively secreted by immune system cells and other cell types. It is the core factor of contact between immune system cells and other types of cells. Cytokines can change the characteristics of secretory cells. They also affect cellular processes through regulating specific cell membrane receptors (Zhang, 2008).

According to their major functions, cytokines can be grouped in different categories such as Interleukin (IL), Colony-stimulating factor (CSF), Interferon (IFN), Tumour necrosis factor (TNF), the transforming growth factor-beta superfamily (TGF-beta superfamily), Growth factor (GF) and the chemokine family (Zhang, 2008). Among the groups of cytokines, the TGF-beta superfamily is the one that this project focuses on.

1.1.2 The TGF-beta signalling pathway

The TGF-beta signalling pathway is the pathway that the ligands in the TGF-beta superfamily mainly follow, which was first identified over 30 years ago. It is a pathway where secreted proteins transform cells and tissues (Pang, 2011). During organ development, the TGF-beta family is required for dorso-ventral patterning, mesoderm induction and patterning, limb bud formation, bone and cartilage formation, neuron differentiation and the development of a variety of different tissues and organs. Ligands of the TGF-beta superfamily produce dimers that bind to heterodimeric receptor complexes composed of type I and type II receptor subunits having serine/threonine kinase domains. After the ligands are bound, the type II receptor phosphorylates and activates the type I receptor to create a Smad-dependent signalling cascade that induces or represses transcriptional activity. This pathway evolved in the early evolution of metazoans (Pang, 2011).

The TGF-beta superfamily signalling pathway includes TGF-beta superfamily ligands, receptors and SMADs (Herpin, 2004). A ligand is a kind of biomolecule that has its own bioactivity and is able to bind to a biomolecule

(which is called a receptor) and form a complex with it to express a specific biological effect. A ligand can be a peptide or other small molecules, such as a neurotransmitter, a hormone, a pharmaceutical drug or a toxin. After binding to a receptor, a ligand will cause the change of cell interstitials to let signalling factors pass between cells and amplify the signalling (Zhang, 2008). This project focuses on the ligands of the TGF-beta superfamily which interact with serine/threonine-specific protein kinase receptors and SMADs.

A receptor is a kind of biomolecule that is located in the plasma membrane or the cytoplasm of a cell, and is attachable to one or more specific kinds of biomolecules (the singular of which is called a ligand) (Zhang, 2008). Usually a cell has many different kinds of receptors. There are a limited number of receptors in the body. If the number of ligands that occupy the available receptors has reached the maximum number, no matter how many ligands are further added, the number of the ligands that are affected with receptors will not change. Each kind of receptor can only bind certain ligand shapes. After forming a complex, the ligand and receptor can dissociate from each other. The structure of ligands and receptors will not be changed after binding and dissociation. According to where the receptor is located, it can be divided into three categories. One is the transmembrane receptor which is embedded in the plasma membrane, such as cholinergic receptors, adrenergic receptors and insulin receptors. The second, called the cytosolic receptor, is in the cytoplasm, such as hormone receptors and glucocorticoid receptors. The third, whose name is the nuclear receptor, is located in the nucleus, e.g. thyroid hormone receptors (Zhang, 2008).

Upon ligand binding, the receptor passes the signal through downstream substrates, which are called signalling molecules, to the effector proteins such as transcription factors or other functional proteins. SMAD is a kind of biomolecule that acts as an intracellular signalling molecule and is able to regulate the activity of ligands in the TGF-beta superfamily (Heldin et al. 1997; Derynck et al. 1998). After being activated by a ligand-bound receptor, a SMAD often forms a complex with other SMADs/CoSMAD, then translocation into the nucleus occurs and it acts as a transcription factor to regulate the expression of target genes (Dijke and Arthur, 2007; Massagué et al. 2005).

There are three kinds of SMAD: the receptor-regulated Smads (R-SMAD), the common-mediator Smads (co-SMAD) and the inhibitory Smads (I-SMAD, which are also called antagonistic Smads). R-SMAD includes SMAD1, SMAD2, SMAD3, SMAD5 and SMAD8/9. SMAD2 and SMAD3 are effectors for TGF-beta or Activin signals. SMAD1, SMAD5 and SMAD8 are effectors for BMP signals (Wu et al. 2001). Co-SMAD only includes SMAD4. It binds to activated R-SMADs and forms a complex to accumulate in the nucleus and regulate the expression of target genes (Shi et al. 1997) I-SMAD including SMAD6 and SMAD7. They act as inhibitors of R-SMADs and Co-SMADs by competing with SMAD4 to bind to R-SMADs. By so doing, I-SMADs can block the activation of R-SMADs and co-SMADs (Itoh et al. 2001).

As shown in Figure 1.1, TGF-beta superfamily signalling is initiated when the ligands bind to cell surface receptor serine/threonine kinases (type II and type I receptors). First, the ligands bind to a type II receptor. Then the type II receptor recruits and phosphorylates a type I receptor to make it activated. After that, the type I receptor then phosphorylates and activates receptor-regulated SMADs (R-SMADs). The Phosphorylated R-SMADs form complexes with the coSMAD (e.g. SMAD4). Next, the complexes accumulate in the nucleus. Finally, the complexes act as transcription factors and cooperate with transcription factors, co-activators and co-repressors to regulate the target gene expression. Inhibitor molecules can work at every stage of the signalling pathway. If Smads entered the nucleus, the specific transcriptional co-repressors would prevent the response to TGF-beta (Powers et al. 2010).

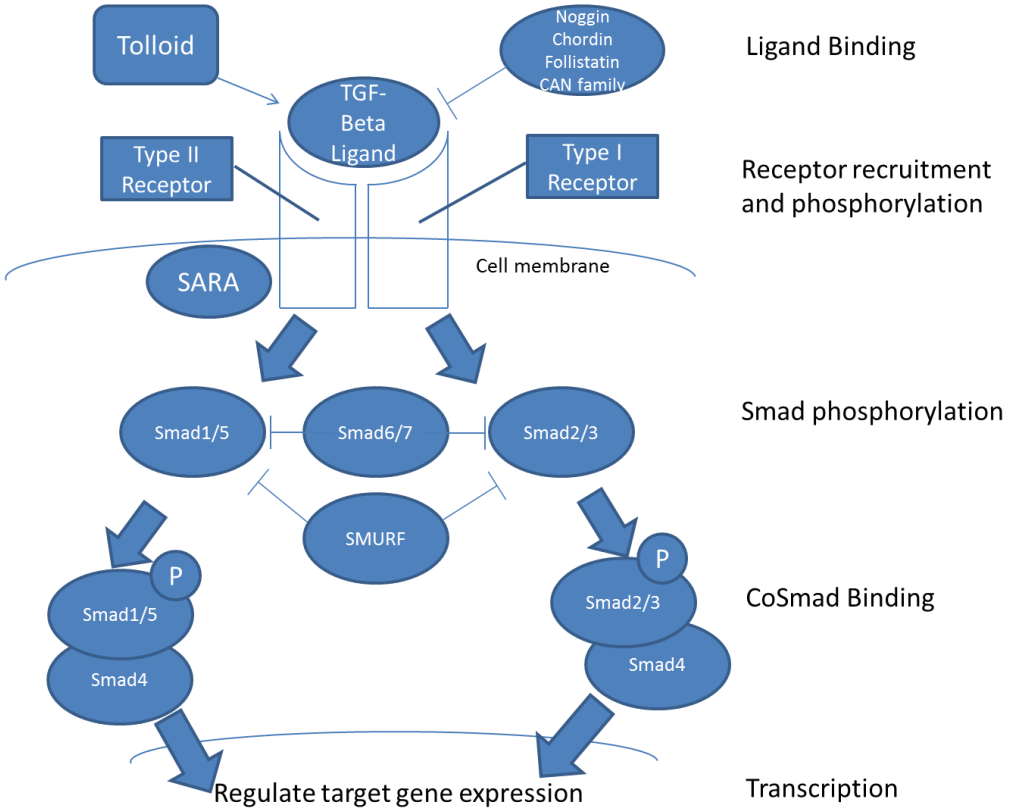

Figure 1.1 Overview of the TGF-beta signalling pathway.
Signalling is initiated by the binding of the Type II receptor and ligand. Activation of the receptor-Smad (Smad2/3, Smad1/5) is triggered by the sequestering of Type I receptors. This complex, in combination with Co-Smad, (Smad4) activates a transcription of target genes after entering the nucleus. The intercellular or extracellular antagonists can inhibit the pathway through SMURF ubiquitin ligase or Inhibitor-Smad (Smad6/7).

The TGF-beta precursor protein is divided into three main distinct regions, namely the signal peptide, the propeptide or latency associated peptide and the mature peptide. Each region has different functions; for example, the signal peptide is responsible for targeting TGF-beta to the endoplasmic reticulum and secretion. Essentially, the mature peptide is cleaved from the precursor protein and is responsible for signal transduction. Unlike the propeptide, the mature peptide is conserved across different families. The mature peptide is mainly cleaved by Furin, which is a convertase, at a dibasic arginine-X-X-arginine site (RXXR). The Homodimer or heterodimer is formed by an active peptide and

binds to a specific TGF-beta Type II receptor. Then, the TGF-beta Type I receptor is recruited by the TGF-beta Type II receptor, wherein its phosphorylated sites are activated by threonine/serine kinase. Following this, phosphorylated TGF-beta Type I receptors phosphorylate and activate receptor-associated Smad proteins (R-Smads), Smad2/3, and Smad1/5 (Pang, 2011). R-Smad proteins are divided into two major functional domains, namely Mad-homology domains 1 and 2 (MH1 and MH2). TGF-beta-like signalling is associated with Smad2/3, while BMP-like signalling is primarily associated with Smad1/5. Membranes are associated with inactive R-Smads through a Smad anchor for the receptor activation (SARA) protein. The Smad anchor for receptor activation contains the FYVE domain, which is a zinc finger domain. After activation, R-Smads are released into the cytosol for interaction with the common-mediator Smad (Smad4 aka Co-Smad). It is later translocated into the nucleus. TGF-beta target genes are thereafter regulated by the heteromeric complex through interaction with transcription factors, including Myc, Fos/Jun or co-activators such as Creb-binding protein (CBP). The MH1 domain can interact with DNA while the MH2 domain can interact with Type I receptors. The target gene is also involved in the protein-protein interactions, for instance Co-Smad/R-Smad binding (Derynck & Zhang, 2003).

TGF-beta signalling inhibition can occur at different levels, for instance in the nucleus, cytoplasm and extracellular matrix. Receptors binding with ligands are impaired by the extracellular diffusible antagonists, due to the fact that they act as ligand traps, for example Follistatin, Noggin, Chordin, and the CAN family (Gremlin/DAN/Cerberus). Thereafter, zinc metalloprotease Tolloid is

activated to cleave to Chordin, and in so doing releases BMPs. This process shows that there are numerous regulation levels of TGF-beta signalling. Apart from cleaving Chordin, Tolloid also cleaves pro-collagens of the extracellular matrix and other proteoglycans. Furthermore, some Tolloid is also involved in the binding of TGF-beta ligands (Pang, 2011). SMURF may also degrade Type I receptors after being recruited by I-Smads in the membrane. TGF-beta signalling can also be regulated in the nucleus as when co-repressors Sno/Ski bind (Liu, et al. 2001). These proteins can recruit repressors to block TGF-beta target gene activation.

In a cell, the TGF-beta signalling pathway can also be inhibited at different levels. For instance, at the receptor level the GS domain binding with Type I receptor phosphorylation can be blocked by FKBP12 (Chen, et al. 1997). A second example is the formation of a receptor complex, which is caused after Type II and Type I receptors bind. In this example, a pseudo receptor, BAMBI, may prevent Type I and Type II receptors from binding (Onichtchouk, et al. 1999). Furthermore, Inhibitor-Smads (Smad6/7, I-Smad) can also cause pathway modulation because they have an MH2 domain, and can bind with Type I receptors to prevent phosphorylation and binding of R-Smad. Co-Smads binding with R-Smad can also be hindered due to competition from I-Smads binding with Co-Smads. TGF-beta signalling can also be regulated by the E3 ubiquitin ligase, SMURF, which targets R-Smads for degradation (Zhu, et al. 1999).

1.1.3 Characteristics of the TGF-beta superfamily ligand sequences

Whether a protein is a family member or not is determined by the presence of the RXXR cleavage site, and the 7 cysteine residues in the mature domain. All the TGF-beta superfamily ligands have a dibasic or RXXR cleavage site. The pro-domain before the cleavage site of TGF-beta is poorly conserved across different family members, although it is well conserved within a particular family member from a different species. The mature domain is more highly conserved than the pro-domain. It contains most of the sequence landmarks. In the mature region, there are 7 cysteine residues that are highly-conserved and hardly changed through all the family members (Figure 1.2). The Cysteine site is missing in GDF-3 and GDF-9.

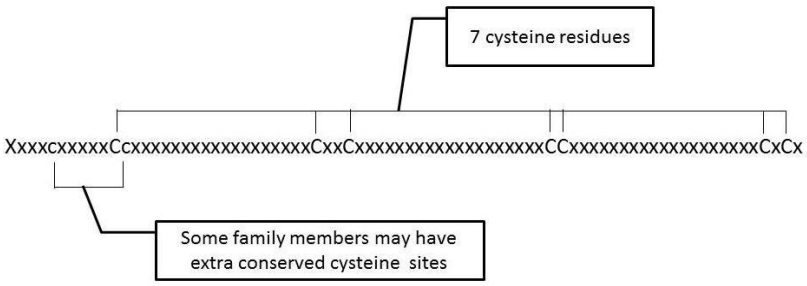

Figure 1.2 The 7 conserved cysteine residues in the TGF-beta superfamily

1.1.4 Four groups of subfamilies in the TGF-beta superfamily

There are dozens of families belonging to the TGF-beta superfamily, which can be divided into two major classes: a protein-like bone morphogenetic class (BMP class) and a TGF-beta-like class. The former includes the following members: Bmp5-8, Bmp2/4/Dpp, Gdf2, Bmp3, ADMP, Nodal, Univin/Vg1 and Gdf5-7, whereas the latter includes lefty, TGF-beta sensu stricto, inhibin/activin and Gdf8/Myostatin (Pang, 2011).

In Herpin's review, the author gave a general review of the TGF-beta superfamily. He introduced the ligands, SMADs and reporters of the TGF-beta superfamily. In the ligand section, he grouped the TGF-beta superfamily into 4 groups according to ligand functions. Figure 1.3 illustrates the phylogenetic relationships of the TGF-beta superfamily in Herpin's review (Herpin et al. 2004).

As shown in Figure 1.3, when grouped by functions, the ligands of the TGF-beta superfamily can be divided into four major subfamilies: (1) The decapentaplegic-Vg-related (DVR) related subfamily – also known as the BMP subfamily. (2) The activin/inhibin subfamily. (3) The TGF-beta sensu stricto and related factor subfamily. (4) A group of various divergent members. (Herpin et al. 2004).

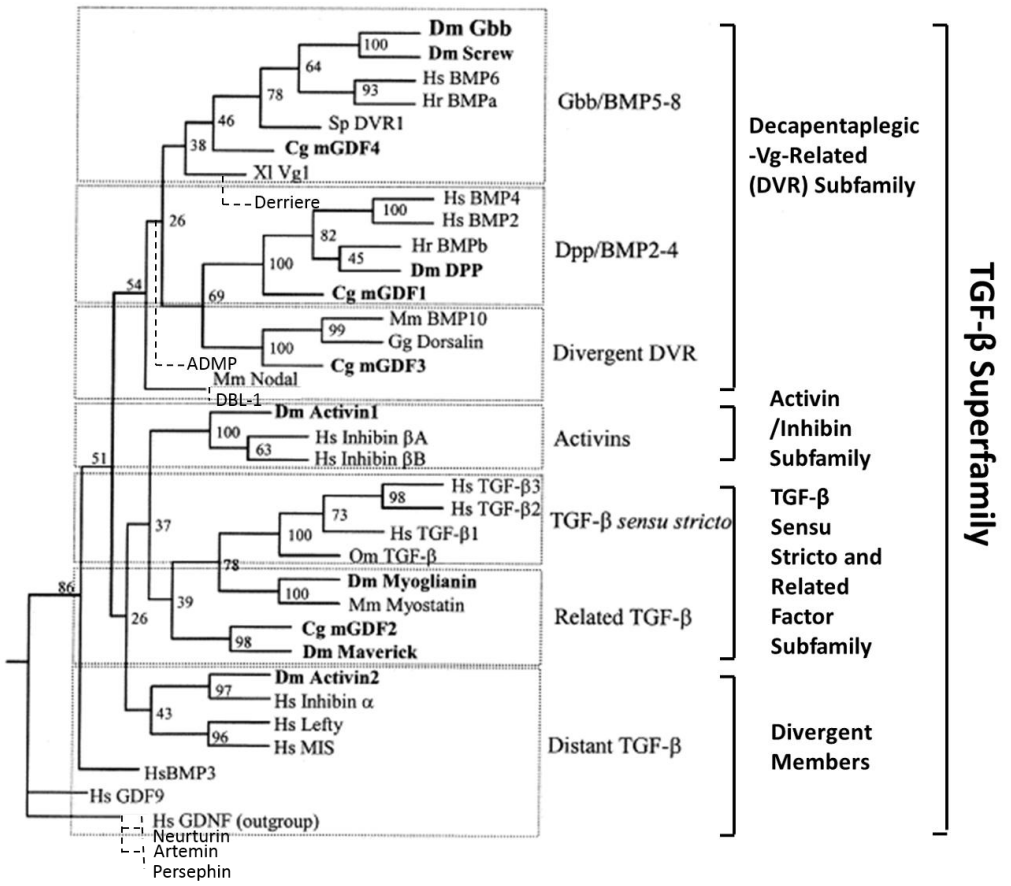

Figure 1.3 The Phylogenetic relationships between the TGF-beta superfamily of ligands.

The 4 square brackets within the TGF-beta superfamily in the diagram represent the four major distinct ligand subfamilies of the TGF-beta superfamily. The first group is the DVR subfamily, which includes GBB/BMP5-8, DPP/BMP2/4 and Divergent DVR. The second group is the activin/inhibin subfamily. The third group is TGF-beta sensu stricto and related factor subfamily, which includes the TGF-beta sensu stricto and related TGF-beta ligands. The last group is a group representing various divergent members in the superfamily which illustrates distant TGF-beta. In this diagram, the numbers at each branch node represent the percentage values given by bootstrap analysis. Protostome sequences are indicated in bold. The GDNF (Glial Derived Neurotrophic Factor) is used as an out group. The tree is based on 120 amino acids. In this figure, the branches drown with dashed lines show a list of TGF-Beta superfamily ligands not included in Herpin's phylogeny but included in other researchers' assumptions. Derriere, ADMP and DBL-1 are said to be in DVR subfamily, Neurturin, Artemin and Persephin are said to be with GDNF.

(1) The decapentaplegic-Vg-related (DVR) subfamily

This subfamily comprises growth and differentiation factors (GDFs) which consist mainly of GDF3, GDF4 and GDF1, Nodal, Gbb, Dpp, Dorsalin, Decapentaplegic-Vg-related (DVR), Screw and most of the bone morphogenetic proteins (BMPs). Derriere, ADMP and DBL-1 are also in this subfamily, but they are not included inHerpin's theory. Among the ligands listed above, Nodal is the ligand that the present project focuses on.

(2) The activin/inhibin subfamily

The activin subfamily includes Activins and Inhibins. There are two kinds of Activin sub-units: sub-unit ßA and sub-unit ßB. Depending on their sub-unit, there are three types of activins: Activin A (composed of ßA ßA), Activin B (composed of ßB ßB) and Activin AB (composed of ßA ßB) (van Zonneveld et al. 2003). Both activins and inhibins are para/autocrine regulators of cell function (Chen et al. 2006).

(3) The TGF-beta sensu stricto and related factor subfamily

The TGF-beta sensu stricto includes TGFB 1-5, whereas the TGF-beta related factor includes the Maverick (Mav), GDF2, Myoglianin and Myostatin. TGFB is involved in embryogenesis, cell differentiation, extracellular matrix neogenesis, immunosuppression, apoptosis as well as other processes (Nguyen et al. 2000).

(4) A group of various divergent members

Proteins included in this group are less similar to other members in the TGF-beta superfamily, but bear the typical architecture of the ligands. In Figure 1.3, the divergent members include the Anti-Müllerian Hormone (AMH, or Müllerian Inhibiting Substance, MIS), Lefty, Daf7, Unc-129 and GDNF (Glial cell-derived neurotrophic factor). It is additionally shown in Figure 1.3 that some activins, inhibins, BMPs and GDFs also fall into this group.

Neurturin (NRTN, NTN), Artemin (ARTN, Enovin) and Persephin (PSPN, PSP) are also divergent members of the TGF-beta superfamily, but they are not included in Herpin's review. GDNF together with NRTN, ARTN and PSPN belong to the GDNF family of ligands (GFL). GFLs affect internal cell survival, neurite outgrowth, cell differentiation and cell migration. The members of the GDNF family belong to the TGF-beta superfamily, but the amino-acid sequence homology is less than 20% of GDNF family members with other members of the TGF-beta superfamily (between the members of the GDNF family, the amino-acid sequence homology is between 40 and 50%) (Airaksinen et al. 2002; Saarma, 2000).

1.2 NODAL

1.2.1 General background

Nodal is the ligand that this project focuses on in the TGF-beta superfamily. It plays an important role in the formation of the left-right axis in the development of vertebrates. It is additionally essential to the formation of the mesoderm and anterior-posterior axis. Nodal is first found expressed in the

node (the organizer for gastrulation in vertebrates), so this gene was named Nodal (Garcia-Fernàndez , et al. 2007; Zhou, et al. 1993). Nodal is primarily found in chordates, but not in ecdysozoa, for example, the nematode or fruit fly. It has also been proved that it is found in deuterostomes such as sea urchins and Chordates and the protostome group, such as Lophotrochozoa. Nodal protein consists of a mature ligand domain and prodomain, and is translated as proproteins (Schier, 2009; Bianco, et al. 2010). Nodal signals are part of the TGF-beta superfamily, and are essential for the determination of the left-right axis and induction of the endoderm and mesoderm. Nodal signals can also act as morphogens because they have concentration-dependent effects and are able to act at a distance from the production source (Schier, 2009). Nodal regulates FoxH1 gene expression and induces the transcription of mRNAs that are involved in cell differentiation, left and right axis specification and mesoderm and endoderm induction(Hamada, et al. 2002). In most species, Nodal gene expresses on the left side of the body in the lateral plate mesoderm and brain region (Ito, et al. 2006).

1.2.2 Nodal signalling pathway

Figure 1.4 shows the Nodal signalling pathway. Nodal ligands, as with other TGF-beta signals, activate threonine/serine kinase receptors which are responsible for the phosphorylation of Smad proteins. Nodal signals are mainly received by EGF-CFC co-receptors and Type II and I Activin receptors. The activation of receptors is followed by the phosphorylation of transcription factors Smad3 and Smad2. This further leads to the binding to the nuclear translocation factor, Smad4, and association with more transcription factors

that regulate target genes. This core pathway is mainly regulated by antagonists that process enzymes and extracellular proteins. Furthermore, Nodal signalling is also regulated by miRNAs. These are responsible for receptor trafficking and intracellular molecules, for example transcriptional cofactors. A more in-depth understanding of Nodal signal transduction's molecular basis enhances the understanding of regulation of Nodal morphogen activity (Schier, 2009).

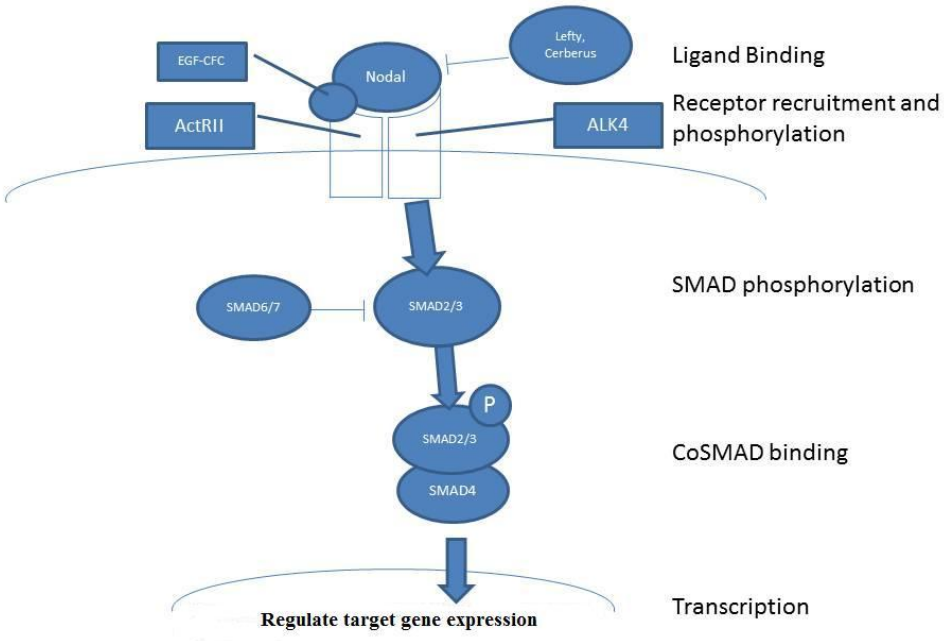

Figure 1.4 Nodal signalling pathway.
After the convertases processes Nodal precursor, Nodal transfers signals via EGF-CFC co-receptors and activin receptors. Lefty and Cerberus mainly act as the extracellular inhibitors. Lefty mRNAs and Nodal are targeted by MicroRNAs that belong to the miR-430 family, and they are responsible for repression and degradation. The Type II activin receptor is repressed by Mir-15/16. Activin receptors are recycled by Rap2, while activin receptor complexes are targeted by Dapper 2 in the lysosome for degradation. Activation of the pathway is mediated by Smad2 phosphorylation and Smad4

association with Smad2 and Mixer, p53 and FoxH1 transcription. On the other hand, the PPM1A dephosphorylated Phospho-Smad4 is later exported by RanBP3 from the nucleus. Deubiquitinase FAM/Usp9x and ubiquitinase Eactodermin regulate the stability and activity of Smad4.

Nodal signals are assembled by receptor complexes, and they consist of both type II and type I activin receptors (ActRIB; ActRIIA/B), which function as serine/threonine kinases (Schier, 2009). EGF-CFC proteins are linked to GPI factors, which are required for Nodal signalling and embryogenesis. For example, an absence in the EGF-CFC protein in one-eyed pinheads renders an embryo resistant to Nodals and inactivates the pathway. Moreover, it is thought that EGF-CFC proteins serve as co-receptors by binding type I activin receptors and Nodals. Recent tissue culture studies have highlighted the need for ligands acting in conjunction with receptor trafficking in Nodal signalling. For example, the mammalian EGF-CFC protein Cripto may be used effectively to promote Nodal signalling through linking the processing and trafficking of Nodal. Cripto can be used to form a complex in conjunction with convertases and Nodal precursors on the surface of cells that will respond by facilitating Nodal and translocation to early endosomes and processing (Schier, 2009).

1.2.3 Extracellular antagonists, convertases and Nodal signals

A model developed by Serra and Chang to show how Nodal and lefty affect left-right patterning is shown in Figure 1.5. Most Nodals express on the left side because of the regulation of other ligands and inhibitors. On the left side, Nodal is regulated by Vg1/GDF1 and early BMP and is inhibited by BMP and Lefty. Oversecretion of Nodal will lead to the expression of Lefty on the left

side to downregulate Nodal. There are Lefties expressed on the midline. There, the Lefties act as a midline barrier to stop Nodal moving into the right side of the body. If Nodal appears on the right side, both the Activin/ActRIIA and BMP/ALK2 can stop it (Serra & Chang, 2003).

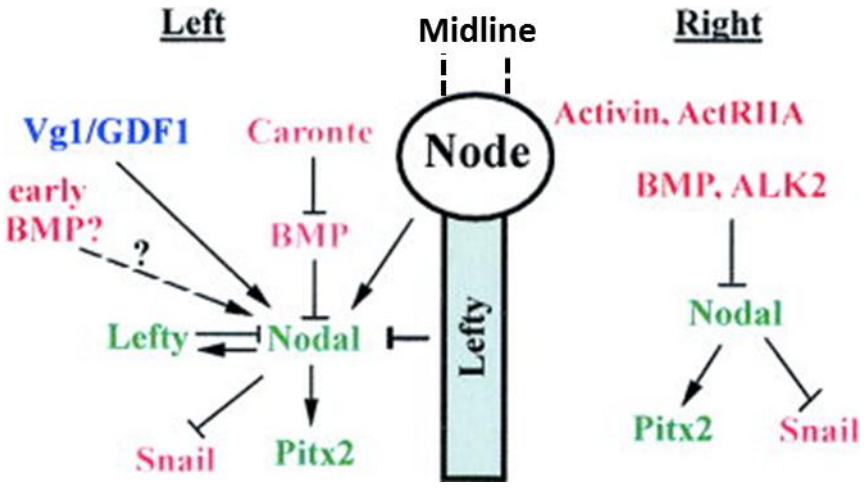

Figure 1.5 Left-right patterning model by lefty and Nodal.
On the left, an early signal from the node causes the expression of Nodal and Lefty in the left lateral plate mesoderm. Vg1/GDF1 is expressed on both sides, but it can only be activated early on the left to regulate Nodal expression. Caronte is a BMP inhibitor. It is also expressed on the left side and antagonizes the function of BMP of inhibiting Nodal. Downstream of Nodal signalling pathway on the left side, transcription factor Pitx2 is turned on and Snail is inhibited. Midline expression of lefty is necessary to prevent Nodal going into the right side. On the right side, BMP signals through ALK2 and Activin through Activin A type II receptors will inhibit Nodal signalling. Downstream of Nodal signalling pathway on the right side, Snail is activated and Pitx2 is shut off.

1.2.4 Number of copies of Nodal in different species

The number of Nodal genes in different species is varied. Nodal paralogs are described as “Nodal-related” in the zebra fish, frog, Japanese newt and Japanese killifish. Mice and humans possess only one Nodal gene, but the zebra fish has three Nodal paralogs: squint, cyclops and southpaw (SPAW). In

the African clawed frog, there are six Nodal genes, known as xnr (*Xenopus laevis* Nodal-related) 1~6 (Swiers, 2010). And in the Western clawed frog, 2 kinds of xtnr (*Xenopus (Silurana) tropicalis* Nodal-related), Xtnr1 and Xtnr3(which has three forms: 3-A, 3-B, and 3-C) were discovered (Haramoto, et al. 2004; Klein, et al. 2002). In Japanese killifish (also known as the Medaka or Japanese rice fish), there are two: onr (*Oryzias latipes* Nodal-related) 1 and 2. (Soroldoni, et al. 2007). In the Japanese fire belly newt, Nodal-related gene is called CyNodal (*Cynops pyrrhogaster* Nodal) (Ito, et al. 2006).

Nodal homologs in different species are very similar in terms of their amino acid sequence structure, yet they have different effects. In the zebra fish, Squint and Cyclops are important for mesendoderm formation, while SPAW plays a vital role in asymmetric heart morphogenesis and visceral left-right asymmetry (Baker, et al. 2008). In the frog, Xnr1, Xnr2 and Xnr4 have mesoderm induction activity. Xnr3 cannot induce mesoderm, but Xnr3 has neural induction activity (Takahashi, et al. 2000).

In the support material for The Genome of the Western Clawed Frog *Xenopus tropicalis* (Hellsten, 2010), the author indicates that there are two Nodal loci in vertebrates. One is between eif4ebp2 and ash2l, and the other is between eif4ebp1 and paladin. In some species such as frogs or fish, the Nodal gene may be amplified and show several copies. In some species such as mammals or birds, one of the loci may be deleted. The bird loses the Nodal locus adjacent to paladin, while the mammal loses the Nodal locus adjacent to ash2l.

Other transcription factors that relate to mesoderm and endoderm development also have multiple copies.

1.3 AIMS AND OBJECTIVES

The purpose of this study is to investigate the evolution of the TGF-beta superfamily with the main focus on the Nodal gene. That is: (1) to determine the relationship within the Nodal family; (2) to examine whether Nodal is duplicated during evolution. To achieve this, whether Nodal is monophyletic and the relationship of Nodal with other ligands in the TGF-beta superfamily will be examined first.

CHAPTER 2 METHODOLOGY

Summarized in this chapter are general methodologies that are referred to in the succeeding chapters. A brief description, along with some basic concepts, will be shown in this chapter. Sequences used in this project were downloaded from GenBank and Ensembl and aligned within the Genetic Data Environment 2.4 Macintosh Edition (MacGDE) (Smith et al. 1994). The sequences were then checked for saturation before being subjected to phylogenetic estimation. To this end, the optimal model that best fitted the dataset is first identified, and then a phylogenetic tree is constructed by using that model with the Maximum-likelihood method.

2.1 ASSEMBLING A DATASET

The DNA sequences used in this analysis were obtained from GenBank and Ensembl through a detailed search of every member of ligands of the TGF-beta superfamily. The analysis tried to include as many Nodal sequences as possible. The DNA sequences were translated into amino acid sequences to provide protein information for building amino acid trees.

Ensembl is a joint project between the European Molecular Biology Laboratory (EMBL), the European Bioinformatics Institute (EBI) and the Welcome Trust Sanger Institute (WTSI). The aim of this joint project is to automatically annotate the selected eukaryotic genomes and to maintain and provide the information in the form of an on-line database (Flicek, 2011).

GenBank is a general genetic sequence database run by the National Center for Biotechnology Information (NCBI), which collects genes from all publicly available DNA sequences. It is a commonly used on-line gene database (Benson, et al. 2009).

Nodal sequences from GenBank were identified by reviewing the literature to ascertain whether they were proven by experiments or only by BLAST searching. It should be noted that as Nodal sequences from Ensembl were automatically annotated with high confidence and no literature information was provided, in this analysis Nodal sequences from Ensembl were not manually checked.

2.2 MULTIPLE SEQUENCE ALIGNMENT

The dataset was aligned through a combination of automatic and manual methods. The on-line MUSCLE service on the EBI website was used to automatically align the dataset. Based on the results of the automatic alignment, the manual alignment was done through the program Genetic Data Environment 2.4 Macintosh Edition (MacGDE). After the alignment, marker files were made to inform which sites were unambiguously aligned that could therefore be used in building the phylogenetic trees.

Multiple Sequence Comparison by Log-Expectation (MUSCLE) is a multiple alignment program for both amino acid and DNA sequences, which is more accurate and efficient than Clustal and T-Coffee (Edgar, 2004). MacGDE is a

multiple phylogeny platform for alignment and phylogenetic analysis which can read a wide range of file formats (Smith et al. 1994).

2.3 CHOICE OF THE DATASETS USED

After the alignment, a dataset that will be brought into phylogenetic analysis needs to be chosen. It needs to contain sufficient sites to build a tree, and needs to contain enough ligands from different subfamilies to show the relationships within the TGF-beta superfamily. Then, the dataset will be brought into a saturation test and further phylogenetic analysis.

2.4 SATURATION TEST

After the dataset was chosen, a saturation test is taken to test the accuracy of the results. Saturation is caused by multiple changes at one site in the alignment (Farrell, 2011). Testing for saturation can be done in different ways, for instance transition distance vs. transversion distance, and transition and transversion distance vs. uncorrected distance, among others (Morisson, 2006; Tsigenopolous et al. 2002).

A transition (ti) refers to the change of a purine nucleotide to another purine (A↔G) or pyrimidine nucleotide to another pyrimidine (C↔T). A transversion (tv) is a nucleotide-pair substitution type that involves a purine replacement with a pyrimidine, or a pyrimidine replacement with a purine. (Collins & Jukes, 1994). Transition (ti) occurs more frequently than transversion (tv).

There are different ways to determine whether a dataset is saturated. There are two tests used in this project: the transition (ti) and transversion (tv) distance plotted against uncorrected distance, and transition (ti) distance plotted against transversion (tv) distance. These methods use different ways to determine the saturation of a dataset.

In the test with the Transition (ti) and transversion (tv) distance vs. the uncorrected distance method, if there were no saturation, there would be two straight lines, as shown in Figure 2.1(a). Due to the fact that transition is more frequent than transversion, the line of the transition will be higher than that of transversion in the saturation test of transition (ti) and transversion (tv) distances against pairwise total uncorrected distances. This is because of the following points: firstly, the saturation is caused by the multiple changes at one site in the database; secondly, transition happens more frequently than transversion. That means for one site, transition is more likely to happen than tranversion. Thus, transition experiences saturation more easily than transversion. When saturation occurs, the transition line is usually a curve in the diagram, while transversion is depicted as a straight line (Figure 2.1 (b)). As a curve is equated with saturation in this dataset, the result based on the dataset may not be accurate. The earlier the ti line crosses the tv line, the more saturation there will be. (Morisson, 2006)

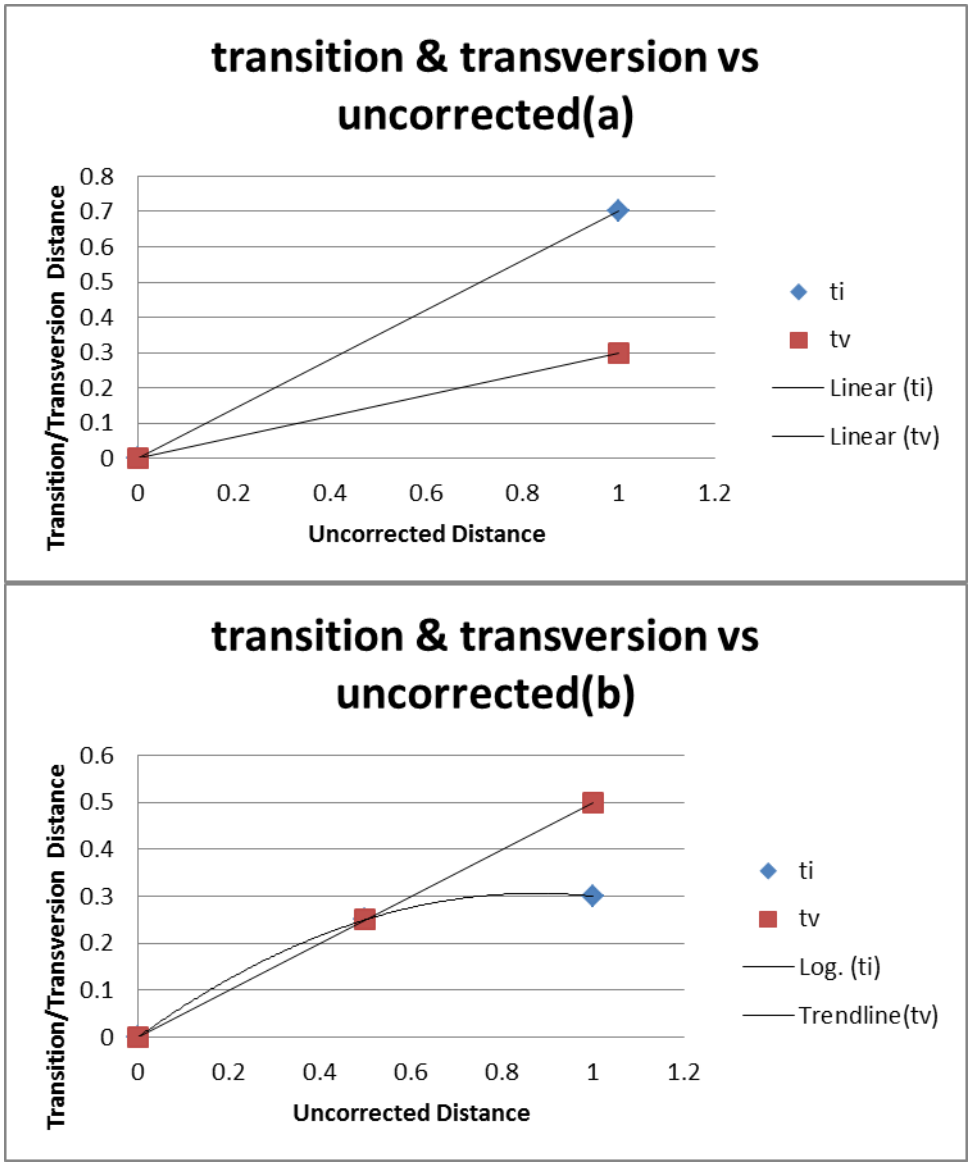

Figure 2.1 Transition and transversion vs. uncorrected distance method.
The diagram illustrates the saturation test results obtained from an uncorrected pairwise transition (ti) and transversion (tv) distance against the total uncorrected distance. (a)The diagram above shows a straight line for both transition and transversion. This indicates that no saturation is observed in the dataset. Both transition and transversion are not saturated, reflecting accuracy in the results. (b) The diagram clearly shows a straight line for transversion, whereas a curve for transition. This indicates that the dataset is saturated. In the case that the dataset is saturated, it can be interpreted that the results may be inaccurate.

In the test with transition (ti) distance vs. transversion (tv) distance model, with the points on the midline y=x it can be interpreted that transversion (tv) and

transition (ti) are equal. It can also be observed that if one point is above the line, transversion (tv) distance is shorter than the transition (ti) distance. The points are most likely to appear above the y=x line because transition (ti) occurs more frequently than transversion (tv). Generally, saturation is observed in the dataset in the case that there are multiple points under the y=x line, which may give inaccurate results (Tsigenopolous et al. 2002).

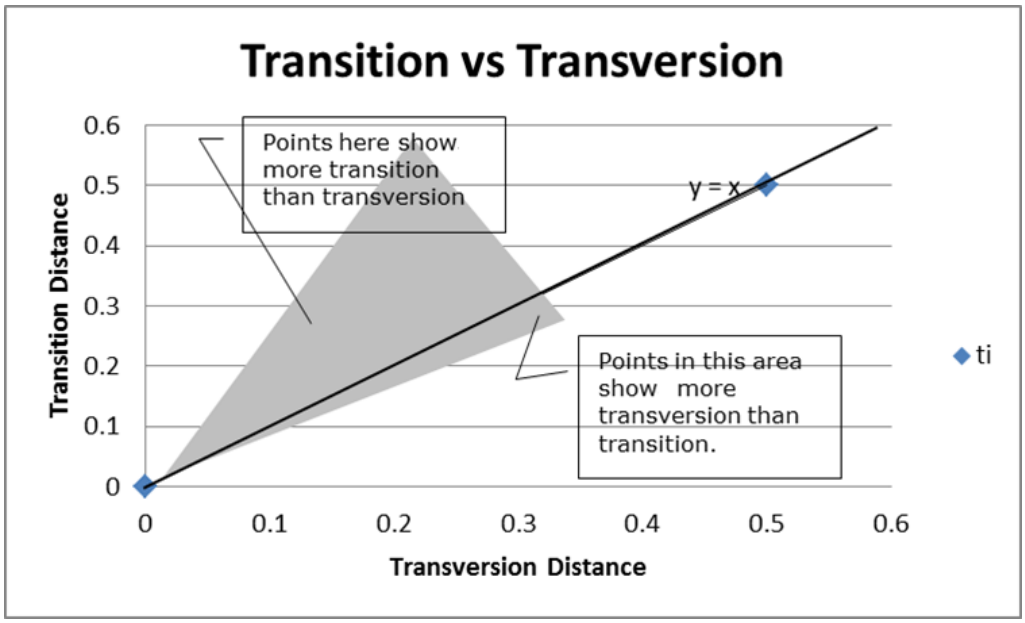

Figure 2.2 Transition (ti) distance vs. transversion (tv) distance model.
The above diagram illustrates anticipated results from a saturation test by using a transition (ti) distance vs. Transversion (tv) distance model. There are different observations expected. For instance, all points along the midline y=x suggest that transversion (ti) and transition (ti) are equal. One point above the midline y=x suggests that transversion (tv) distance is shorter than the transition (ti) distance. Furthermore, most points above the midline y=x indicate that transversions are less frequent than transitions. In this case, the points are most likely to appear above the midline y=x, which shows that transitions are greater than transversions, and it can be interpreted that the dataset is not saturated. On the other hand, if most of the points are below the midline y=x, it can be interpreted that the dataset is saturated.

During the saturation test, Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0 Beta (PAUP) (Swofford, 1998) was used to calculate the

uncorrected distances and the transversion/transition distances. Then, the diagrams are generated from the distance data by using Microsoft Excel.

2.5 PHYLOGENY RECONSTRUCTION

After conducting the saturation test, the datasets chosen are used to build phylogenetic trees. Neighbor-joining (NJ) is a distance method that works quickly. In this method, the evolutionary distance is calculated between sequences, the distance data is collated to form a distance matrix and then a tree is drawn from the matrix. It assumes that the distances are additive, but does not require the data to be ultrametric (Saitou and Nei, 1987). With this method, a general view of the tree could be shown quickly, but the result may not be as accurate as the character - state methods. Maximum Likelihood (ML) is a character - state method that considers the probability of each nucleotide changing in sequence alignment in each group. Then, the tree that gets the largest sum of the probability is that which most likely reflects the true situation of the phylogenetic tree (Cavalli-Sforza and Edwards, 1967; Felsenstein, 2004). It has more statistical flexibility than Maximum Parsimony (MP). In addition, compared to the Bayesian Inference, ML can provide satisfactory, accurate results although will take a longer time. Thus, ML is chosen to be the method used in this project.

2.6 CALCULATION OF DISTANCES

Uncorrected p distances, in which distances were calculated as the number of substitutions divided by sequence length with no correction for multiple substitutions, were calculated in PAUP.

Corrected distances in which distances were corrected to account for multiple hits at the same site leading to an underestimate of the actual amount of change, were calculated using GTR models in PhyML. The GTR model is short for the general time-reversible model. It requires 6 substitution rate parameters and assumes that all six pairs of substitutions have different rates and the base frequencies are not equal. But it considers that all nucleotide sites are equally likely to change, all nucleotide sites change independently and the base composition is at equilibrium among all sequences (Tavaré, 1986; Rodríguez, et al.1990). According to Yang and Kosakovsky's work, this model is considered to be the most complex model that fits for the appropriate set of characters (Yang and Nielsen, 1998; Kosakovsky, et al. 2007).

In different packages, the GTR model may have different names. For example, in the nucleotide package in PhyML, it is called the GTR model; while in the amino acid package in PhyML, it is called the REV model. The MtRev model, which was utilised for protein sequences in this project, is a special GTR model. In order to account for rate variation between sites the Gamma-distribution (G) and invariant sites (I), can be added to the GTR model. In this project, "GTR+G+I" means using the chosen model with a Gamma-distribution, along with invariant sites.

2.7 BOOTSTRAP ANALYSIS

After a phylogenetic tree was built, bootstrap analysis was used to test the confidence. Bootstrap analysis is a resampling technique used to estimate the confidence level of hypotheses in a phylogenetic tree. It was raised by Bradley Efron in 1979 to test the possibility of variation of results. It was a simple but effective method, and it generated random samplings from the original dataset with replacement. A measure of support for the branches in the tree is provided by bootstrap values. Each time a random sample of sites from the original data set is taken, the sample is subjected to the phylogeny estimation procedure, so that, for example, 100 trees are generated from 100 re-sampled data sets. A bootstrap value shows the number of trees, from this 100, which contain that particular branch of the tree. Usually the bootstrap value was underestimated. >95% was the confidence interval, which meant with a value within this interval, the result might hardly change and it had the highest credibility to be believed as the true structure. Usually >70% was a satisfactory result and the structure was stable (Graur & Li, 2000; Zvelebil & Baum, 2008).

2.8 PRESENTING THE FIGURES

After the phylogenetic reconstruction, the results as well as the bootstrap values are presented as figures. The programs used for generating tree pictures were: Tree Explorer, Archaeopteryx, Photoshop and PowerPoint. Tree Explorer was used to modify the trees with branches only. Bootstrap values

and node names can be added by using PowerPoint. In this stage, a ligand or group of ligands are selected to be the root of the phylogenetic tree.

When building the phylogenetic tree, BioEdit Sequence Alignment Editor (Version 7.0.5.3 10/28/05), MacGDE and Geneious were used to change file formats for different programs.

CHAPTER 3 PHYLOGENETIC ANALYSIS OF NODAL WITHIN THE TGF-BETA SUPERFAMILY

3.1 INTRODUCTION

Previous work by Herpin et al has divided the TGF-beta superfamily into four main groups (Herpin et al. 2004): the DVR subfamily, the activin/inhibin subfamily, the TGF-beta sensu stricto and related factor subfamily and a group of various divergent members, Nodal is in the DVR-subfamily. The phylogenetic tree shown in Herpin's review is shown in Figure 3.1 in this chapter. The phylogeny shows the relationships among the genes in the TGF-beta superfamily. In Figure 3.1, Nodal remains unaccompanied on a single branch. However, the low bootstrap values suggest that the structure of the tree may not be that reliable. A bootstrap value of 54% supports the position of Nodal within the DVR subfamily. This hints to the fact that Nodal may shift around subfamilies (Herpin et al. 2004).

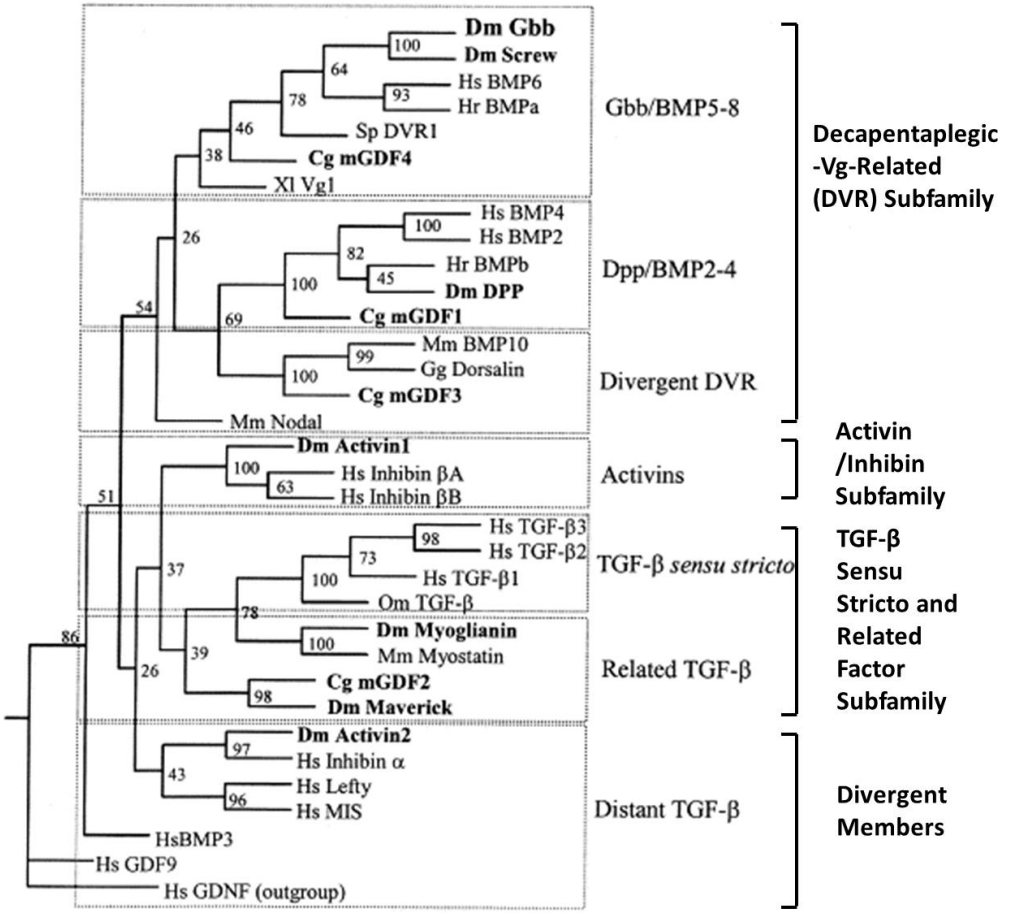

Figure 3.1 Phylogenetic relationships among the TGF-beta superfamily of ligands.

The 4 groups in the diagram represent the four major distinct ligand subfamilies of the TGF-beta superfamily. The first group is the DVR subfamily which includes GBB/BMP5-8, DPP/BMP2/4 and Divergent DVR. The second group is the activin/inhibin subfamily. The third group is TGF-beta sensu stricto and related factor subfamily which includes the TGF-beta sensu stricto and related TGF-beta ligands. The last group is a group representing various divergent members in the superfamily which illustrates distant TGF-beta. In this diagram, numbers at each branch node represent the percentage values given by bootstrap analysis. Protostome sequences are indicated in bold. GDNF (Glial Derived Neurotrophic Factor) is used as an out group. The tree is based on 120 amino acids.

What's more, findings from other researchers (Figure 3.2) are somewhat incompatible with Herpin's review about the position of Nodal in the TGF-beta superfamily. In the paper that first reported Nodal, the author states that Nodals are detached externally to a group of GDFs, BMPs, DPP and VG-1

(Zhou, et al. 1993). According to Bengtsson (2001), Nodal is an out-branch of a group of BMPs and GDFs. Similar to the position of Nodal in Figure 3.1, Bengtsson's paper suggests that Nodal does not abide with the Activin subfamily and TGFB subfamily (Bengtsson, 2001). In Newfeld's work (Newfeld, et al. 1999), Nodal is with DBL-1 and the group of Nodal and DBL-1 remains with the DVR subfamily with a quite low bootstrap number of 10, which suggests that structure may change. Next, in Ponce's study (Ponce, et al. 1999), Nodal stays with DBL-1 on a branch - but the relationship of this branch with other ligands is unclear. In Nguyen's paper (Nguyen, et al. 2000), Nodal forms a group with GDF10, BMP3 and Maverick. Figure 3.1 shows Nodal is in the DVR subfamily, Maverick is in the TGFB subfamily and BMP3 is a divergent TGF-beta superfamily member. The research to date therefore indicates that the position of Nodal within the TGF-beta superfamily remains uncertain, and there is no clear answer about which ligand or group of ligands is closest to Nodal. Moreover, although research on Nodal suggests that it is a monophyletic group, whether Nodal is truly a monophyletic group still needs to be ascertained. In most of the research mentioned above, the author only used one single Nodal sequence or Nodal sequences from one species to show the phylogeny of ligands of the TGF-beta superfamily. In this way, whether Nodal is monophyletic cannot be tested. Therefore, this project aims to: (1) try to bring as many Nodal sequences as possible to show if Nodal is monophyletic. (2) further study the Nodal gene in order to offer more information on the TGF-beta superfamily.

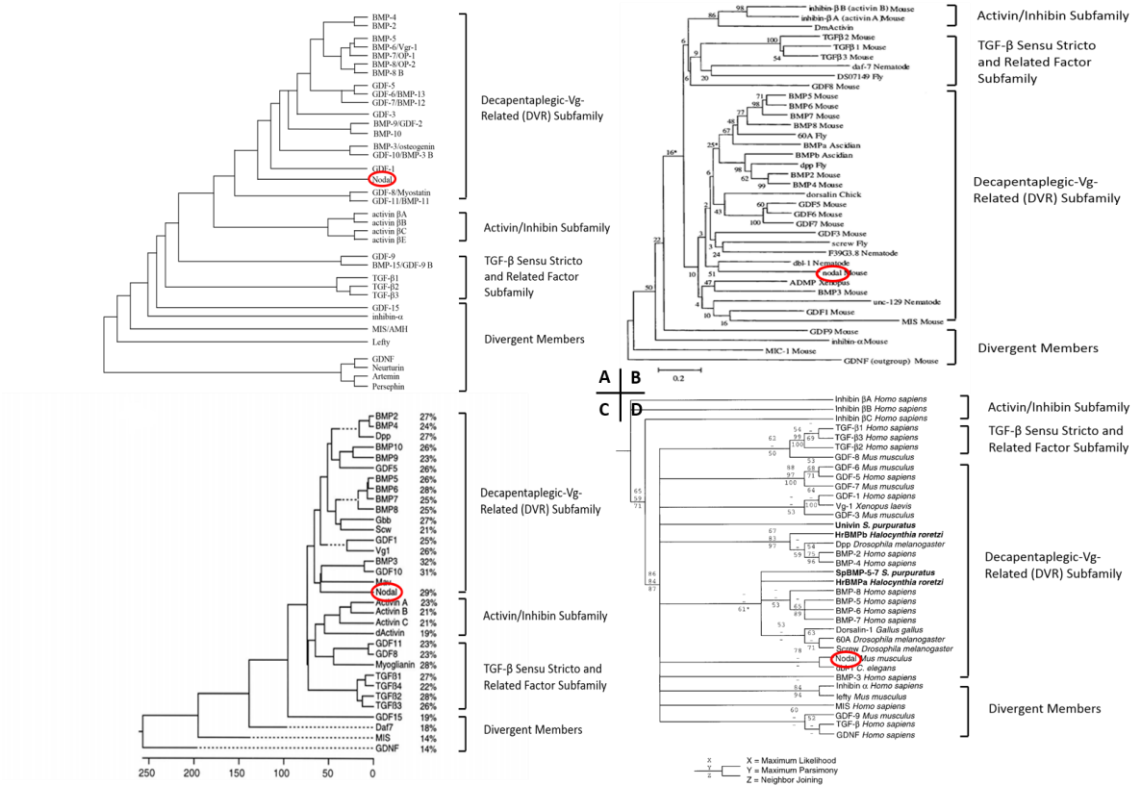

Figure 3.2 Different models from other researchers.

Those 4 figures illustrated the other researchers' theory of the location of Nodal within the TGF-beta superfamily. (A) Bengtsson's theory. Nodal is a single branch in DVR subfamily. (B) Newfeld's theory. Nodal stays along with DBL-1. (C) Nguyen's theory. Nodal stays with GDF10, BMP3 and Maverick. (D) Ponce's theory. Nodal stays with DBL-1 without support.

Based on the fact that in different researchers' work the position of Nodal in the TGF-beta superfamily may change, a phylogenetic tree which contains Nodal and other ligands of the TGF-beta superfamily will be produced to show the position of Nodal among the whole superfamily and determine relationships among ligands. This tree can provide answers to the following two objectives: (1) to examine whether Nodal is monophyletic; (2) to determine the relationship of Nodal with other ligands in TGF-beta superfamily.

3.2 MATERIALS AND METHODS

3.2.1 Sequence analysis

The nucleotide sequences used in this project were collected from GenBank and Ensembl through a detailed search of all the members of the ligands of the TGF-beta superfamily. In practice, the aim was to include as many Nodal sequences as possible. In this project, the DNA sequences were translated into Amino Acid sequences to provide protein information for building Amino Acid trees by the Mac Genetic Data Environment (MacGDE).

In this analysis, 711 sequences from the TGF-beta superfamily were brought into the alignment. 659 of the sequences were downloaded from GenBank and 52 of them were from Ensembl. 142 of them were Nodal sequences. The typical length of the nucleotide sequences in the whole TGF-beta superfamily was about 1000 to 1200bp. The common length of the nucleotide sequences of Nodal was about 800 to 1200bp.

3.2.2 Multiple sequence alignment

The dataset was aligned through a combination of automatic and manual alignment. First, the on-line Muscle service on the EBI website was used to automatically align the dataset. Then, based on the results of automatic alignment, the dataset was manually aligned through the program MacGDE. During the alignment, Nodal sequences were first brought into the database. Then, the ligand most similar to Nodal was brought in and aligned, then the

next most similar. And this was repeated until all the sequences were brought into the dataset and aligned.

After alignment, the sequences and the sites to be used in phylogenetic analysis were carefully selected. In this stage, sequence alignment markers were made to distinguish which sites were to be used in building phylogenetic trees. After the alignment, there are groups where sequences within a group are more similar than between groups. For example, TGFB1, TGFB2, TGFB3 and TGFB5 sequences show high similarity, so when making marker files they are seen as one group and together as one marker file. The marker file is used to show if the site marked will be included to build a phylogenetic tree. Only the unambiguously aligned sites were decided to be used in the further analysis.

3.2.3 Phylogeny Reconstruction

i. Choosing suitable datasets to build a phylogenetic tree

After making the marker files, a dataset that keeps a reasonable amount of the available sites was chosen to undertake analysis. In order to show the position of Nodal within the TGF-beta superfamily, the dataset must also contain a suitable number of ligands as well.

After choosing the dataset, before building the phylogenetic tree, some partial sequences were deleted from the dataset, because if those sequences were kept in, a large number of sites would be lost in tree reconstruction. In this stage, some sequences that would make long branches in the phylogenetic tree would also be removed.

ii. Saturation test

After the datasets were chosen, datasets used to build phylogenetic trees were tested for saturation. For nucleotide sequences, dataset 13 in Table 3.1 excluding the UNC-129 sequences was chosen to build trees. UNC-129 was removed because it would make a long branch in the phylogenetic tree. Before building trees, saturation tests were carried out for all three codon positions and only the1st and 2nd codon positions of the dataset sequence.

iii. PhyML

After the saturation test, the chosen datasets were used to build the final phylogenetic trees under maximum likelihood methods by PhyML. In practice, the MtRev model was utilised for protein sequences and GTR model was utilised for nucleotide sequences.

In PhyML, the model for nucleotide sequences was set as GTR+G+I. For amino acid sequences the model was MtREV+G+I.

If the dataset including all the ligands in the TGF-beta superfamily could be used in further phylogenetic analysis, the tree would be rooted at GDNF as Herpin did in his review (Herpin et al. 2004). This was done because the members of the GDNF family belong to the TGF-beta superfamily, but the amino-acid sequence homology is less than 20% for GDNF family members with other members of the TGF-beta superfamily (Airaksinen et al. 2002; Saarma, 2000).

If the dataset including all the ligands in TGF-beta superfamily was not used in further phylogenetic analysis. Then, since in Herpin's review (Herpin et al. 2004), Nodal was in the DVR subfamily, members from another subfamily other than the DVR subfamily could be chosen as the root. For example, if the dataset containing the DVR subfamily, the TGFB subfamily and some other ligands that were not included in Herpin's review (such as ADMP, DBL-1, UNC-129 and so on) were used to build a phylogenetic tree, that tree could be rooted on the TGFB subfamily. In this situation, all the other sequences are either in the DVR subfamily or not included in Herpin's review, so whether Nodal is monophyletic and its position could still be tested.

iv. Bootstrap

After building the phylogenetic trees in PhyML, bootstrap analyses were used to test the credibility of the result. The number of replicates of the non-parametric bootstrap analysis was set as 100 for both amino acid sequences and nucleotide sequences.

3.3 RESULTS

Sequences of the TGF-beta superfamily were aligned after being downloaded. As mentioned in Chapter 3.2.2, during the alignment, the whole database was divided into groups. The whole dataset was divided into 21 datasets, and each of the datasets had a marker file indicating the alignable sites for that dataset. The situation of ligands and numbers of aligned sites are shown in Table 3.1.

Dataset number	Number of aligned nucleotide sites	Number of aligned nucleotide sites after removal of UNC-129	Ligands included
1	462	462	Nodal
2	393	393	ADMP
3	372	372	GDF5, GDF6
4	318	Removed	UNC-129
5	279	315	DBL-1, DPP
6	279	300	VG1, Derriere, GDF9, GDF10
7	261	273	DVR1
8	261	273	GBB
9	261	273	BMP2-8/10/15
10	261	273	SCREW
11	243	255	GDF1, GDF2, GDF3, GDF7
12	240	240	DAF-7
13	240	240	TGFB1, TGFB2, TGFB3, TGFB5
14	234	234	MYOGLIANIN, Myostatin, GDF11
15	234	234	Maverick
16	234	234	GDF15
17	228	228	Activin/Inhibin
18	144	144	GDNF
19	84	84	AMH
20	63	63	TGFB4, LEFTY
21	9	9	MGDF

Table 3.1 Number of markers in each dataset.

Column “Dataset number” shows the serial number of the dataset. Column “Number of aligned sites” shows the number of nucleotide sites to be used in the phylogenetic analysis if that dataset is chosen. Column “Ligands included” shows the ligands included in addition to those in the previous dataset. The first ligand to be included is Nodal. In this table, dataset number k contains all the genes listed in row k plus all the ones listed in earlier rows i<k.

The dataset that will be used to build the phylogenetic trees to show the relationships within the TGF-beta superfamily then needs to be selected. Obviously it is better to use the whole TGF-beta superfamily to build a tree when examining the relationships within the superfamily, but as shown in

Table 3.1, if all the ligands are included to make a tree, it would present too few sites to build a useful tree (as shown in Table 3.1, Dataset No.20 or No.21). In order to keep a balance of including as many ligands as possible while including a reasonable amount of sites, Dataset No.13 was chosen to build a tree showing the relationship of Nodal and other TGF-beta superfamily members. Based on Table 3.1, the amount of sites is not so few as Dataset No.20 or No.21 in Table 3.1, and based on the reference tree shown in Figure 1.3, all genes that are added into the datasets above dataset 13 except GDF9 are either in the DVR subfamily in Figure 3.1 or not included in Herpin's review (such as ADMP, DBL-1, UNC-129 and so on). In Figure 3.1, the author groups the whole superfamily into 4 groups: the DVR subfamily; TGFB subfamily; Activin/Inhibin subfamily; and distant TGF-beta ligands. In dataset 13, Nodal, DPP, DVR1, GBB, SCREW, BMP2-8/10, VG1 are in the DVR subfamily in Figure 3.1; TGFB1, TGFB2, TGFB3, TGFB5 are in TGFB subfamily in Figure 3.1; GDF9 is in the distant TGF-beta ligands group; ADMP, GDF5, GDF6, UNC-129, DBL-1, Derriere, BMP15, DAF-7,GDF1, GDF2, GDF3, GDF7, GDF10 are not included in Herpin's review. So to use dataset 13, the position of Nodal among the whole superfamily could be found and the relationship of Nodal with the DVR subfamily and TGFB subfamily can be observed. Furthermore, a tree based on dataset 13 can show whether Nodal is monophyletic as well as showing the ligand that is closest to Nodal.

In this situation, the final tree of Nodal with some other TGF-Beta superfamily members would be rooted at the TGFB subfamily. In this tree, all the other sequences except the root groups would be either in the DVR subfamily or not

included in Herpin's review. Then whether Nodal is monophyletic and its position could still be tested.

As mentioned in Chapter 3.2.3, UNC-129 was removed because it would make a long branch in the phylogenetic tree. UNC-129 is removed because although UNC-129 is a nematode TGF-beta gene, it is very different from other TGF-beta ligands both in its sequence and its functional pathway. Nematodes do not require conventional TGF-beta receptors and Smads and the TGF-beta pathway is different in nematodes from the TGF-beta pathway in other species. (Padgett & Patterson, 2006; Colavita et al. 1998). As shown in Table 3.1, to remove UNC-129 could bring in more sites in dataset 5~11, but did not affect the number of aligned sites in datasets 12 and 13.

Some fragmentary sequences, that only contained a short partial region which would sharply reduce the number of sites included in further analysis, were also removed from the dataset. The sequences removed from dataset No.13 were listed in Table 3.2. Those sequences were deleted because they were partial sequences or they could cause a long-branch problem.

Ligands	Sequence Name in Dataset	NCBI ID	Ensembl ID
Nodal	frog_N_3		ENSXETG00000016778
Nodal	Sloth_N		ENSCHOG00000010347
Nodal	hedgehog1_N		ENSETEG00000013276
Nodal	Pig_N_3	AM072821.1	
Nodal	pig_N_2		ENSSSCG00000010265
Nodal	Alpaca_N		ENSVPAG00000002635
Nodal	Chicken_Nr1_1	AF486810.1	
Nodal	Onr1_3	AB116041.1	
Nodal	Onr2_3	AB116642.1	
Nodal	Onr1_1	EF206724.1	
Nodal	Onr2_1	EF206725.1	
Nodal	finch_NH	XM_002194155.1	
Nodal	CyNodal_2	AB114684.1	
ADMP	Mouse_ADMP	AF365876.1	
ADMP	Human_ADMP2	AF458592.1	
ADMP	Human_ADMP	AK312144.1	
ADMP	salmon_ADMP2	BT057114.1	
ADMP	salmon_ADMP	NM_001146504.1	
ADMP	wasp_ADMP	XM_001604676.1	
ADMP	tick_ADMP	XM_002402657.1	
ADMP	Junglefowl_ADMP	XM_422812.2	
GDF5	sea anemone_GDF5_1	AY391717.1	
GDF5	sea anemone_GDF5_2	AY496945.1	
UNC-129	nematode_UNC-129_1	AF029887.1	
UNC-129	nematode_UNC-129_2	NM_069165.4	
DPP	sludgeworm_DPP	AB192888.1	
DPP	millipede_DPP	AJ843875.1	
DPP	bug_DPP	AY899334.1	
DPP	butterfly_DPP	EU233806.1	
TGF-beta 2	nematode_TGFB2	AF104016.1	
TGF-beta 2	hookworm_TGFB2	AY942844.1	

Table 3.2 Deleted sequences.

The sequences listed were the sequences that were deleted after alignment because they were partial sequences or they could cause a long-branch problem.

The final dataset that would be used in the phylogenetic analysis was made based on dataset No.13 but removing some sequences listed in Table 3.2. The sequences that were included in the phylogenetic analysis were: 129 Nodal gene sequences, 16 sequences of ADMP, 12 sequences of GDF5, 11 sequences of GDF6, 29 sequences of DPP, 2 sequences of DBL-1, 8 sequences of VG1, 5 sequences of Derriere, 27 sequences of GDF9, 11 sequences of GDF10, 3 sequences of DVR1, 6 sequences of GBB, 5 sequences of BMP2, 5 sequences of BMP 3, 11 sequences of BMP 4, 8 sequences of BMP 5, 9 sequences of BMP 6, 4 sequences of BMP 7, 10 sequences of BMP 8, 5 sequences of BMP 10, 9 sequences of BMP 15, 2 sequences of SCREW, 10 sequences of GDF1, 18 sequences of GDF2, 12 sequences of GDF3, 11 sequences of GDF7, 5 sequences of DAF-7, 7 sequences of TGFB1, 3 sequences of TGFB2, 13 sequences of TGFB3, 3 sequences of TGFB5.

3.3.1 Saturation Test

After the dataset was selected, the dataset was examined for evidence of substitution saturation to analyse the accuracy of the phylogenetic tree. In the saturation test, uncorrected pairwise transition (ti) and transversion (tv) distances were plotted against pairwise total uncorrected distances, and uncorrected pairwise transition distances were plotted against transversion distances for all three codon positions and only the 1st and 2nd codon positions.

As mentioned in Chapter 2.4, when examining the uncorrected pairwise transition and transversion distances against pairwise total uncorrected

distances, if both the transition line and transversion line are straight lines, it suggests there is no saturation in the dataset. If either line is curved, it suggests that the dataset is saturated. Usually it is the transition line curved and crossing transversion line. When examining the uncorrected pairwise transition distances against transversion distances, if most points are set above the line y=x, it suggests there is no saturation in the dataset.

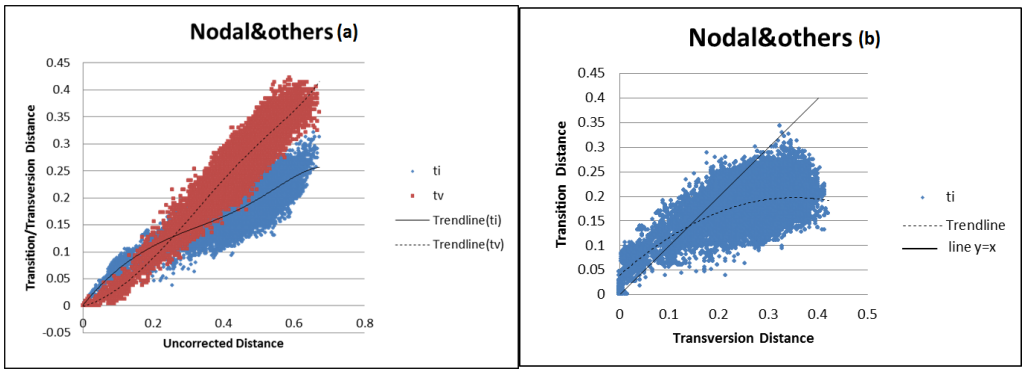

Figure 3.3 Saturation test for all three codons of Nodal with some other TGF-beta superfamily members.

(a)Uncorrected pairwise transition (ti) and transversion (tv) distances against pairwise total uncorrected distances for Nodal and other ligands in the TGF-beta superfamily. (b) Uncorrected pairwise transition (ti) distances against transversion (tv) distances for Nodal and other ligands in the TGF-beta superfamily. The diagonal stands for the line y=x.

Figure 3.3 (a) showed the saturation test result of uncorrected pairwise transition (ti) and transversion (tv) distances against pairwise total uncorrected distances for Nodal and other ligands in the TGF-beta superfamily. In Figure 3.3 (a), the transition line formed a curve and crossed the transversion line. This suggested that the transitions are saturated. Figure 3.3 (b) showed the saturation test result of uncorrected pairwise transition (ti) distances against transversion (tv) distances for Nodal and other ligands in the TGF-beta superfamily. In Figure 3.3 (b), most points were in the area under the line y=x, which clearly showed that the dataset is saturated. As shown in the Figure 3.3,

the dataset of Nodal and other ligands was saturated. Trees developed from the dataset with all 3 codons may therefore be inaccurate.

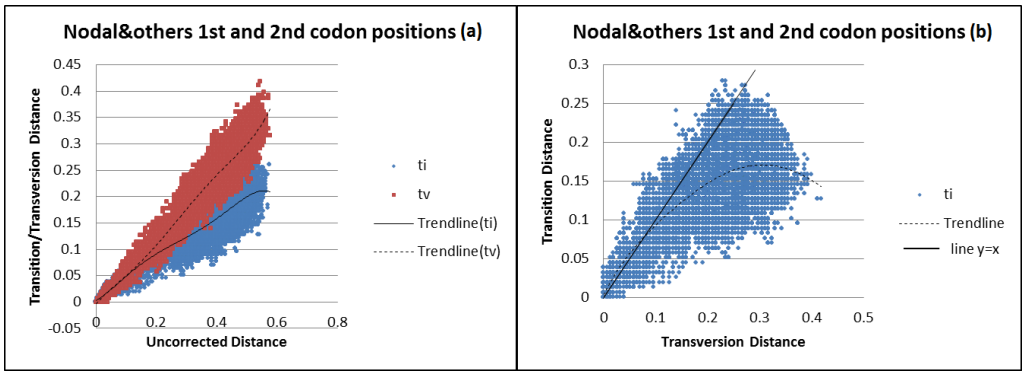

Figure 3.4 Saturation test for the 1st and 2nd codon positions of Nodal with some other TGF-beta superfamily members.
(a)Uncorrected pairwise transition (ti) and transversion (tv) distances against pairwise total uncorrected distances for the 1st and 2nd codon positions of Nodal and other ligands in the TGF-beta superfamily. (b) Uncorrected pairwise transition (ti) distances against transversion (tv) distances for the 1st and 2nd codon positions of Nodal and other ligands in the TGF-beta superfamily. The diagonal stands for the line y=x.

Figure 3.4 (a) showed the saturation test result of uncorrected pairwise transition (ti) and transversion (tv) distances against pairwise total uncorrected distances for the 1st and 2nd codon positions of Nodal and other ligands in the TGF-beta superfamily. In Figure 3.4 (a), the transition line formed a curve and crossed the transversion line at quite an early stage. Figure 3.4 (b) showed the saturation test result of uncorrected pairwise transition (ti) distances against transversion (tv) distances for the 1st and 2nd codon positions of Nodal and other ligands in the TGF-beta superfamily. Figure 3.4 (b) showed that most points were in the area under the line y=x, which suggested that saturation happened. As shown in Figure 3.4, the dataset with 1st and 2nd codon positions was saturated; this suggests trees developed from this dataset with 1st and 2nd codon positions may also be inaccurate.

The saturation test showed that the two datasets of nucleotide sequences with all three codons positions and 1st/2nd codon positions were all saturated. The tree developed from those datasets may be inaccurate. But those datasets were still used to build phylogenetic trees for three reasons: First of all, a phylogenetic tree is still needed to show the relationships of Nodal among the TGF-beta superfamily. Secondly, using dataset 10 instead means removing the TGFB subfamily ligands and some of the ligands that were not included in Herpin's review but were included in dataset 13. Although it may bring in more sites and may have less saturated data than dataset 13, it would not be possible to tell whether Nodal is within the DVR subfamily or not. Third, the alignment used to make up the datasets is the best one that can be provided. However, the problems of saturated data can be reduced to some extent by using a more complex likelihood model (Farrell, 2011).

3.3.2 Phylogenetic Trees

After the saturation test, with the aim to examine whether Nodal was monophyletic and to find if there was a neighbour ligand or group of ligands for Nodal, phylogenetic analyses based on dataset No.13 in Table 3.1 were carried out. Phylogenetic trees were developed based on protein sequences, which were translated by MacGDE (Figure 3.5), 1st/2nd codon position (Figure 3.6) and all 3 codons of DNA sequences (Figure 3.7) through the maximum likelihood (ML) method, and these show the relationship of Nodal and other ligands in the TGF-beta superfamily. The relationships within Nodal will be discussed in the next chapter (Chapter 4).

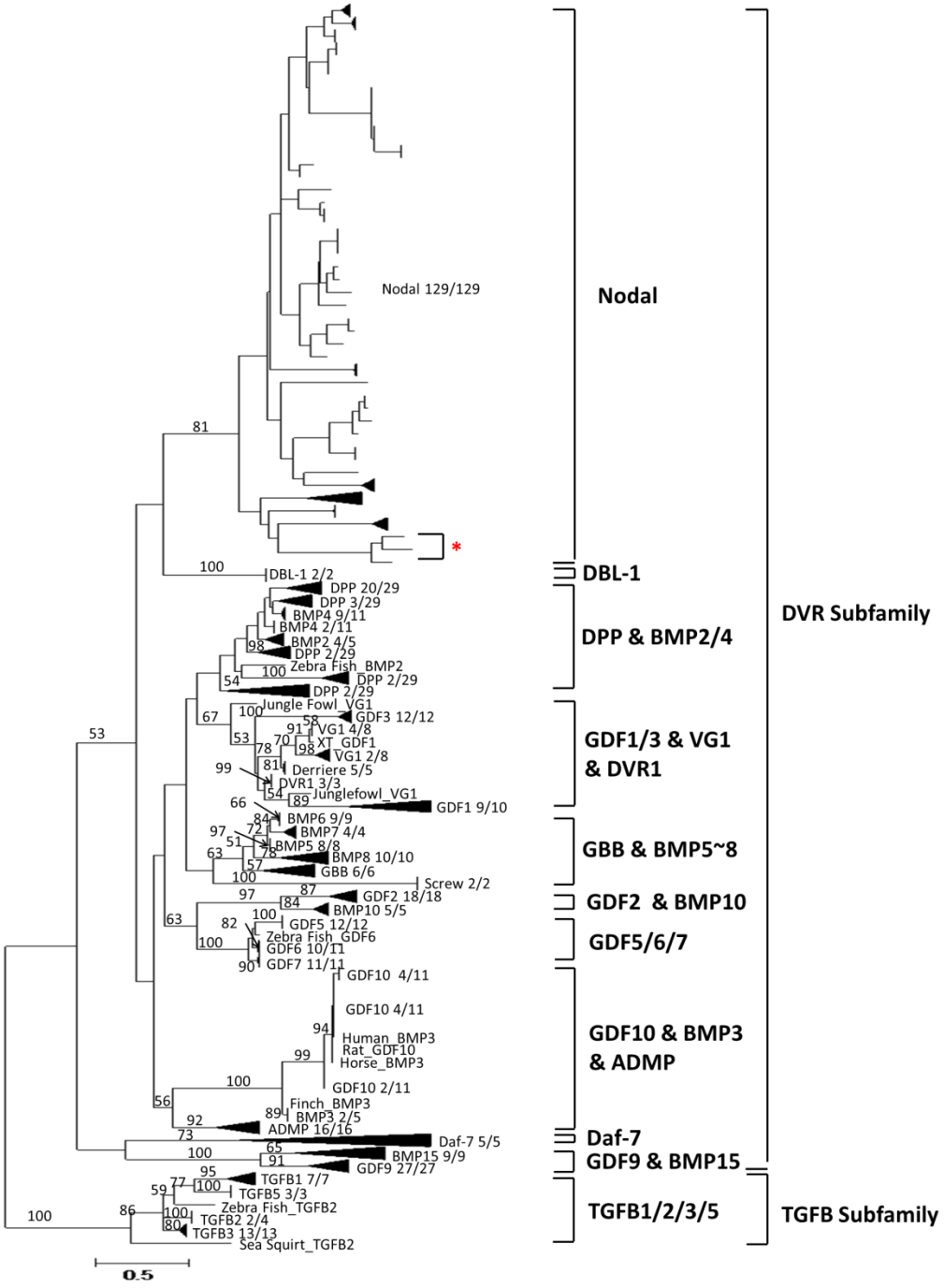

Figure 3.5 Maximum likelihood amino acid phylogenetic tree showing the phylogenetic position of Nodal within the TGF-beta superfamily

This tree is built based on 79 amino acid sites. The scale bar corresponds to 50 changes per 100 nucleotide positions. Numbers on branches represent the bootstrap value of that branch based on 100 replicates. Only values higher than 50% are shown. The tree is rooted on TGFB subfamily.

In the phylogenetic tree built from amino acid sequences of Nodal and other ligands in the TGF-beta superfamily (Figure 3.5), Nodal forms a monophyletic group with a strong bootstrap value of 81%. Figure 3.5 also shows that although the bootstrap value is low (Nei & Kumar, 2000), Nodal is supported with a bootstrap value of 53% to be with the main DVR subfamily members DPP&BMP2/4, GBB&BMP5~8 and BMP10. However, there is insufficient evidence to reveal the relationship of Nodal among other ligands in the DVR subfamily. That is to say, all the other ligands within the DVR subfamily could be nearest to Nodal. So the nearest neighbour ligands cannot be found through this phylogenetic tree.

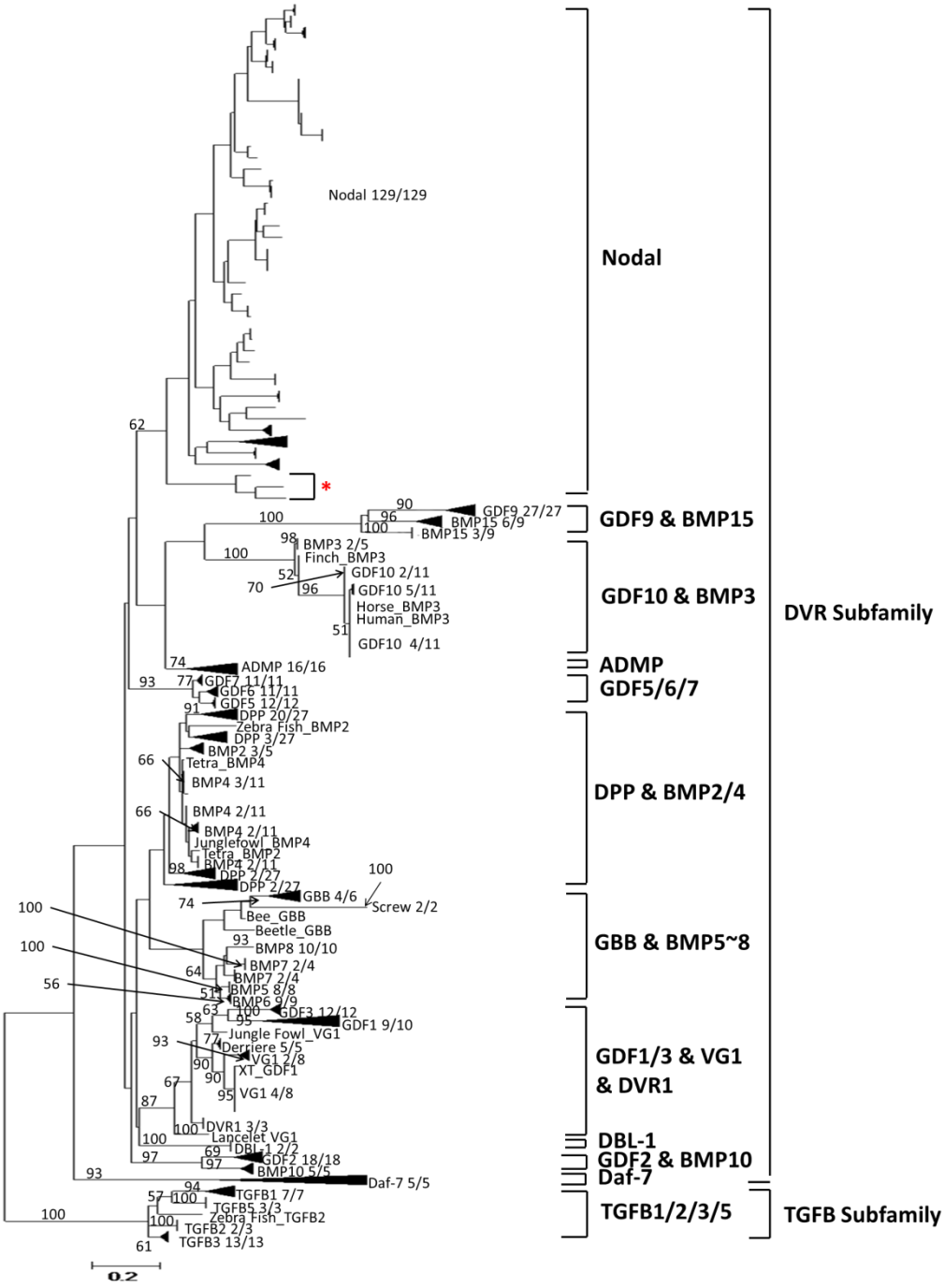

Figure 3.6 Maximum likelihood 1st and 2nd codon nucleotide phylogenetic tree showing the phylogenetic position of Nodal within the TGF-beta superfamily

This tree is built based on 158 nucleotide sites by using 1st and 2nd codon positions only. The scale bar corresponds to 20 changes per 100 nucleotide positions. Numbers on branches represent the bootstrap value of that branch based on 100 replicates. Only values higher than 50% are shown. The tree is rooted on TGFB subfamily.

In the tree built from nucleotide sequences with 1st and 2nd codon positions (Figure 3.6), Nodal is a monophyletic group with a low bootstrap support value of 62%. The relationship between Nodal and other ligands is still uncertain.

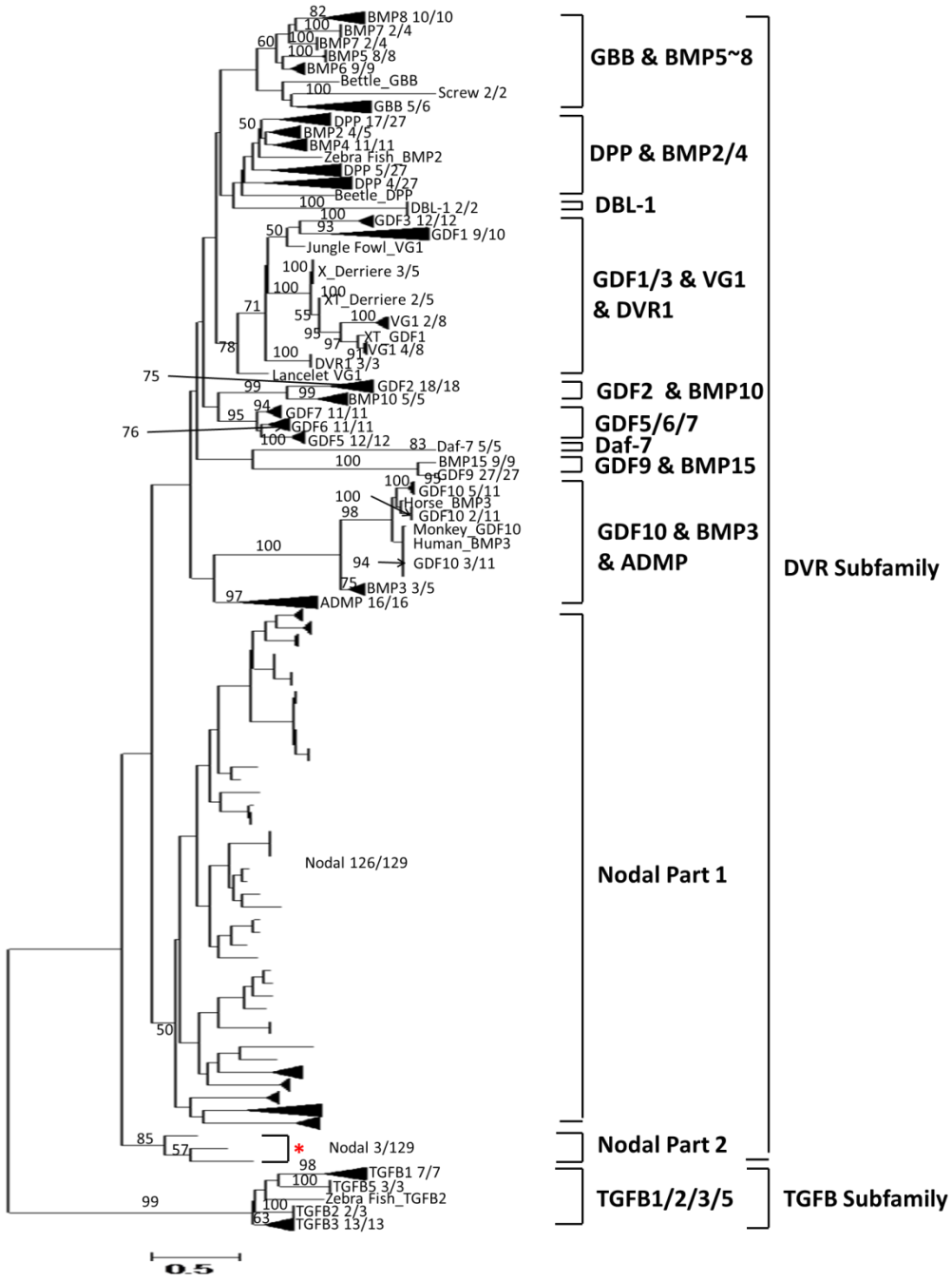

Figure 3.7 Maximum likelihood nucleotide phylogenetic tree showing the phylogenetic position of Nodal within the TGF-beta superfamily

This tree is built based on 237 nucleotide sites by using all three codon positions. The scale bar corresponds to 50 changes per 100 nucleotide positions. Numbers on branches represent the bootstrap value of that branch based on 100 replicates. Only values higher than 50% are shown. The tree is rooted on TGFB subfamily. *: The highlighted group of Nodal that named as Nodal Part 2 in Figure 3.7 is also marked by "*" in Figure 3.5 and Figure 3.6. Those sequences are Nodal of limpet, snail and sea slug.

In the tree of nucleotide sequences with all three codons (Figure 3.7), Nodal is not monophyletic. There are two groups of Nodal, one contains Limpet, Snail and Sea Slug sequences and another contains other Nodals. Again, bootstrap values of basic structures in the tree shown in Figure 3.7 are low.

From the 3 figures it can be seen that there is some support to indicate that Nodal is monophyletic (the value is 81% in amino acid dataset, Figure 3.5 and with a value of 62% in the nucleotide dataset included 1st/2nd codon positions, Figure 3.6). However, there is also limited support to suggest that Nodal is not monophyletic in the nucleotide dataset including all three codon positions (Figure 3.7). Moreover, as the 3rd codon positions will change more frequently than the 1st and 2nd codon positions, the 3rd codon positions will be more easily saturated. This suggests that the results of the phylogenetic tree with all three codon positions (Figure 3.7) may be more inaccurate than the one with 1st/2nd codon positions alone (Figure 3.6).

Lastly, it can be seen that Nodal appears more likely to stay in the DVR-subfamily; however, the position of Nodal within that subfamily cannot be tested from the phylogenetic trees (Figure 3.5, Figure 3.6 and Figure 3.7) because of the low support value.

3.4 DISCUSSION

3.4.1 Examination of whether Nodal is monophyletic

Both the Amino Acid tree (Figure 3.5) and the Nucleotide tree with 1st and 2nd codon positions (Figure 3.6) show Nodal is monophyletic. However, it is

only in the Amino Acid tree where there is a high support to state that Nodal is a monophyletic group (In Figure 3.5, the bootstrap value is 81%). In Figure 3.6, the value is lower than 70 % (the value is 62% in Figure 3.6).

However, the nucleotide tree with all 3 codons (Figure 3.7) does not support Nodal being a monophyletic group. In Figure 3.7, Nodal sequences are divided into two groups with low evidence. One of the groups contains the only three Nodal sequences that came from the gastropoda class, while the other group contains all other Nodal sequences. But considering that the 3rd codon positions may change more frequently than the 1st and 2nd codon positions, when including the 3rd codon positions (which means including all three codon positions), the phylogenetic result may be more inaccurate than the one with 1st/2nd codon positions.

Considering the following 2 points: (1) the result of the phylogenetic tree with all three codons of nucleotide sequences (Figure 3.7) may be more inaccurate than the one with 1st/2nd codon positions (Figure 3.6). (2) in the amino acid tree (Figure 3.5), the monophyletic group of Nodal is well supported, I suggest that Nodal is monophyletic.

3.4.2 The position of Nodal within the TGF-beta superfamily

There is some support to indicate that Nodal is in the DVR subfamily, but the support value for this is not high (the value is 53 in Figure 3.5 and lower than 50 in Figure 3.6 and Figure 3.7). Given the low bootstrap value, the exact position of Nodal in the DVR subfamily remains uncertain. Through the three

trees, the single ligand or ligand group nearest to Nodal remains uncertain. In Figure 3.5, DBL-1 seems to be the answer but with an extremely low support value. In Figure 3.6, the groups of GDF-9, BMP15, GDF10, BMP3 and ADMP are next to Nodal, but again with a support lower than 50%.

3.4.3 Comparison of the function of Nodal and other ligands

Nodal is involved in mesoderm differentiation in vertebrates. Nodal plays an important role in mesoderm formation, anterior-posterior axis formation and left-right axis formation in vertebrate development.

BMPs can induce animal or human mesenchymal cells to differentiate into bones, cartilages, ligaments, tendons and nerve tissues. GDFs perform functions predominantly related to development. They play a crucial role in cell differentiation regulation in both adult tissues (such as ovary, thymus and spleen) and embryogenesis. Dorsalin is one type of GDF2. DPP (decapentaplegic) is the skin growth factor of organisms. It affects the skin colour on the back of organisms. It is a functional ortholog of mammalian BMP-2 and BMP-4. Vg1 and DVR1 (decapentaplegic and Vg-related 1) are also named GDF1 in some researchers' papers to make terminology consistent (Helde and Grunwald, 1993). Daf-7 is important to control dauer larva development in Nematode (*Caenorhabditis elegans*) (Matt Crooka, et al. 2005). GBB (Glass bottom boat, 60A) regulates synaptic growth at the Drosophila neuromuscular junction. GBB is a functional ortholog of mammalian BMP5~8. Screw (SCW) is a DPP/GBB like gene. It affects specification of the Drosophila embryo dorsal cell. (Ongkar Khalsa, et al. 1998) Derriere is closely

related to Vg1. It is induced by VG1 in animal cap explants and can rescue the L-R orientation that is changed by VG1. It also plays a role in posterior development in Xenopus. Derriere is involved in earlier molecular pathways developing the L-R asymmetry (Hiroshi Hanafusa, et al. 2000; B.I. Sun, et al. 1999) ADMP (Anti-Dorsalizing Morphogenetic Protein) is most closely related to human BMP-3. From the phylogenetic trees in Chapter 3.3.2 it can be seen that GBB and BMP5~8 stay together to form a GBB&BMP5~8 group, DPP and BMP2/4 stay together to form a DPP&BMP2/4 group.

ADMP is induced by lithium chloride treatment or activin. It has the ability to inhibit the development of dorsoanterior structures and mitigate organizer-associated dorsalizing influences (M. Moos, et al. 1999). During the alignment, it can be seen that the gene structure of ADMP is quite similar to Nodal. But in the phylogenetic trees in Chapter 3.3.2, the relationship of Nodal and ADMP remained unclear because of the low supported branches.

Lefty is an antagonist of Nodal signalling which directly inhibits Nodal signalling by competitive binding to Nodal receptors and plays an important function in L-R patterning in early vertebrate embryos. It is found in the midline structures and serves as a barrier to prevent the crossing of left or right determinants. It is further found in the left lateral plate mesoderm (LPM) to be a negative feedback regulator of Nodal signals to determine the left side identity. Among the whole superfamily, Lefty has the most similar function of Nodal. But in Herpin's tree (Figure 3.1), it stays far from Nodal. In the alignment stage of this project, it also can be seen that Lefty sequence is very

different from Nodal. So the lefty sequences were excluded when building the phylogenetic tree to prevent losing sites.

3.4.4 Relationships within the TGF-beta superfamily

To determine the relationships among the ligands, it is shown in the reference tree in Figure 3.1 that in the DVR subfamily, DPP/BMP2/4 usually forms a group and stays close to the group of GBB/BMP5~8. However, in Figure 3.5, Figure 3.6 and Figure 3.7, the values that support the position of DPP/BMP2/4 and GBB/BMP5~8 were lower than 50.

VG1, DVR1, GDF1 and GDF3 are usually in a branch with good support together with the group of GBB/BMP5~8 and DPP/BMP2/4 in the DVR subfamily. However, this is not as Herpin described in Figure 3.1 where VG1, DVR1, GDF1 and GDF3 seemed to be more likely to be in a separate group instead of forming a group together within the group of Gbb/BMP5~8, however, again, the support of VG1, DVR1, GDF1 and GDF3 were in a separate group from Gbb/BMP5~8 was weak.

3.4.5 Future work

In the phylogenetic analysis in this chapter, the TGF-beta superfamily ligands sequences are downloaded and aligned without any selection in hope to show a full view of the whole TGF-beta superfamily. However, the varied amount of sequences from different species and different kinds of ligands makes the data

saturate and sharply reduce the sites that can be aligned through the whole database.

Based on the work in this chapter, it seems to be sure that Nodal is a ligand in the DVR subfamily. To find the exact position of Nodal within the DVR subfamily, the ligands that are contained in this chapter seems to be the minimum set. They only contained the DVR subfamily members, one other group of TGFB as the out-group to root the DVR subfamily tree and some ligands whose positions remain unclear. This means the sites are the most statistically powerful we can get in this situation. However, there is still insufficient evidence to support the structure within the DVR subfamily.

If a well-supported tree could be built in the future, both the position of Nodal and the neighbour ligand or group of ligands of Nodal could be determined. That may be helpful for further analysis to examine the relationships within Nodal.

3.5 CONCLUSION

In conclusion, according to the phylogenetic analyses presented here, Nodal seems to be a ligand within the DVR subfamily of the TGF-beta superfamily. Nodal is monophyletic but the ligand or ligand group next to it is uncertain.

CHAPTER 4 PHYLOGENETIC TREE OF NODAL TO TEST THE EVOLUTIONARY RELATIONSHIPS OF NODAL GENES FROM DIFFERENT SPECIES

4.1 INTRODUCTION

Previous research showed two ways of grouping Nodal in different species. One stated that there were three groups within Nodal as shown in Figure 4.1: Group C contained humans, the mouse, the opossum, the African clawed frog and the anole lizard; Group B contained 1 copy of bony fish; Group A contained the chicken, other 2 copies of bony fish and the other copies of the anole lizard and the African clawed frog. The base of the tree consisted of various divergent members such as the sea squirt and the lancelet (Kuraku & Kuratani, 2011). The other one demonstrated that there were two major groups of Nodals as shown in Figure 4.2. In this theory, Group B and Group C stayed on the same branch. The base of the tree consists of various divergent members such as the sea squirt and the lancelet (Fan & Dougan, 2007).

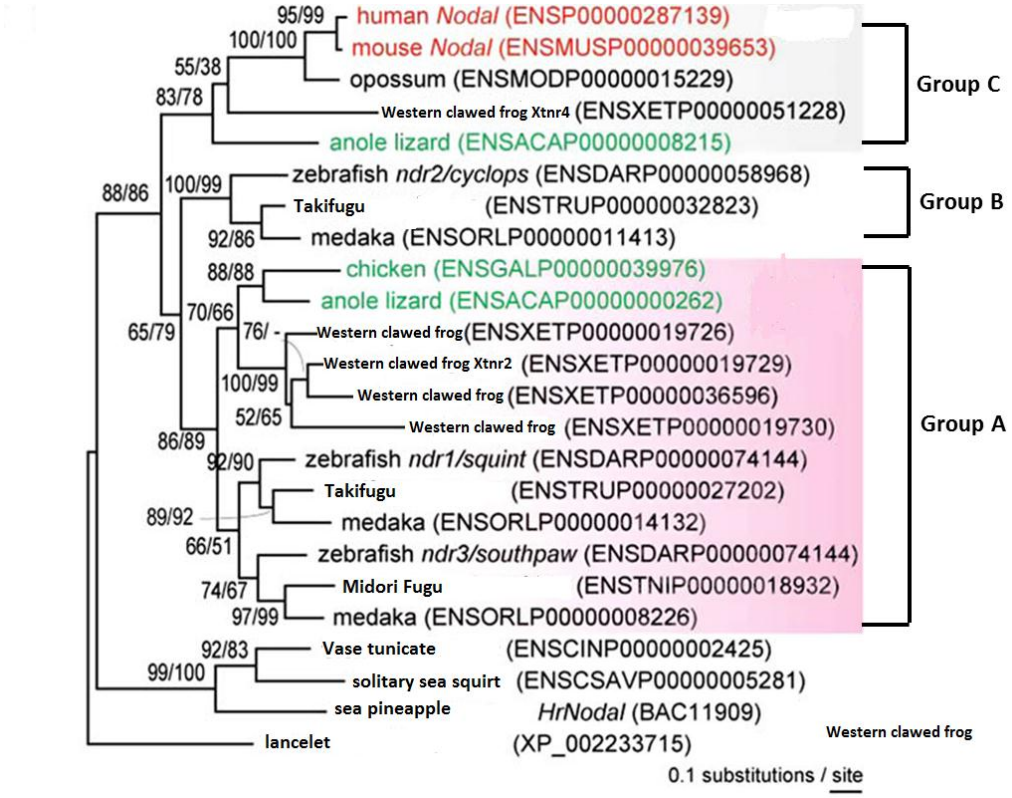

Figure 4.1 Kuraku's theory which suggests three groups within Nodal

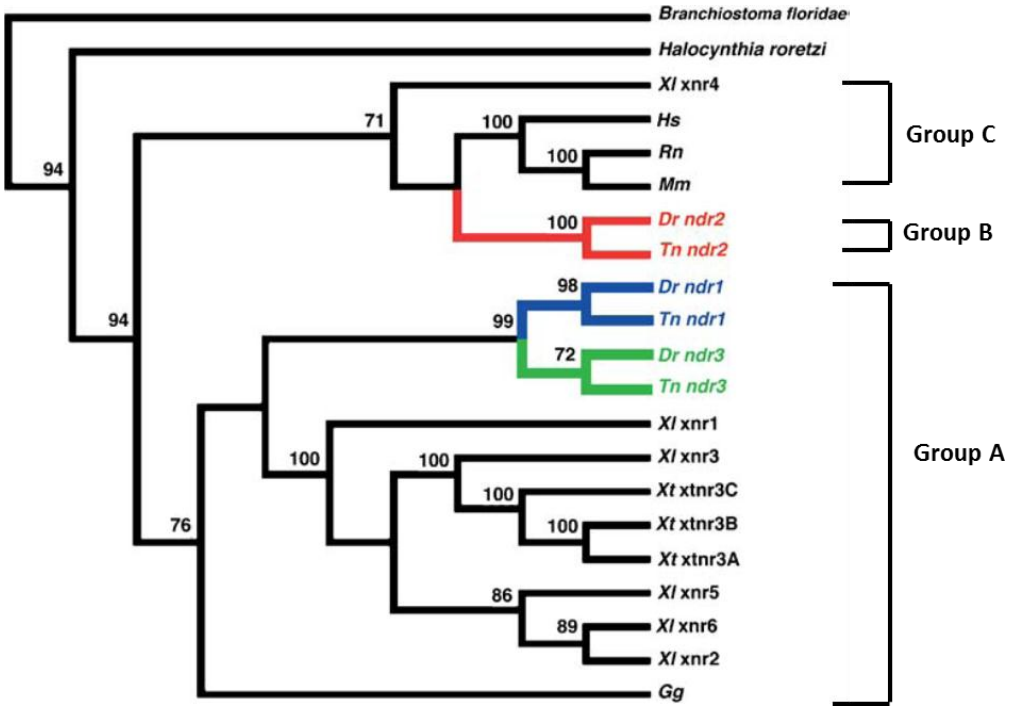

Figure 4.2 Fan's theory which suggests two groups within Nodal

The summarised Latin names in this figure refer to: Xl= *Xenopus laevis*; Hs= *Homo sapiens*; Rn= *Rattus norvegicus*; Mm= *Mus musculus*; Dr= *Danio rerio*; Tn= *Tetraodon nigroviridis*; Gg= *Gallus gallus*.

The research to date therefore suggests that the Nodal genes can be divided into two major groups, with various other divergent members at the base of the tree. Nevertheless, which species are in which group remains uncertain. Moreover, although in Hellsten's supplementary material, he mentioned there were 2 types of copies of Nodal, he didn't provide phylogenetic support to this hypothesis (Hellsten, 2010). So how Nodal is duplicated in evolution still remains to be ascertained. In the research referred to above, the authors use only some of the Nodal genes available to demonstrate the phylogeny of Nodal from some species. In this way, the relationship in Nodal genes across all species cannot be tested. This project attempts to bring in as many Nodal sequences as possible in order to establish a general view of the relationship among Nodal genes of different species, and to ascertain whether Nodal is duplicated in different species. Compared to the previous work, many species were included in this project, such as: three-spined stickleback fish, turkey, axolotl, limpet, snail, sea urchin and a large number of mammals such as hedgehog, pig, horse, monkey, chimpanzee, rhesus monkey, dog, cattle, tarsier, marmoset, cavy, armadillo, kangaroo rat, cat, gorilla, elephant, kangaroo, opossum, lemur, bat, rabbit, galago, orangutan, dolphin, alpaca, flying fox, shrew, rock hyrax and squirrel.

This study has the following objectives: (1) to determine the evolutionary relationship within Nodal; (2) to examine whether Nodal is duplicated in different species.

To achieve these objectives, a phylogenetic tree which contains all Nodal genes is used to determine the relationship among Nodal sequences.

4.2 MATERIAL AND METHOD

4.2.1 Sequence analysis

There were 142 Nodal genes that were found and downloaded from the online database. Among them, 90 sequences were from GenBank and 52 were from Ensembl. In the 90 Nodal genes from GenBank, 68 of them were proven to be Nodal sequences by an experiment that had been reported and published. 15 of them were predicted by a search engine, 22 of them remained unknown (3 of them were submission only, and 19 of them were unpublished).

After the alignment, some sequences that were included in the analysis in Chapter 3 would be excluded from further analysis in this chapter. Those sequences were excluded because they were partial sequences. In the analysis in Chapter 3, they would not have affected the number of aligned sites in dataset 13. However, in this chapter, if those sequences had been included, the number of sites would have been sharply reduced.

Finally, the genes of the dataset that would be used in the final phylogenetic analysis that were listed in Table 4.1 were: 42 mammalian Nodal genes, of which 5 were human, 3 were mouse, 2 were hedgehog, 5 were pig, 2 horse, 2 monkey, and 1 sequence for the chimpanzee, rhesus monkey, dog, cattle, tarsier, marmoset, cavy, armadillo, kangaroo rat, cat, gorilla, elephant, kangaroo, opossum, lemur, bat, rabbit, galago, orangutan, dolphin, alpaca,

flying fox, shrew, rock hyrax, squirrel and rat; 3 bird Nodal genes, of which 2 were chicken and 1 turkey; 16 fish Nodal genes, of which 3 were three-spined stickleback fish, 6 were fugu, 2 were Japanese killifish and 5 were zebrafish; 52 amphibian Nodal genes, of which 49 were frog, 2 were axolotl and 1 newt; 2 gastropoda Nodal genes, of which 1 was from the limpet and 1 from the snail; 6 sea urchin Nodal genes, 2 lancelet Nodal genes and 2 sea squirt Nodal genes were included as well.

NCBI ID	Ensembl ID	Name in tree	Class	Species	bp
NM_001085796.1		Xnr1_1	Amphibian	African Clawed Frog	1515
U29447.1		Xnr1_2	Amphibian	African Clawed Frog	1515
BC169388.1		Xnr2_1	Amphibian	African Clawed Frog	1338
BC169392.1		Xnr2_2	Amphibian	African Clawed Frog	1338
NM_001087967.1		Xnr2_3	Amphibian	African Clawed Frog	1459
U29448.1		Xnr2_4	Amphibian	African Clawed Frog	1459
U25993.1		Xnr3_1	Amphibian	African Clawed Frog	1634
BC169689.1		Xnr3_2	Amphibian	African Clawed Frog	1388
BC169691.1		Xnr3_3	Amphibian	African Clawed Frog	1379
NM_001085790.1		Xnr3_4	Amphibian	African Clawed Frog	1634
NM_001088347.1		Xnr4_1	Amphibian	African Clawed Frog	1746
U79162.1		Xnr4_2	Amphibian	African Clawed Frog	1746
NM_001097061.1		Xnr5_1	Amphibian	African Clawed Frog	1606

AB219843.1		Xnr5_10	Amphibian	African Clawed Frog	1622
AB219845.1		Xnr5_11	Amphibian	African Clawed Frog	1593
AB219847.1		Xnr5_12	Amphibian	African Clawed Frog	1782
AB219848.1		Xnr5_13	Amphibian	African Clawed Frog	1634
AB219849.1		Xnr5_14	Amphibian	African Clawed Frog	1603
AB219850.1		Xnr5_15	Amphibian	African Clawed Frog	1621
AB219851.1		Xnr5_16	Amphibian	African Clawed Frog	1686
AB219852.1		Xnr5_17	Amphibian	African Clawed Frog	1616
BC169725.1		Xnr5_18	Amphibian	African Clawed Frog	1498
BC169727.1		Xnr5_19	Amphibian	African Clawed Frog	1500
AB219855.1		Xnr5_2	Amphibian	African Clawed Frog	1606
NM_001085585.1		Xnr5_20	Amphibian	African Clawed Frog	1782
BC169822.1		Xnr5_3	Amphibian	African Clawed Frog	1495
BC169824.1		Xnr5_4	Amphibian	African Clawed Frog	1495
BC169866.1		Xnr5_5	Amphibian	African Clawed Frog	1495
BC170152.1		Xnr5_6	Amphibian	African Clawed Frog	1495
AB219846.1		Xnr5_7	Amphibian	African Clawed Frog	1648
AB038133.1		Xnr5_8	Amphibian	African Clawed Frog	1589
AB219842.1		Xnr5_9	Amphibian	African Clawed Frog	1594
BC169659.1		Xnr6_1	Amphibian	African Clawed Frog	1233
BC169661.1		Xnr6_2	Amphibian	African Clawed Frog	1233
AB038134.1		Xnr6_3	Amphibian	African Clawed Frog	1137
NM_001085564.1		Xnr6_4	Amphibian	African Clawed Frog	1137
BC170314.1		Xnr6_5	Amphibian	African Clawed Frog	1137

GU256638		AxNodal_1	Amphibian	Axolotl	2109
GU256639		AxNodal_2	Amphibian	Axolotl	1559
AB212661.1		CyNodal_1	Amphibian	Newt	1616
	ENSXETG00000009008	frog_N_1	Amphibian	Western Clawed Frog	2569
	ENSXETG00000016779	frog_N_2	Amphibian	Western Clawed Frog	6722
	ENSXETG00000025789	frog_N_4	Amphibian	Western Clawed Frog	4619
	ENSXETG00000023748	frog_N_5	Amphibian	Western Clawed Frog	7819
	ENSXETG00000017442	frog_N_6	Amphibian	Western Clawed Frog	1264
NM_001016321.2		Xt_NH	Amphibian	Western Clawed Frog	1499
BC171037.1		Xtnr1_1	Amphibian	Western Clawed Frog	1466
AB093329.1		Xtnr3_1	Amphibian	Western Clawed Frog	1573
NM_001112906.1		Xtnr3_2	Amphibian	Western Clawed Frog	1573
AB093327.1		Xtnr3_3	Amphibian	Western Clawed Frog	1619
AB093328.1		Xtnr3_4	Amphibian	Western Clawed Frog	1648
NM_203533.1		Xtnr3_5	Amphibian	Western Clawed Frog	1648
XM_424385.2		chicken_NH	Bird	Chicken	875
	ENSGALG00000003209	chicken_NR_2	Bird	chicken	963
	ENSMGAG00	turkey_N	Bird	turkey	960

	00000 2207				
	ENSTN IG000 00013 237	Fugu2_1	Ray-finned Fish	Fugu (Green Spotted Puffer)	1581
	ENSTN IG000 00015 847	Fugu2_2	Ray-finned Fish	Fugu (Green Spotted Puffer)	1607
	ENSTN IG000 00005 578	Fugu2_3	Ray-finned Fish	Fugu (Green Spotted Puffer)	2545
	ENSTR UG000 00010 779	Fugu1_1	Ray-finned Fish	Fugu (Takifugu)	1704
	ENSTR UG000 00012 437	Fugu1_2	Ray-finned Fish	Fugu (Takifugu)	2505
	ENSTR UG000 00012 942	Fugu1_3	Ray-finned Fish	Fugu (Takifugu)	2659
	ENSOR LG000 00011 275	ONr1_2	Ray-finned Fish	Japanese Killifish	1320
	ENSOR LG000 00009 098	ONr2_2	Ray-finned Fish	Japanese Killifish	3986
	ENSGA CG000 00002 333	Three-spined sticklebac k_1	Ray-finned Fish	Three-spined stickleback Fish	3089
	ENSGA CG000 00008 499	Three-spined sticklebac k_2	Ray-finned Fish	Three-spined stickleback Fish	1895
	ENSGA CG000 00017 712	Three-spined sticklebac k_3	Ray-finned Fish	Three-spined stickleback Fish	1473
NM_1 39133. 1		Znr1_1	Ray-finned Fish	Zebrafish	1514
U8775 8.1		Znr1_2	Ray-finned Fish	Zebrafish	1506

NM_130966.1		Znr2_2	Ray-finned Fish	Zebrafish	1480
	ENSDARG00000014309	zebrafish_SPAW	Ray-finned Fish	Zebrafish	6958
AF056327.1		Znr1_3	Ray-finned Fish	Zebrafish	1480
	ENSDNOG00000017851	armadillo_N	Mammal	Armadillo	2373
	ENSMLUG00000015297	bat_N	Mammal	bat	7312
	ENSFCAG00000001230	cat_N	Mammal	cat	4728
XM_609225.2		Cattle_NH	Mammal	Cattle	1041
	ENSCPOG00000025772	Cavy_N	Mammal	Cavy	6526
XM_521502.2		chimpanzee_NH	Mammal	Chimpanzee	2330
XM_546146.2		Dog_NH	Mammal	Dog	1047
	ENSTTRG00000003182	Dolphin_N	Mammal	Dolphin	6642
	ENSLAFG00000021867	Elephant_N	Mammal	Elephant	7819
	ENSPVAG00000000104	Flying Fox_N	Mammal	Flying Fox	6693
	ENSOGAG00000005716	Galago_N	Mammal	Galago	9246
	ENSGGOG00000002581	Gorilla(Ape)_N	Mammal	Gorilla(Ape)	9783

	ENSEEUG00000011834	Hedgehog2_N	Mammal	Hedgehog	9015
	ENSECAG00000017055	horse_N	Mammal	horse	6898
XM_001503737.1		Horse_NH	Mammal	Horse	1047
BC039861.1		Human_Nh_1	Mammal	Human	1372
NM_018055.4		Human_NH_2	Mammal	Human	2086
BC104976.1		Human_NH_3	Mammal	Human	1284
BC112025.1		Human_NH_4	Mammal	Human	1296
BC033585.1		Human_Nh_5	Mammal	Human	1680
	ENSMEUG00000011841	kangaroo_N	Mammal	kangaroo	3017
	ENSDORG00000011307	KangarooRat_N	Mammal	Kangaroo Rat	1775
	ENSMICG00000015080	Lemur_N	Mammal	Lemur	7859
	ENSCJAG00000016288	Marmoset_N	Mammal	Marmoset	16830
XM_001108074.1		Monkey_NH_1	Mammal	Monkey	1551
XM_001108137.1		Monkey_NH_2	Mammal	Monkey	1709
BC128018		Mouse_N_1	Mammal	Mouse	1070
NM_013611.3		Mouse_N_2	Mammal	Mouse	1065
X70514.1		Mouse_N_3	Mammal	Mouse	2160

	ENSMODG00000012158	opossum_N	Mammal	opossum	6699
	ENSPPYG00000002370	orangutan_N	Mammal	Orangutan	8635
	ENSSSCG00000010269	pig_N_1	Mammal	Pig	7348
XM_001928024.1		Pig_NH_4	Mammal	Pig	1047
XM_001927851.1		Pig_Nh_5	Mammal	Pig	1047
	ENSOPRG00000015824	Pika_N	Mammal	Pika	6605
	ENSOCUG00000008685	Rabbit_N	Mammal	Rabbit	6506
NM_001106394.1		Rat_N	Mammal	Rat	2034
	ENSMMUG00000023170	rhesus monkey_N	Mammal	rhesus monkey	10345
	ENSPCAG00000015506	Rock Hyrax_N	Mammal	Rock Hyrax	6935
	ENSSARG00000000846	Shrew_N	Mammal	Shrew	7400
	ENSSTOG00000012946	squirrel_N	Mammal	Squirrel	6799
	ENSTSYG00000005226	Tarsier_N	Mammal	Tarsier	2818
AB097411.1		Lancelet_N	Leptocardii	Lancelet	2481

AY083838.1		Lancelet_Nr	Leptocardii	Lancelet	1931
NM_001078532.1		Squirt1_N	Ascidiacea	Sea Squirt	1367
AB069969.1		Squirt2_N	Ascidiacea	Sea Squirt	1676
AY442295.1		Urchin_N_1	Echinozoa	Sea Urchin	2210
DQ017963.1		Urchin_N_2	Echinozoa	Sea Urchin	2326
EU812569.1		Urchin_N_3	Echinozoa	Sea Urchin	1227
EF036514.1		Urchin_N_4	Echinozoa	Sea Urchin	1353
NM_001098449.1		Urchin_N_5	Echinozoa	Sea Urchin	1353
EU812568.1		Urchin_N_6	Echinozoa	Sea Urchin	2322
EU394708.1		Limpet_N	Gastropoda	Limpet	1471
EU394707.1		Snail_N	Gastropoda	Snail	1329

Table 4.1 125 Nodal sequences included in the phylogenetic analyses

4.2.2 Phylogeny Reconstruction

i. Choosing suitable datasets to build the phylogenetic tree

To illustrate the relationships within Nodal, dataset No.1 in Table 3.1 was accessed to build the phylogenetic tree. After choosing the dataset, before building the phylogenetic tree, various partial sequences were deleted from the dataset; had those sequences been kept in, a large number of sites would have been lost in the tree reconstruction.

ii. Saturation test

After choosing the dataset, the dataset used to build phylogenetic trees was tested for saturation. In the saturation test, the uncorrected pairwise transition (ti) and transversion (tv) distances , plotted against the pairwise total uncorrected distances, as well as the uncorrected pairwise ti distances against tv distances, were tested for DNA sequences with all three codon positions and the 1st and 2nd codon positions of the chosen dataset No.1 in Table 3.1.

iii. PhyML

After the saturation test, the chosen datasets were used to build phylogenetic trees under maximum likelihood methods by PhyML. In practice, the MtRev model was utilised for the protein sequences and the GTR model was utilised for the nucleotide sequences.

In PhyML, the model for the nucleotide sequences was set as GTR+G+I. For amino acid sequences the model was MtREV+G+I.

The analysis in Chapter 3 shows that there is not a ligand or a group of ligands in the TGF-beta superfamily that definitely constitute an out-group for Nodal. In the amino acid tree showing the phylogenetic position of Nodal within the TGF-beta superfamily (Figure 3.5), DBL-1 is nearest to Nodal with low support. In the nucleotide tree of Nodal and other ligands in the TGF-beta superfamily using the 1st and 2nd codon positions only (Figure 3.6), the group formed by GDF9 & BMP15, GDF10 & BMP3 and ADMP seems nearest to Nodal with low support. In the nucleotide tree of the TGF-beta superfamily

members using all three codon positions (Figure 3.7), Nodal is not monophyletic. As there is not a ligand or a group of ligands in the TGF-beta superfamily that definitely constitute an out-group for Nodal tree, the Nodal-only tree needs to be rooted by Nodal itself. In the analyses in Chapter 3, the group of sea slugs, snails and limpets is the most distantly related group of species within Nodal. This suggests that the group of sea slugs, snails and limpets could be chosen to be the root of the Nodal tree in the analysis in this chapter.

iv. Bootstrap

After building the phylogenetic trees in PhyML, bootstrap analyses were used to test the credibility of the results. The number of replicates of the non-parametric bootstrap analysis was set as 100 for both the amino acid sequences and the nucleotide sequences.

4.3 RESULTS

4.3.1 Saturation Test

After the dataset had been chosen, the dataset was examined for evidence of substitution saturation to ascertain whether the phylogenetic tree would be accurate.

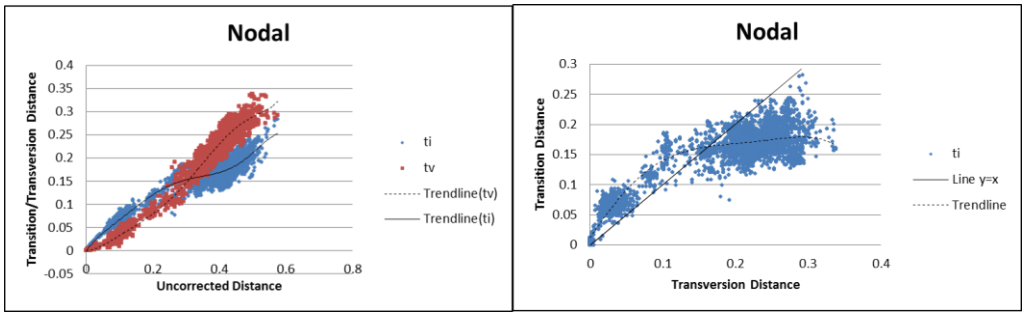

Figure 4.3 Saturation test for all three codons of Nodal.
(a)Uncorrected pairwise transition (ti) and transversion (tv) distances against the pairwise total uncorrected distances for Nodal only. (b) Uncorrected pairwise transition (ti) distances against transversion (tv) distances for Nodal only. The diagonal stands for the line y=x.

Figure 4.3 (a) shows the saturation test results of the uncorrected pairwise transition (ti) and transversion (tv) distances against the pairwise total uncorrected distances for Nodal only. In Figure 4.3 (a), the transition line forms a curve and crosses the transversion line. This suggests that the transitions are saturated. Figure 4.3 (b) shows the saturation test results of the uncorrected pairwise transition (ti) distances against the transversion (tv) distances for Nodal. In Figure 4.3 (b), most points are in the area under the line y=x, which clearly shows that the dataset is saturated. As shown in Figure 4.3, the dataset of Nodal with all three codon positions is saturated. Trees developed from the dataset with all three codons may therefore be inaccurate.

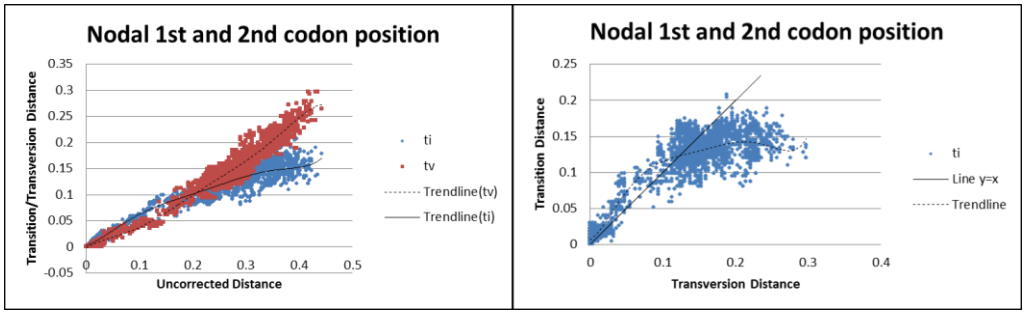

Figure 4.4 Saturation test for 1st and 2nd codon positions of Nodal.
(a)Uncorrected pairwise transition (ti) and transversion (tv) distances against the pairwise total uncorrected distances for the 1st and 2nd codon positions of Nodal only. (b) Uncorrected pairwise transition (ti) distances against transversion (tv) distances for the 1st and 2nd codon positions of Nodal only. The diagonal stands for the line y=x.

Figure 4.4 (a) shows the saturation test results of the uncorrected pairwise transition (ti) and transversion (tv) distances against the pairwise total uncorrected distances for the 1st and 2nd codon positions of Nodal. In Figure 4.4 (a), the transition line forms a curve and crosses the transversion line. Figure 4.4 (b) shows the saturation test results of the uncorrected pairwise transition (ti) distances against transversion (tv) distances for the 1st and 2nd codon positions of Nodal. Figure 4.4 (b) shows that most points are in the area under the line y=x, which suggests that saturation occurred. As shown in Figure 4.4, the dataset with the 1st and 2nd codon positions is saturated; this suggests that trees developed from this dataset with the 1st and 2nd codon positions may also be inaccurate.

The saturation test showed that the datasets of nucleotide sequences with all three codon positions and the 1st/2nd codon positions were all saturated. The tree developed from those datasets may be inaccurate. Nevertheless, these datasets were still used to build phylogenetic trees for two reasons: Firstly, a

phylogenetic tree is still needed to show the relationships of Nodal. Secondly, the alignment used to make up the datasets is the best one that can be provided. However, the problems of saturated data can be reduced to some extent by using a more complex likelihood model (Farrell, 2011).

4.3.2 Phylogenetic trees

After the saturation test, with the aim of examining the relationships within Nodal, phylogenetic analyses based on dataset No.1 in Table 3.1 were carried out. In the analysis, phylogenetic trees were developed based on protein sequences (Figure 4.5), the 1st/2nd codon position (Figure 4.6) and all 3 codon positions of the DNA sequences (Figure 4.7) through the maximum likelihood (ML) method to determine the relationship among Nodal genes from different species.

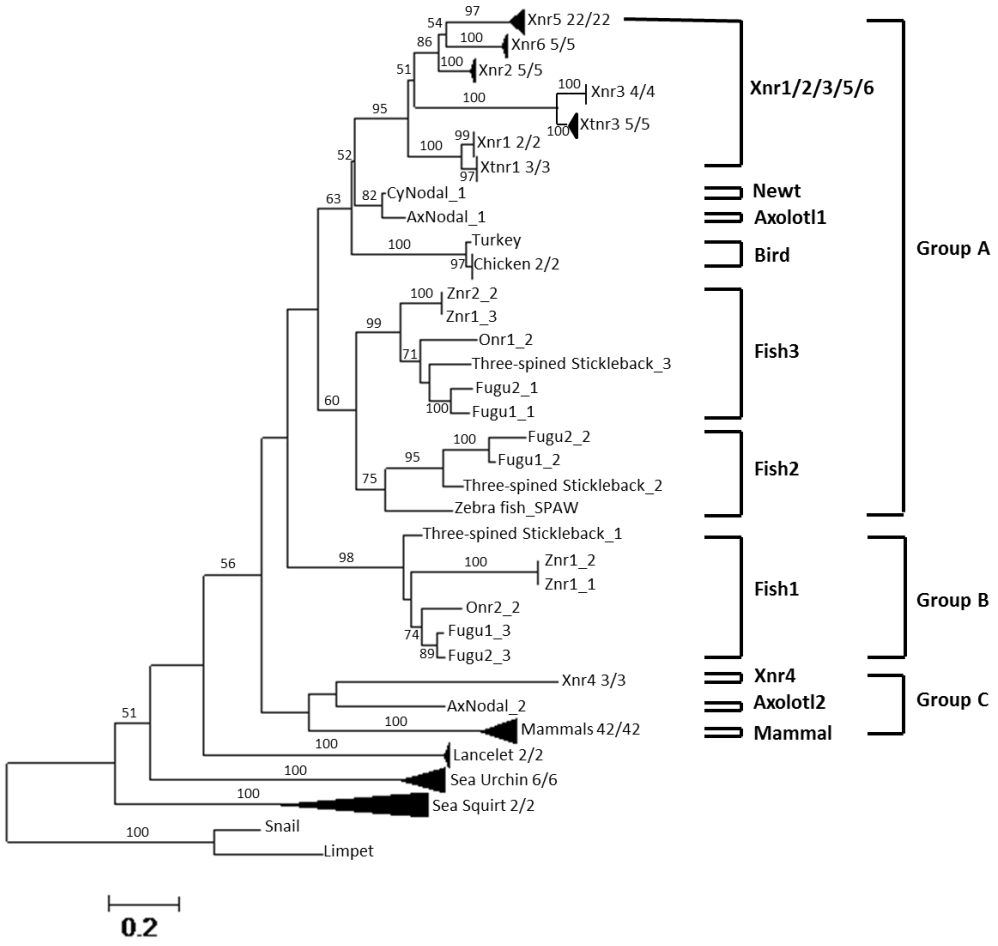

Figure 4.5 Maximum likelihood amino acid phylogenetic tree of Nodal
This tree is built based on 137 amino acid sites. The scale bar corresponds to 20 changes per 100 amino acid positions. The numbers on branches represent the bootstrap support for that branch based on 100 bootstrap replicates. Only bootstrap values higher than 50% are shown. The tree is rooted on snail and limpet Nodals shown in the base of Nodal tree in earlier analysis (chapter 3).

In the phylogenetic tree built from amino acid sequences of Nodal (Figure 4.5), Nodal falls in three main groups. Group A is supported with a bootstrap number lower than 50%. Group B, which contains Fish 1, is supported with a bootstrap number of 98%. In this figure, Group B stays with Group A with a support lower than 50%. Xnr4, Axolotl 2 and the Mammal Nodal form a Group C with a support lower than 50%. The base of the tree is lancelet, sea

urchin, sea squirt, snail and limpet. The tree is rooted on the group of snail and limpet.

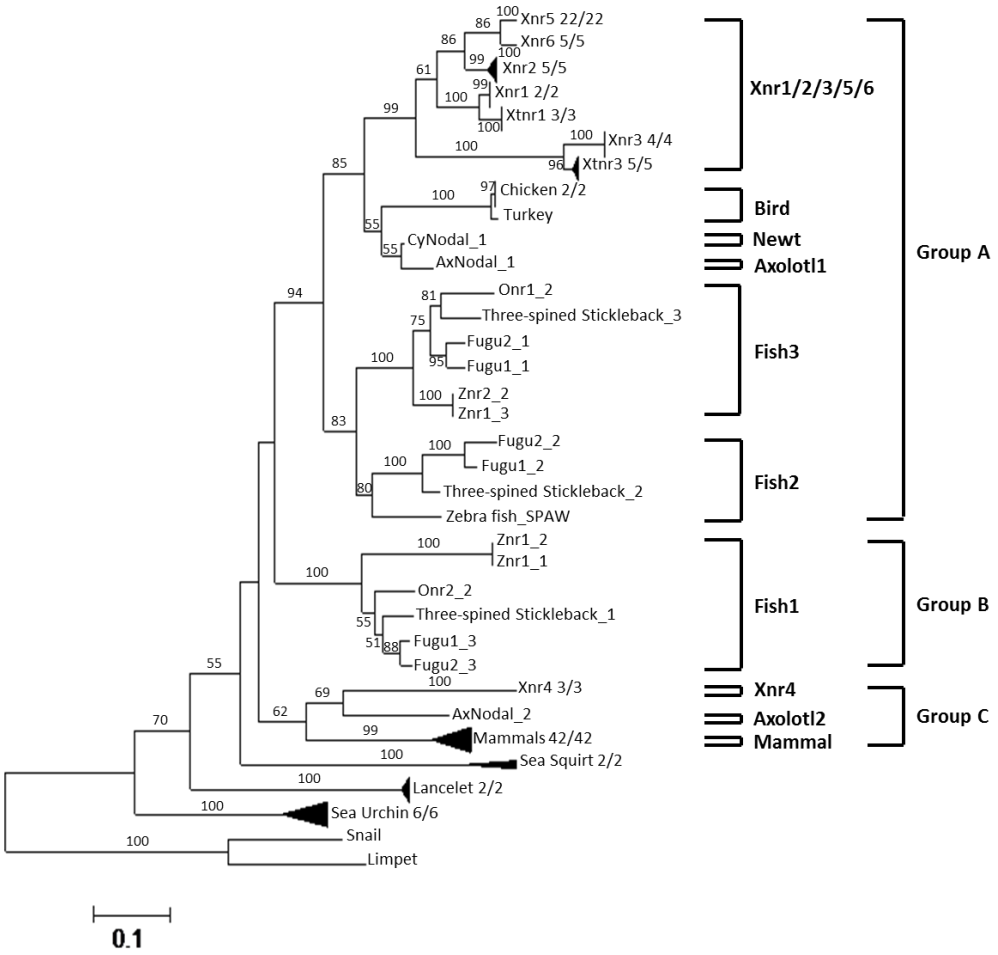

Figure 4.6 Maximum likelihood 1st and 2nd codon position phylogenetic tree of Nodal

This tree is built based on 276 nucleotide sites by using the 1st and 2nd codon positions only. The scale bar corresponds to 10 changes per 100 nucleotide positions. The numbers on branches represent the bootstrap support for that branch based on 100 bootstrap replicates. Only bootstrap values higher than 50% are shown. The tree is rooted on the snail and limpet Nodals shown in the base of Nodal tree in earlier analysis (chapter 3).

In the tree built from nucleotide sequences with the 1st and 2nd codon positions (Figure 4.6), Nodal falls into three main groups. Group A is supported with a high bootstrap value of 94%. The bootstrap value to support

Group B is 100%. In Figure 4.6, Group B also stays with Group A with a support lower than 50%. Group C contains the same members as shown in Figure 4.5, with a support of 62%. The base of the tree is lancelet, sea urchin, sea squirt, snail and limpet. The tree is rooted on the group of snail and limpet.

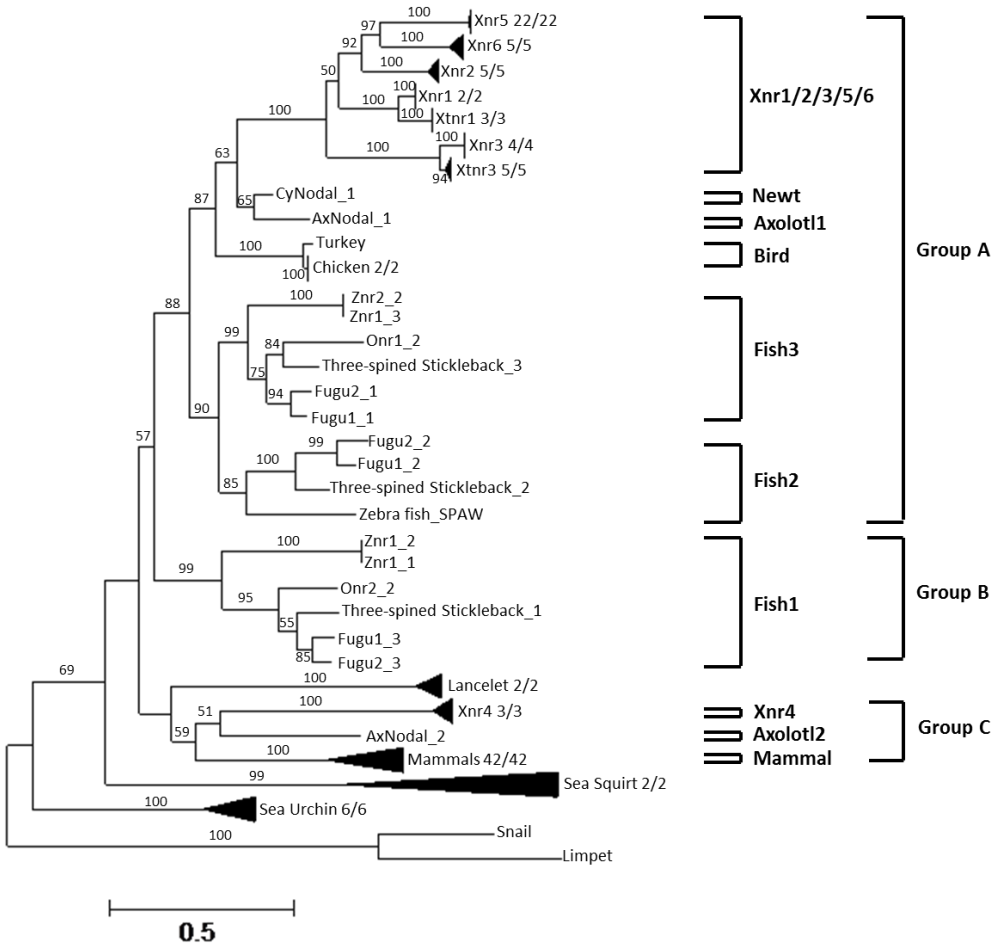

Figure 4.7 Maximum likelihood nucleotide phylogenetic tree of Nodal

This tree is built based on 414 nucleotide sites by using all three codon positions. The scale bar corresponds to 50 changes per 100 nucleotide positions. The numbers on branches represent the bootstrap support for that branch based on 100 bootstrap replicates. Only bootstrap values higher than 50% are shown. The tree is rooted on the snail and limpet Nodals shown at the base of the Nodal tree in earlier analysis (chapter 3).

In the tree of nucleotide sequences with all three codon positions (Figure 4.7), Nodal falls in three main groups. Group A is supported with a high bootstrap

value of 88%. The support value of Group B is 99%. In this figure, Group B stays with Group A with a bootstrap value of 57%. In Figure 4.7, Group C contains the same members as shown in previous two figures with a support lower than 50%. The base of the tree is sea urchin, sea squirt, snail and limpet. The tree is rooted on the group of snail and limpet.

From the 3 figures it can be seen that there is evidence to indicate that Xnr1/2/3/5/6, newts, axolotl 1, fish 2 and 3 and birds form a group which is shown as group A in the tree figures. It is probable that Group B may stay with Group A with a low support. Xnr4, axolotl 2 and mammals form a Group C. There is only one copy of Nodal gene in mammals, birds, sea squirts, sea urchins and gastropoda, but there are two or more copies in amphibians and fish. In the three figures, the branch is separated when the species evolve from lancelet to vertebrates.

4.4 DISCUSSION & CONCLUSION

4.4.1 The relationships within Nodal

It can be seen from the results that there are six copies of Nodal in the frog, two copies in urodele amphibians, three copies in fishes and only one copy in mammals, birds, sea squirts, lancelets, sea urchins, snails and limpets. Among them, birds form a group which is called group A in the result section, with one copy of amphibians, one group of copies of frog (which is Xnr1/2/3/5/6) and two copies of fish with a valid support. However, previous work of Fan showed two groups of Nodal (Fan & Dougan, 2007). But in this project a

converse result shows that Group B (Fish 1) stays with Group A instead of Group C.

In Fan's paper (Fan & Dougan, 2007), the authors suggest that Group B along with group C forms a group which contains Fish 3, Xnr4 and the Mammal. In Kuraku's study (Kuraku & Kuratani, 2011), Kuraku's Nodal tree looks like Nodal trees in this project (Figure 4.5, Figure 4.6 and Figure 4.7). They suggest that Group B departs from group C. Although this structure has limited support in this project, the structure is supported well in Kuraku's study (Fan & Dougan, 2007. Kuraku & Kuratani, 2011).

4.4.2 Duplication of Nodal in different species

As outlined in Chapter 1.2.4, the number of Nodal genes in different species is varied. In some species, such as mammals or birds, one of the loci may be deleted. In some species, such as amphibians, fishes or lizards, there may be several copies of Nodal. In this project, there exists some evidence to support that there is duplication within Nodal. Although there are low bootstrap values for group B, there is valid support for the assertion that group A stays alone and does not combine well with the other copies of vertebrates' Nodal sequences. It can be seen that duplication occurred when vertebrates evolved from Urochordata (which is the sea urchin in this chapter). In most vertebrates, Nodal genes can be grouped into two groups. However, deletion occurs in birds and mammals, so there is only one copy of the mammal Nodal and the bird Nodal.

4.4.3 Further examination of the Nodal locus in fish groups

In the support material for The Genome of the Western Clawed Frog *Xenopus tropicalis* (Hellsten, 2010), the author indicates that there are two Nodal loci in vertebrates. One is between eif4ebp2 and ash2l, and the other is between eif4ebp1 and paladin. The bird loses the Nodal locus adjacent to paladin, while the mammal loses the Nodal locus adjacent to ash2l. In the analyses in this chapter, Group C contains Nodal genes near paladin and Group A contains Nodal genes near ash21.

As discussed in Chapter 4.4.1, the researcher of this project initially envisaged Group B (Fish 1) being with Group C, as Fan described in his study. However, the result is that Group B seems to be much closer to Group A than Group C. Thus, it is particularly interesting to look into the 3 fish groups.

As checked in Ensembl, Fish 1 is located near paladin, the locus of the Nodal gene of Group Fish 2 is between DGUOK and ANK1 (1of2) and the Group Fish 3 is near CLDN23 (usually between eif4ebp1/ANK1 (2of2) and CLDN23). The loci of Fish 2 and Fish 3 are not far apart, but the location of Fish 1 is far from Fish 2 and 3. The loci of Nodal of those fish groups seems to suggest the Group B may be with Group C, because their loci are all located around paladin. Nevertheless, this hypothesis is not supported by the phylogeny results in this chapter.

4.4.4 Future work

For the analysis of Nodal phylogeny, since the sequences are well aligned and the sites are chosen carefully, the trees shown in this chapter may be the best results based on that number of sequences. To remove snails, limpets, sea urchins and lancelets to construct a vertebrates' Nodal tree and to root that tree on the sea squirt may be worthwhile in order to view the relationship specifically within the vertebrates' Nodal. Nevertheless, it seems unlikely that it would make a great improvement to the structure and bootstrap support by simply removing 10 sequences.

Conversely, it may well be interesting to use one sequence from each class to build a Nodal tree. Nevertheless, as the amphibian group (which contains newts and axolotl 1 in this chapter) and the Fish 2 group are not so strongly supported, it is doubtful whether that tree would show a true picture.

Because Lefty has the most similar function to Nodal and ADMP has the most similar sequence structure to Nodal, it may also be interesting to download as many sequences of ADMP and Lefty as possible in order to construct an ADMP tree and a Lefty tree. Then, the relationship of species within ADMP and Lefty can be determined, and that result can be compared with the Nodal tree to ascertain whether they have the same situation as Nodal, for example duplication and deletion during evolution.

4.5 CONCLUSION

In conclusion, according to the phylogenetic analyses presented in this chapter, there are two different types of Nodals in vertebrates. Duplication occurred when vertebrates evolved from Urochordata. Furthermore, deletion occurred in birds and mammals.

CHAPTER 5 SUMMARY

Nodal is a ligand of the TGF-beta superfamily. It has the function of determining the left-right axis and inducing the endoderm and mesoderm. Nodal signals can also act as morphogens (Schier, 2009). In Herpin's review, the TGF-beta superfamily is divided into four subfamilies: the DVR subfamily, the activin/inhibin subfamily, the TGF-beta subfamily and a group of divergent members. Furthermore, Nodal is in the DVR subfamily (Herpin et al. 2004). As Hellsten described in his study (Hellsten, 2010), there are two loci of Nodal in vertebrates, and deletion occurs later in birds and mammals. Previous studies show two different views of the relationships within Nodal. One suggests that the fishes are divided into two groups, as are most other vertebrate species. The other suggests that all fish Nodal genes are in the group in which the bird Nodal is located (Fan & Dougan, 2007; Kuraku & Kuratani, 2011).

In this project, the phylogeny of the TGF-beta superfamily was investigated further, using 407 taxa based on nucleotide sequences and amino acid sequences. This study demonstrates the monophyly of Nodal, but its neighbour ligand or ligand group is nonetheless uncertain. According to the phylogenetic analyses presented in Chapter 3, Nodal seems to be a ligand within the DVR subfamily of the TGF-beta superfamily, as Herpin demonstrated in his study, but the bootstrap value to support it is limited (Herpin et al. 2004).

In this project, the phylogeny of Nodal was also investigated, using 131 taxa across 46 species based on nucleotide sequences and amino acid sequences.

This study demonstrates that the fish sequences are all in the Group A in which the bird Nodal is located, but the support is not particularly valid. In addition, when checked by gene loci, Fish 1 group seemed to be within Group C because their loci were all near the paladin gene. According to the phylogenetic analyses presented in Chapter 4, duplication occurred when vertebrates evolved from Urochordata. In addition, deletion occurred in birds and mammals.

There are several limitations in this project. Firstly, this research was limited to include Nodal genes from all species, due to the need to obtain a certain amount of sites. Partial gene sequences are the biggest limitation to the data collection. For example, some representative Nodal genes such as lizard Nodal genes were excluded in the test of Nodal-only tree because they were partial sequences. Secondly, this project was limited to include all TGF-beta superfamily ligands to show the whole view of the relationships between the members of the whole superfamily. This was also due to the need to obtain a certain amount of sites. When trying to include all members, only very limited sites could be used in the phylogeny. Thirdly, this project tries to include all Nodal sequences found online, therefore there may be several sequences from one species. For example, there were 5 human Nodal sequences included in this project. If one was to choose one sequence from each group of species based on the suggested tree in chapter 4, more sites would be provided and a more valid supported tree may exist. Finally, this research has demonstrated the relationships of species within Nodal and Nodal with other ligands in the DVR subfamily. It is recommended to further research a comparison of Nodal

with those ligands that have sequence similarity (such as ADMP) or functional similarity (such as Lefty).

References

Airaksinen M, Saarma M (2002). The GDNF family: signalling, biological functions and therapeutic value. Nat Rev Neurosci. 2002 May; 3(5):383-94.

Alexander F. Schier. Nodal Morphogens. Cold Spring Harb Perspect Biol. 2009 November; 1(5): a003459. doi: 10.1101/cshperspect.a003459

Amaury Herpin, Christophe Lelong, Pascal Favrel (2004). Transforming growth factor-beta-related proteins: an ancestral and widespread superfamily of cytokines in metazoans. Developmental and Comparative Immunology 28 (2004) 461–485.

Antonio Colavita, Srikant Krishna, Hong Zheng, Richard W. Padgett, Joseph G. Culotti (1998). Pioneer axon guidance by UNC-129, a C. elegans TGF-beta.. SCIENCE VOL 281 31 JULY 1998.

B.I. Sun, S.M. Bush, L.A. Collins-Racie, E.R. LaVallie, E.A. DiBlasio-Smith, N.M. Wolfman, J.M. McCoy and H.L. Sive (1999). Derriere: a TGF-beta family member required for posterior development in Xenopus. Development 126, 1467-1482 (1999).

Baker K, Holtzman NG, Burdine RD. (2008). Direct and indirect roles for Nodal signaling in two axis conversions during asymmetric morphogenesis of the zebrafish heart. Proc Natl Acad Sci U S A. 2008 Sep 16; 105(37):13924-9. Epub 2008 Sep 10.

Behringer RR. (1994). The in vivo roles of müllerian-inhibiting substance. Curr Top Dev Biol. 1994; 29:171-87.

Benson D, et al. (2009). "GenBank". Nucleic Acids Research 37 (Database): D26–D31. doi:10.1093/nar/gkn723. PMC 2686462. PMID 18940867

Caterina Bianco, Maria Cristina Rangel, Nadia P. Castro, Tadahiro Nagaoka, Kelly Rollman, Monica Gonzales, David S. Salomon. Role of Cripto-1 in Stem Cell Maintenance and Malignant Progression. Am J Pathol. 2010 August; 177(2): 532–540. doi: 10.2353/ajpath.2010.100102

Cavalli-Sforza LL, Edwards AW. (1967). Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet. 1967 May;19(3 Pt 1):233-57.

Chapman SC, Bernard DJ, Jelen J, Woodruff TK.Properties of inhibin binding to betaglycan, InhBP/p120 and the activin type II receptors.Mol Cell Endocrinol. 2002 Oct 31; 196(1-2):79-93.

Chen D, Zhao M, Mundy GR. (2004). Bone morphogenetic proteins. Growth Factors. 2004 Dec; 22(4):233-41..

Chen YG, Liu F, Massague J (1997) Mechanism of TGFbeta receptor inhibition by FKBP12. EMBO J 16: 3866–3876.

Chen, Pinjian. Zoology. Science Press China (2001).

Christien Weenen,, Joop S.E. Laven, Anne R.M. von Bergh, Mark Cranfield, Nigel P. Groome, Jenny A. Visser, Piet Kramer, Bart C.J.M. Fauser and Axel P.N. Themmen. Anti-Müllerian hormone expression pattern in the human ovary: potential implications for initial and cyclic follicle recruitment. Molecular Human Reproduction, Vol. 10, No. 2, pp. 77-83, 2004.

Christophe Lelong, Michel Mathieu, Pascal Favrel (2000). Structure and expression of mGDF, a new member of the transforming growth factor-

beta superfamily in the bivalve mollusc Crassostrea gigas. European Journal of Biochemistry Volume 267, Issue 13, pages 3986–3993, July 2000.

Claudine Montgelard, Ellen Forty, Véronique Arnal and Conrad A Matthee, 2008. Suprafamilial relationships among Rodentia and the phylogenetic effect of removing fast-evolving nucleotides in mitochondrial, exon and intron fragments. BMC Evolutionary Biology 2008, 8:321 doi: 10.1186/1471-2148-8-321

Collins DW, Jukes TH (April 1994). "Rates of transition and transversion in coding sequences since the human-rodent divergence". Genomics 20 (3): 386–96.

Derynck R, Jarrett JA, Chen EY, Eaton DH, Bell JR, et al. (1985) Human transforming growth factor-beta complementary DNA sequence and expression in normal and transformed cells. Nature 316: 701–705

Derynck R, Zhang YE (2003) Smad-dependent and Smad-independent pathways in TGF-beta family signaling. Nature 425: 577–584.

Edgar RC (2004). "MUSCLE: a multiple sequence alignment method with reduced time and space complexity". BMC Bioinformatics 5 (1): 113. Doi:10.1186/1471-2105-5-113. PMC 517706. PMID 15318951.

Farrell, 2011. Saturation points & weighting. Harvard Entomology & the Farrell Lab.

Felsenstein, J. (2004). Inferring Phylogenies. Sinauer Associates, Sunderland, Mass.

Flicek P, Amode MR, Barrell D, et al. (November 2010). "Ensembl 2011". Nucleic Acids Res 39 (Database issue): D800–D806. doi:10.1093/nar/gkq1064. PMC 3013672. PMID 21045057

Garcia-Fernàndez J, D'Aniello S, Escrivà H (2007). "Organizing chordates with an organizer". Bioessays 29 (7): 619–24. doi:10.1002/bies.20596

Gemma Swiers, Yi-Hsien Chen, Andrew D. Johnson, Matthew Loose (2010). Evolution of Developmental Control Mechanisms A conserved mechanism for vertebrate mesoderm specification in urodele amphibians and mammals. Developmental Biology 2010 (2010) 138–152.

Graur, D. and Li, WH (2000). Fundementals of molecular evolutioin, 2nd edn. Sinauer Associates, Sunderland.

Haramoto Y, Tanegashima K, Onuma Y, Takahashi S, Sekizaki H, Asashima M. (2004). Xenopus tropicalis nodal-related gene 3 regulates BMP signaling: an essential role for the pro-region.. Dev Biol. 2004 Jan 1;265(1):155-68

Helde, K.A., and Grunwald, D.J. (1993). The DVR-1 (Vg1) transcript of zebrafish is maternally supplied and distributed throughout the embryo. Dev Biol 159, pp. 418–26.

Heldin CH, Miyazono K, ten Dijke P (1997). TGF-beta signalling from cell membrane to nucleus through SMAD proteins. Nature, 390 (6659): 465–71(December 1997). doi:10.1038/37284. .

Henrik Bengtsson (2001). Bone Morphogenetic Protein Receptors in the Nervous System: Neurotrophic Functions with Emphasis on Catecholaminergic Neurons. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1104.

Hiroshi Hamada,Chikara Meno,Daisuke Watanabe, Yukio Saijoh (2002). Establishment of vertebrate left-right asymmetry. Nature Reviews Genetics 3, 103-113 (February 2002).

Hiroshi Hanafusa, Norihisa Masuyama, Morioh Kusakabe, Hiroshi Shibuya & Eisuke Nishida (2000). The TGF-beta family member derrière is involved in regulation of the establishment of left–right asymmetry. EMBO reports 1, 1, 32–39 (2000).

Ito Y, Oinuma T, Takano K, Komazaki S, Obata S, Asashima M. (2006). CyNodal, the Japanese newt nodal-related gene, is expressed in the left side of the lateral plate mesoderm and diencephalon. Gene Expr Patterns. 2006 Mar;6(3):294-8. Epub 2005 Dec 27

Itoh F, Asao H, Sugamura K, Heldin CH, ten Dijke P, Itoh S (2001). Promoting bone morphogenetic protein signalling through negative regulation of inhibitory Smads. EMBO J. 20 (15): 4132–42 (August 2001). doi:10.1093/emboj/20.15.4132. .

Joan Massagué, Joan Seoane and David Wotton (2005). Smad transcription factors. Genes Dev. 2005 19: 2783-2810.

K Tsuchida. (2008) Myostatin inhibition by a follistatin-derived peptide ameliorates the pathophysiology of muscular dystrophy model mice. Acta Myol. 2008 July; 27(1): 14–18.

Kalinovsky A, Boukhtouche F, Blazeski R, Bornmann C, Suzuki N, et al. (2011) Development of Axon-Target Specificity of Ponto-Cerebellar Afferents. PLoS Biol 9(2): e1001013. doi:10.1371/journal.pbio.1001013

Klein SL, Strausberg RL, Wagner L, Pontius J, Clifton SW, Richardson P. (2002). Genetic and genomic tools for Xenopus research: The NIH Xenopus initiative. Dev Dyn. 2002 Dec; 225(4):384-91

Kosakovsky Pond SL, Mannino FV, Gravenor MB, Muse SV, Frost SD. (2007). Evolutionary Model Selection with a Genetic Algorithm: A Case Study Using Stem RNA. Mol Biol Evol. 2007 Jan; 24(1):159-70. Epub 2006 Oct 12.

Lee SJ. (2007). Sprinting without myostatin: a genetic determinant of athletic prowess. Trends Genet. 2007 Oct; 23(10):475-7. Epub 2007 Sep 19.

Liangyi Xue, Kaixian Qian, Hongqin Qian, Lu Li, Qiaoyi Yang and Mingyun Li (2006). Molecular Cloning and Characterization of the Myostatin Gene in Croceine Croaker, Pseudosciaena crocea. Mol Biol Rep. 2006 Jun; 33(2):129-

Liu X, Sun Y, Weinberg RA, Lodish HF (2001) Ski/Sno and TGF-beta signaling. Cytokine Growth Factor Rev 12: 1–8.

M. E. Abbott and J. C. Meakins (1915). "On the differentiation of two forms of congenital dextrocardia". Bulletin of the International Association of Medical Museums: 134–138.

M. Moos, S. Wang and M. Krinks (1995). Anti-dorsalizing morphogenetic protein is a novel TGF-beta homolog expressed in the Spemann organizer. Development 121, 4293-4301 (1995).

Marı́a R. Ponce, Jose L. Micol, Kevin J. Peterson and Eric H. Davidson (1999). Molecular Characterization and Phylogenetic Analysis of SpBMP5–7, a New Member of the TGF-beta superfamily Expressed in Sea Urchin Embryos. Mol. Biol. Evol. 16(5):634–645.1999.

Marketa J. Zvelebil, Jeremy O. Baum, 2008. Understanding Bioinformatics. Garland Science, 2008. ISBN: 0815340249, 9780815340249

Matt Crooka, Fiona J. Thompsona, Warwick N. Grantb and Mark E. Viney (2005). Daf-7 and the development of Strongyloides ratti and Parastrongyloides trichosuri. Molecular and Biochemical Parasitology Volume 139, Issue 2, February 2005, Pages 213-223.

Minh Nguyen, Louise Parker, Kavita Arora (2000). Identification of maverick, a novel member of the TGF-beta superfamily in Drosophila.. Mechanisms of Development 95 (2000) 201-206.

Morgan, T. H. (2001). Regeneration. Oxford: Oxford University Press.

Morrison, D.A. 2006. Phylogenetic analyses of parasites in the new millennium. Advances in Parasitology 63: 1-124.

Nei M & Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Ongkar Khalsa, Jung-won Yoon, Sonia Torres-Schumann and Kristi A. Wharton (1998). TGF-beta/BMP superfamily members, Gbb-60A and Dpp, cooperate to provide pattern information and establish cell identity in the Drosophilawing. Development 125, 2723-2734 (1998).

Onichtchouk D, Chen YG, Dosch R, Gawantka V, Delius H, et al. (1999) Silencing of TGF-beta signaling by the pseudoreceptor BAMBI. Nature 401: 480–485.

Pang K, Ryan JF, Baxevanis AD, Martindale MQ (2011) Evolution of the TGF-beta Signaling Pathway and Its Potential Role in the Ctenophore, Mnemiopsis leidyi. PLoS ONE 6(9): e24152. doi:10.1371/journal.pone.0024152

Patrick C.H. Lo, Manfred Frasch (1999). Sequence and expression of myoglianin, a novel Drosophila gene of the TGF-beta superfamily. Mechanisms of Development 86 (1999) 171-175.

Pei-Yu Wang, Anna Protheroe, Andrew N. Clarkson, Floriane Imhoff, Kyoko Koishi, and Ian S. McLennan (2009). Müllerian inhibiting substance contributes to sex-linked biases in the brain and behavior. Proc Natl Acad Sci U S A. 2009 April 28; 106(17): 7203–7208.

Peter ten Dijke & Helen M. Arthur (2007). Extracellular control of TGFbold beta signalling in vascular development and disease. Nature Reviews

Philippe H, Sörhannus U, Baroin A, Perasso R, Gasse F, Adoutte A (1994). Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J Evol Biol 1994, 7:247-265.

Powers SE, Taniguchi K, Yen W, Melhuish TA, Shen J, Walsh CA, Sutherland AE, Wotton D. (2010). Tgif1 and Tgif2 regulate Nodal signalling and are required for gastrulation. Development. 2010 Jan; 137(2):249-59.

Rik Derynck, Ying Zhang, and Xin-Hua Feng (1998). Smads: Transcriptional Activators of TGF-b Responses. Cell, Vol. 95, 737–740, December 11, 1998, Copyright (C) 1998 by Cell Press.

Rodríguez F, Oliver JL, Marín A, Medina JR. (1990). The general stochastic model of nucleotide substitution. J Theor Biol. 1990 Feb 22; 142(4):485-501.

Rosa Serra, Chenbei Chang (2003). TGF-beta signalling in human skeletal and patterning disorders. Birth Defects Research Part C: Embryo Today: Reviews, Volume 69, Issue 4, pages 333–351, December 2003.

Saarma M. (2000). GDNF - a stranger in the TGF-beta superfamily? Eur J Biochem. 2000 Dec; 267(24):6968-71.

Saitou N and Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987 Jul;4(4):406-25.

Shi Y, Hata A, Lo RS, Massagué J, Pavletich NP (1997). A structural basis for mutational inactivation of the tumour suppressor Smad4. Nature 388 (6637): 87–93 (July 1997). Doi: 10.1038/40431.

Shigehiro Kuraku and Shigeru Kuratani (2011). Genome-Wide Detection of Gene Extinction in Early Mammalian Evolution. Genome Biol. Evol. 3:1449–1462. doi:10.1093/gbe/evr120 Advance Access publication November 17, 2011.

Smith, S.W., Overbeek, R., Woese, C.R., Gilbert, W. and Gillevet, P.M. 1994. The genetic data environment, an expandable GUI for multiple sequence analysis. Computer Applications in the Biosciences 10: 671-675.

Soroldoni D, Bajoghli B, Aghaallaei N, Czerny T. (2007). Dynamic expression pattern of Nodal-related genes during left-right development in medaka.. Gene Expr Patterns. 2007 Jan;7(1-2):93-101. Epub 2006 Jun 6

Strandberg AKK, Salter LA. A comparison of methods for estimating the transition: transversion ratio from DNA sequences. Mol Phylogenet Evol. 2004; 32:495–503. doi: 10.1016/j.ympev.2004.01.013.

Stuart J. Newfeld, Robert G. Wisotzkey and Sudhir Kumar (1999). Molecular Evolution of a Developmental Pathway: Phylogenetic Analyses of Transforming Growth Factor-ß Family Ligands, Receptors and Smad Signal Transducers. Genetics, Vol. 152, 783-795, June 1999, Copyright © 1999.

Swofford, D., 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA

Takahashi S, Yokota C, Takano K, Tanegashima K, Onuma Y, Goto J, Asashima M. (2000). Two novel nodal-related genes initiate early inductive events in Xenopus Nieuwkoop center.. Development. 2000 Dec; 127(24):5319-29

Tavaré S (1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences (American Mathematical Society) 17: 57–86.

Tsigenopolous, C.S., Rab, P. Naran, D. and Berrebi, P. 2002. Multiple origins of polyploidy in thephylogeny of southern African barbs (Cyprinidae) as inferred from mtDNA markers. Heredity 88: 466-473.

Uffe Hellsten, Richard M. Harland, Michael J. Gilchrist, David Hendrix, Jerzy Jurka, Vladimir Kapitonov, Ivan Ovcharenko, Nicholas H. Putnam, Shengqiang Shu, Leila Taher, Ira L. Blitz, Bruce Blumberg, Darwin S. Dichmann, Inna Dubchak, Enrique Amaya, John C. Detter, Russell Fletcher, Daniela S. Gerhard, David Goodstein, Tina Graves, Igor V. Grigoriev, Jane Grimwood, Takeshi Kawashima, Erika Lindquist, Susan M. Lucas, Paul E. Mead, Therese Mitros, Hajime Ogino, Yuko Ohta, Alexander V. Poliakov, Nicolas Pollet, Jacques Robert, Asaf Salamov, Amy K. Sater, Jeremy Schmutz, Astrid Terry, Peter D. Vize, Wesley C. Warren, Dan Wells, Andrea Wills, Richard K. Wilson, Lyle B. Zimmerman, Aaron M. Zorn, Robert Grainger, Timothy Grammer, Mustafa K. Khokha, Paul M. Richardson, and Daniel S. Rokhsar (2010). Supporting Online Material for The Genome of the Western

Clawed Frog Xenopus tropicalis. Science 30 April 2010: Vol. 328. no. 5978, pp. 633 - 636

Van Zonneveld P, Scheffer GJ, Broekmans FJ, Blankenstein MA, de Jong FH, Looman CW, Habbema JD, te Velde ER. (2003). Do cycle disturbances explain the age-related decline of female fertility? Cycle characteristics of women aged over 40 years compared with a reference population of young women. Hum Reprod. 2003 Mar; 18(3):495-501.

W. Newton Suter, 2012. Chapter 12: Qualitative Data, Analysis, and Design. Introduction to Educational Research-A Critical Thinking Approach Second Edition. SAGE Publications, Inc. 2012

Wang H, Jiang JY, Zhu C, Peng C, Tsang BK. (2006). Role and Regulation of Nodal/Activin Receptor-Like Kinase 7 Signaling Pathway in the Control of Ovarian Follicular Atresia. Mol Endocrinol. 2006 Oct; 20(10):2469-82. Epub 2006 May 18.

Wang, J., Tokarz, R., Savage-Dunn, C. (2002). The expression of TGFbeta signal transducers in the hypodermis regulates body size in C. elegans. . Development 129, 4989-4998 (2002).

Wu JW, Hu M, Chai J, Seoane J, Huse M, Li C, Rigotti DJ, Kyin S, Muir TW, Fairman R, Massagué J, Shi Y (2001). Crystal structure of a phosphorylated Smad2. Recognition of phosphoserine by the MH2 domain and insights on Smad function in TGF-beta signalling. Cell 8 (6): 1277–89 (December 2001). Doi: 10.1016/S1097-2765(01)00421-X.

Xiang Fan & Scott T. Dougan (2007). The evolutionary origin of nodal-related genes in teleosts. Dev Genes Evol (2007) 217:807–813 DOI 10.1007/s00427-007-0191-y.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol. 1998 Dec; 15(12):1600-11.

Ye-Guang Chen, Qiang Wang, Shi-Lung Lin, C. Donald Chang, Jody Chung and Shao-Yao Ying (2006). Activin Signalling and Its Role in Regulation of Cell Proliferation, Apoptosis, and Carcinogenesis. Exp. Biol. Med. 2006; 231:534-544.

Zhang, Xiaolian (2008). Medical Immunology. Wuhan University Press, 2008.1 P61.

Zhou X, Sasaki H, Lowe L, Hogan BL, Kuehn MR (1993). Nodal is a novel TGF-beta-like gene expressed in the mouse node during gastrulation.. Nature. 1993 Feb 11; 361(6412):543-7.

Zhu H, Kavsak P, Abdollah S, Wrana JL, Thomsen GH (1999) A SMAD ubiquitin ligase targets the BMP pathway and affects embryonic pattern formation. Nature 400: 687–693.

APPENDIX

APPENDIX 1. SEQUENCE IMFORMATION

Change Name	NCBI ID	NCBI classification	EMBL ID	EMBL classification
Squirt_N	AB069969.1	nodal		
Lancelet_Na	AB097411.1	nodal		
Pig_Nc	AM072821.1	nodal		
Urchin_Na	AY442295.1	nodal		
Mouse_Na	BC128018	nodal		
Urchin_Nb	DQ017963.1	nodal		
Urchin_Nd	EF036514.1	nodal		
Snail_N	EU394707.1	nodal		
Limpet_N	EU394708.1	nodal		
Urchin_Nf	EU812568.1	nodal		
Urchin_Nc	EU812569.1	nodal		
Urchin_Ne	NM_001098449.1	nodal		
Mouse_Nb	NM_013611.3	nodal	ENSMUSG00000037171	Nodal
Mouse_Nc	X70514.1	nodal		
Human_Nhme	BC033585.1	Nodal		
Human_Nhma	BC039861.1	Nodal		
Human_NHmc	BC104976.1	Nodal		
Human_NHmd	BC112025.1	Nodal		
Xt_NHm	NM_001016321.2	Nodal homolog		

Rat_N	NM_001106394.1	Nodal	ENSRNOG00000000556	Nodal
Human_NHmb	NM_018055.4	Nodal	ENSG00000156574	Nodal
Horse_NHm	XM_001503737.1	Nodal homolog		
Pig_Nhme	XM_001927851.1	Nodal homolog		
Pig_NHmd	XM_001928024.1	Nodal homolog		
finch_NHm	XM_002194155.1	Nodal	ENSTGUG00000004739	Nodal
chicken_NHm	XM_424385.2	Nodal		
Dog_NHm	XM_546146.2	Nodal homolog	ENSCAFG00000014052	Nodal
Cattle_NHm	XM_609225.2	Nodal homolog		
Monkey_Nhma	XM_001108074.1	nodal precursor		
Monkey_Nhmb	XM_001108137.1	nodal precursor		
chimpanzee_NHm	XM_521502.2	nodal	ENSPTRG00000002592	Nodal
SeaSlug_Nlike	FJ616286.1	nodal like		
Xnr5ae	AB038133.1	xnr5		
Xnr6c	AB038134.1	xnr6		
Xtnr3c	AB093327.1	xtnr3		
Xtnr3d	AB093328.1	xtnr3		
Xtnr3a	AB093329.1	xtnr3		
Newt_CyNodalb	AB114684.1	CyNodal		
Onr1	AB116041.1	ONr1		
Onr2	AB116642.1	ONr2		

Newt_CyNodala	AB212661.1	CyNodal		
Xnr5bb	AB219842.1	xnr5		
Xnr5ad	AB219843.1	xnr5		
Xnr5ai	AB219845.1	xnr5		
Xnr5g	AB219846.1	xnr5		
Xnr5ac	AB219847.1	xnr5		
Xnr5af	AB219848.1	xnr5		
xnr5ah	AB219849.1	xnr5		
Xnr5aa	AB219850.1	xnr5		
xnr5ag	AB219851.1	xnr5		
Xnr5ba	AB219852.1	xnr5		
Xnr5b	AB219855.1	xnr5		
Znr2c	AF003699.1	znr2		
Znr1c	AF056327.1	NDR2		
Chicken_Nr1a	AF486810.1	nodal-related		
Lancelet_Nr	AY083838.1	nodal-related		
Xnr2a	BC169388.1	xnr2		
Xnr2b	BC169392.1	xnr2		
Xnr6a	BC169659.1	xnr6		
Xnr6b	BC169661.1	xnr6		
Xnr3b	BC169689.1	xnr3		
Xnr3c	BC169691.1	xnr3		
Xnr5bd	BC169725.1	xnr5		
Xnr5bc	BC169727.1	xnr5		
Xnr5c	BC169822.1	xnr5		
Xnr5d	BC169824.1	xnr5		
Xnr5e	BC169866.1	xnr5		

Xnr5f	BC170152.1	xnr5		
Xnr6e	BC170314.1	xnr6		
Xtnr1a	BC171037.1	xtnr1		
Onr1a	EF206724.1	onr1		
Onr2a	EF206725.1	onr2		
Xnr6d	NM_001085564.1	xnr6		
Xnr5ab	NM_001085585.1	xnr5		
Xnr3d	NM_001085790.1	xnr3		
Xnr1a	NM_001085796.1	xnr1		
Xnr2c	NM_001087967.1	xnr2		
Xnr4a	NM_001088347.1	xnr4		
Xnr5a	NM_001097061.1	xnr5		
Xtnr3b	NM_001112906.1	xtnr3		
Znr2b	NM_130966.1	NDR1	ENSDARG00000057096	znr2
Znr1a	NM_139133.1	NDR2	ENSDARG00000040299	znr1
Xtnr3e	NM_203533.1	xtnr3		
Xnr3a	U25993.1	xnr3		
Xnr1b	U29447.1	xnr1		
Xnr2d	U29448.1	xnr2		
Xnr4b	U79162.1	xnr4		
Znr1b	U87758.1	znr1		

Squirt_Na	NM_001078532.1	Nodal		
anole_N			ENSACAG00000008399	Nodal
Marmoset_N			ENSCJAG00000016288	Nodal
Cavy_N			ENSCPOG00000025772	Nodal
Sloth_N			ENSCHOG00000010347	Nodal
zebrafish_SPAW			ENSDARG00000014309	SPAW
armadillo_N			ENSDNOG00000017851	Nodal
Kangaroo Rat_N			ENSDORG00000011307	Nodal
Lesser hedgehog1_N			ENSETEG00000013276	Nodal
horse_N			ENSECAG00000017055	Nodal
Hedgehog2_N			ENSEEUG00000011834	Nodal
cat_N			ENSFCAG00000001230	Nodal
chicken_NR1	XM_424385		ENSGALG00000003209	Nr1
Gorilla(Ape)_N			ENSGGOG00000002581	Nodal
Three-spined stickleback(fish)_Na			ENSGACG00000002333	Nodal

Three-spined stickleback(fish)_Nb			ENSGACG00000008499	Nodal
Three-spined stickleback(fish)c			ENSGACG00000017712	Nodal
Elephant_N			ENSLAFG00000021867	Nodal
rhesus monkey_N	XM_001108074.1,XM_001108137.1		ENSMMUG00000023170	Nodal
kangaroo_N			ENSMEUG00000011841	Nodal
Lemur_N			ENSMICG00000015080	Nodal
turkey_N			ENSMGAG00000002207	Nodal
opossum_N			ENSMODG00000012158	Nodal
bat_N			ENSMLUG00000015297	Nodal
Pika_N			ENSOPRG00000015824	Nodal
Rabbit_N			ENSOCUG00000008685	Nodal
Galago_N			ENSOGAG00000005716	Nodal
medaka_N			ENSORLG00000006553	Nodal
medaka_Nr1			ENSORLG00000011275	Nr1

medaka_Nr2			ENSORLG00000009098	Nr2
orangutan_N			ENSPPYG00000002370	Nodal
RockHyrax_N			ENSPCAG00000015506	Nodal
FlyingFox_N			ENSPVAG00000000104	Nodal
Shrew_N			ENSSARG00000000846	Nodal
squirrel_N			ENSSTOG00000012946	Nodal
pig_Na			ENSSSCG00000010269	Nodal
pig_Nb			ENSSSCG00000010265	Nodal
Tarsier_N			ENSTSYG00000005226	Nodal
Fugu1_Na			ENSTRUG00000010779	Nodal
Fugu1_Nb			ENSTRUG00000012437	Nodal
Fugu1_Nc			ENSTRUG00000012942	Nodal
Fugu2_Na			ENSTNIG00000013237	Nodal
Fugu2_Nb			ENSTNIG00000015847	Nodal
Dolphin_N			ENSTTRG00000003182	Nodal
Alpaca_N			ENSVPAG00000002635	Nodal
frog_Ne			ENSXETG00000023748	Nodal
frog_Na			ENSXETG00000009008	Nodal

frog_Nb			ENSXETG00000016779	Nodal
frog_Nc			ENSXETG00000016778	Nodal
Xtnr3	AB093328		ENSXETG00000009009	Nr3
frog_Nd			ENSXETG00000025789	Nodal
Fugu2_Nc			ENSTNIG00000005578	Nodal
frog_Nf			ENSXETG00000017442	Nodal
AxNodal-1	GU256638	AxNodal-1		
AxNodal-2	GU256639	AxNodal-2		
Human_GDNF15	AJ001897.1	GDNF		
Human_GDNF16	AJ001898.1	GDNF		
Human_GDNF17	AJ001899.1	GDNF		
Human_GDNF18	AJ001900.1	GDNF		
horse_GDNF	XM_001497180.2	GDNF		
Human_MGDF	U11025.1	MGDF		
chimpanzee_MGDF	XM_001136518.1	MGDF		
opossum_MGDF	XM_001376801.1	MGDF		
horse_MGDF	XM_001498257.1	MGDF		
eel_ACTA	AB025356.1	activin B		
sea urchin_ACTA	EU526314.1	activin B		

sea urchin_ACTA2	NM_001128068.1	activin B		
X_ACTA2	D49543.1	activin D		
Finch_ACTA	XM_002199879.1	activin D		
X_ACTA	BC169414.1	activin D precursor		
X_ACTA3	NM_001085864.1	activin D precursor		
X_ACTB	NM_001090586.1	Activin-beta B		
X_ACTB2	S61773.1	Activin-beta B		
goldfish_ACTB	AF004669.1	Activin-beta B precursor		
carp_ACTB	DQ340764.1	Activin-beta B precursor		
rat_ACTB2	AF089825.1	Activin-beta E		
rat_ACTB	AF140032.1	Activin-beta E		
Human_ACTB	AF412024.1	Activin-beta E		
Mouse_ACTB	U96386.1	Activin-beta E		
rhesus monkey_ACTB	XM_001115949.1	Activin-beta E		
rhesus monkey_ACTB2	XM_001115958.1	Activin-beta E		
horse_ACTB	XM_001488790.1	Activin-beta E		
chimpanzee_ACTB	XM_509161.2	Activin-beta E		
cattle_ACTB	XM_595759.3	Activin-beta E		
dog_ACTB	XM_844366.1	Activin-beta E		

Junglefowl_ADMP3	AF082178.1	ADMP		
Mouse_ADMP	AF365876.1	ADMP		
Zebrafish_ADMP4	AF418564.1	ADMP		
Zebrafish_ADMP2	AF420475.1	ADMP		
Human_ADMP2	AF458592.1	ADMP		
Zebrafish_ADMP	AJ315468.1	ADMP		
Human_ADMP	AK312144.1	ADMP		
X_ADMP2	BC130130.1	ADMP		
salmon_ADMP2	BT057114.1	ADMP		
Worm_ADMP2	DQ431039.1	ADMP		
XT_ADMP	NM_001045692.1	ADMP		
sea squirt_ADMP	NM_001078517.1	ADMP		
X_ADMP	NM_001088323.1	ADMP		
X_ADMP4	NM_001097118.1	ADMP		
salmon_ADMP	NM_001146504.1	ADMP		
Worm_ADMP	NM_001164922.1	ADMP		
Zebrafish_ADMP3	NM_131876.2	ADMP		
Junglefowl_ADMP4	NM_204822.1	ADMP		
X_ADMP3	U22155.1	ADMP		
Platypus_ADMP	XM_001506733.1	ADMP		

wasp_ADMP	XM_001604676.1	ADMP		
tick_ADMP	XM_002402657.1	ADMP		
Junglefowl_ADMP	XM_422812.2	ADMP		
Junglefowl_ADMP2	XM_426514.2	ADMP		
Japanese killifish_AMH	AB166790.1	AMH (MIS, MIF)		
flounder_AMH	AB166791.1	AMH (MIS, MIF)		
Japanese killifish_AMH2	AB214971.1	AMH (MIS, MIF)		
Boar_AMH	AF006570.1	AMH (MIS, MIF)		
Alligator_AMH	AF180294.1	AMH (MIS, MIF)		
Possum_AMH	AF503621.1	AMH (MIS, MIF)		
mole_AMH	AJ550376.1	AMH (MIS, MIF)		
seabass_AMH	AM232701.1	AMH (MIS, MIF)		
seabass_AMH2	AM232703.1	AMH (MIS, MIF)		
seabass_AMH3	AM232704.1	AMH (MIS, MIF)		
turtle_AMH	AY235424.1	AMH (MIS, MIF)		
kangaroo_AMH	AY346371.1	AMH (MIS, MIF)		
Quail_AMH	AY633648.1	AMH (MIS, MIF)		

zebrafish_AMH	AY677080.2	AMH (MIS, MIF)		
zebrafish_AMH2	AY721604.1	AMH (MIS, MIF)		
salmon_AMH	AY722411.1	AMH (MIS, MIF)		
pejerrey_AMH	AY763406.2	AMH (MIS, MIF)		
zebrafish_AMH3	AY881649.1	AMH (MIS, MIF)		
Japanese killifish_AMH3	AY899282.1	AMH (MIS, MIF)		
Japanese killifish_AMH4	AY899283.1	AMH (MIS, MIF)		
duck_AMH	AY904047.1	AMH (MIS, MIF)		
quail_AMH2	AY904049.1	AMH (MIS, MIF)		
mouse_AMH	AY911505.1	AMH (MIS, MIF)		
human_AMH	BC049194.1	AMH (MIS, MIF)		
mouse_AMH2	BC150477.1	AMH (MIS, MIF)		
mouse Synthetic construct_AMH	BC167250.1	AMH (MIS, MIF)		
Tilapia_AMH	DQ257618.1	AMH (MIS, MIF)		
Tilapia_AMH2	DQ257619.1	AMH (MIS, MIF)		
Pejerrey_AMH2	DQ441594.2	AMH (MIS, MIF)		

Japanese killifish_AMH 5	DQ523689.1	AMH (MIS, MIF)		
carp_AMH	EU136185.1	AMH (MIS, MIF)		
carp_AMH2	EU136186.1	AMH (MIS, MIF)		
fox_AMH	EU371740.1	AMH (MIS, MIF)		
hamster_AMH	EU564707.1	AMH (MIS, MIF)		
boradllo_AMH	FJ587489.1	AMH (MIS, MIF)		
stickleback_AMH	FJ773241.1	AMH (MIS, MIF)		
Human_AMH 2	NM_000479.3	AMH (MIS, MIF)		
zebrafish_AMH4	NM_001007779.1	AMH (MIS, MIF)		
Japanese killifish_AMH 6	NM_001104728.1	AMH (MIS, MIF)		
salmon_AMH 2	NM_001123585.1	AMH (MIS, MIF)		
mouse_AMH 3	NM_007445.2	AMH (MIS, MIF)		
rat_AMH	NM_012902.1	AMH (MIS, MIF)		
cattle_AMH	NM_173890.1	AMH (MIS, MIF)		
chicken_AMH	NM_205030.1	AMH (MIS, MIF)		
pig_AMH	NM_214310.1	AMH (MIS, MIF)		
chicken_AMH 2	U61754.1	AMH (MIS, MIF)		

pig_AMH2	U80853.1	AMH (MIS, MIF)		
chicken_AMH3	X89248.1	AMH (MIS, MIF)		
chimpanzee_AMH	XM_001172985.1	AMH (MIS, MIF)		
opossum_AMH	XM_001372305.1	AMH (MIS, MIF)		
platypus_AMH	XM_001520602.1	AMH (MIS, MIF)		
dog_AMH	XM_542190.2	AMH (MIS, MIF)		
rhesus monkey_AMH	XR_014624.1	AMH (MIS, MIF)		
Junglefowl_BMP10A	AJ581667.1	BMP10		
human-Synthetic construct_BMP10	AY890696.1	BMP10		
Finch_BMP10	XM_002196574.1	BMP10		
Junglefowl_BMP10	XM_417667.1	BMP10		
cattle_BMP10	XM_583418.2	BMP10		
mouse_BMP15A	AF082348.1	BMP15		
seabass_BMP15	AM933668.1	BMP15		
mouse_BMP15	BC055363.1	BMP15		
human_BMP15	BC069155.1	BMP15		
zebrafish_BMP15A	BC124106.1	BMP15		
zebrafish_BMP15	BC164703.1	BMP15		

zebrafish_BMP15B	NM_001020484.1	BMP15		
mouse_BMP15B	NM_009757.4	BMP15		
rat_BMP15	NM_021670.1	BMP15		
chimpanzee_BMP15	XM_529247.2	BMP15		
tetra_BMP2	DQ915172.1	BMP2		
Japanese killifish_BMP2	DQ915174.1	BMP2		
HUMBMP2A	M22489.1	BMP2		
Japanese killifish_BMP2A	NM_001104908.1	BMP2		
zebrafish_BMP2	XM_001342061.2	BMP2		
salmon_BMP2	BT059611.1	BMP2 precursor		
human_BMP3	D49492.1	BMP3		
junglefowl_BMP3	NM_001034819.1	BMP3		
horse_BMP3	XM_001494773.2	BMP3		
Finch_BMP3	XM_002190452.1	BMP3		
junglefowl_BMP3A	DQ097308.1	BMP3 precursor		
Japanese killifish_BMP4	AF538055.1	BMP4		
zebrafish_BMP4B	BC078423.1	BMP4		
zebrafish_BMP4a	D49972.1	BMP4		

opossum_BMP4a	DQ192517.1	BMP4		
tetra_BMP4	DQ915173.1	BMP4		
duck_BMP4	EF540749.1	BMP4		
salmon_BMP4	NM_001139844.1	BMP4		
zebrafish_BMP4	U82231.1	BMP4		
junglefowl_BMP4	X75915.1	BMP4		
opossum_BMP4	XM_001362554.1	BMP4		
Finch_BMP4	XM_002200411.1	BMP4		
salmon_BMP4a	BT044754.1	BMP4 precursor		
mouse_BMP5	AK033362.1	BMP5		
mouse_BMP5C	BC100751.1	BMP5		
mouse_BMP5E	BC100752.1	BMP5		
mouse_BMP5A	BC100754.1	BMP5		
mouse_BMP5D	BC141283.1	BMP5		
mouse_BMP5B	L41145.1	BMP5		
rat_BMP5	NM_001108168.1	BMP5		
mouse_BMP5F	NM_007555.3	BMP5		
mouse_BMP6	AK041210.1	BMP6		
rat_BMP6	AY184240.1	BMP6		
mouse_BMP6C	BC138593.1	BMP6		
mouse_BMP6B	BC138595.1	BMP6		
mouse_BMP6A	NM_007556.2	BMP6		
rat_BMP6A	NM_013107.1	BMP6		

boar_BMP6	XM_001925853.1	BMP6		
boar_BMP6A	XM_001928395.1	BMP6		
cattle_BMP6	XM_869844.3	BMP6		
zebrafish_BMP7	AF201379.1	BMP7		
XT_BMP7	BC063373.1	BMP7		
zebrafish_BMP7A	NM_131321.1	BMP7		
XT_BMP7A	NM_203866.1	BMP7		
mouse_BMP8D	AK082895.1	BMP8		
mouse_BMP8	AK157978.1	BMP8		
mouse_BMP8B	BC052168.1	BMP8		
mouse_BMP8E	BC137890.1	BMP8		
zebrafish_BMP8	NM_001044971.1	BMP8		
rat_BMP8	NM_001109432.1	BMP8		
mouse_BMP8A	NM_007558.2	BMP8		
mouse_BMP8F	NM_007559.4	BMP8		
mouse_BMP8C	U39545.1	BMP8		
rat_BMP8A	XM_001054775.1	BMP8		
Nematode_DAF-7	AY672707.1	daf-7		
Worm_DAF-7	DQ058687.1	daf-7		
Nematode_DAF-7A	EF514232.1	daf-7		
Nematode_DAF-7B	NM_064864.3	daf-7		
Nematode_DAF-7C	U72883.1	daf-7		

X_DER	AF065135.1	derriere		
X_DER2	BC073508.1	derriere		
XT_DER	BC080341.1	derriere		
XT_DER2	NM_001007904.1	derriere		
X_DER3	NM_001087497.1	derriere		
Chicken_DORSALIN	L12032.1	dorsalin		
cricket_DPP	AB044710.1	DPP		
spider_DPP	AB096072.1	DPP		
fly_DPP	AB121072.1	DPP		
sludge worm_DPP	AB192888.1	DPP		
oyster_DPP	AB379969.1	DPP		
Nematode_DPP	AF004395.1	DPP		
coral_DPP	AF285166.1	DPP		
Locust_DPP	AF374725.1	DPP		
sea snail_DPP	AF499914.1	DPP		
spider_DPP2	AJ518936.1	DPP		
millipede_DPP	AJ843875.1	DPP		
clam worm_DPP	AM114782.1	DPP		
sea anemone_DPP	AY391716.1	DPP		
bug_DPP	AY899334.1	DPP		
X_DPP	BC059286.1	DPP		
sea squirt_BMPb	D85464.1	BMPb		
butterfly_DPP	EU233806.1	DPP		

silkworm_DPP	FJ572058.1	DPP		
beetle_DPP	NM_001039451.1	DPP		
silkworm_DPP2	NM_001145329.1	DPP		
fruit fly_DPP	NM_057963.4	DPP		
fruit fly_DPP2	NM_164485.1	DPP		
fruit fly_DPP3	NM_164486.1	DPP		
fruit fly_DPP4	NM_164487.1	DPP		
fruit fly_DPP5	NM_164488.1	DPP		
grasshopper_DPP	U23785.1	DPP		
bee_DPP	XM_001122815.1	DPP		
fruit fly_DPP6	XM_001355941.2	DPP		
wasp_DPP	XM_001607627.1	DPP		
mosquito_DPP	XM_001654103.1	DPP		
mosquito_DPP2	XM_001846364.1	DPP		
aphid_DPP	XM_001944112.1	DPP		
aphid_DPP2	XM_001945591.1	DPP		
aphid_DPP3	XM_001945975.1	DPP		
fruit fly_DPP7	XM_001968381.1	DPP		

fruit fly_DPP8	XM_002051935.1	DPP		
fruit fly_DPP9	XM_002077849.1	DPP		
fruit fly_DPP10	XM_002087645.1	DPP		
louse_DPP	XM_002427761.1	DPP		
nematode_DPP2	NM_072308.4	DPP/BMP like		
Zebrafish_DVR1	BC085547.1	DVR1		
Zebrafish_DVR1A	BC164172.1	DVR1		
Zebrafish_DVR1B	NM_130948.1	DVR1		
zebrafish_DVR1C	U00931.1	DVR1		
fruit fly_GBB	M84795.1	GBB		
beetle_GBB	NM_001114341.1	GBB		
fruit fly_GBB2	NM_057992.2	GBB		
aphid_GBB	XM_001947922.1	GBB		
fruit fly_GBB3	XM_002049912.1	GBB		
bee_GBB	XM_394252.1	GBB		
Mouse_GDF1A	BC079555.1	GDF1		
XT_GDF1	BC161554.1	GDF1		
Mouse_GDF1	M57639.1	GDF1		
HUMAN_GDF1A	M62302.1	GDF1		
rat_GDF1	NM_001044240.2	GDF1		

MOUSE_GDF1C	NM_001163282.1	GDF1		
HUMAN_GDF1	NM_001492.4	GDF1		
MOUSE_GDF1B	NM_008107.4	GDF1		
opossum_GDF1	XM_001370408.1	GDF1		
CATTLE_GDF1	XM_585368.3	GDF1		
Human_GDF10A	BC028237.1	GDF10		
Mouse_GDF10B	BC058358.1	GDF10		
Cattle_GDF10A	BC123524.1	GDF10		
Mouse_GDF10A	L42114.1	GDF10		
Cattle_GDF10	NM_001076167.1	GDF10		
Human_GDF10	NM_004962.2	GDF10		
Rat_GDF10	NM_024375.1	GDF10		
Mouse_GDF10C	NM_145741.2	GDF10		
Mouse_GDF10	S82648.1	GDF10		
Rhesus Monkey_GDF10	XM_001109475.1	GDF10		
Chimpanzee_GDF10	XM_001135281.1	GDF10		
Dog_GDF10	XM_848811.1	GDF10		
Zebrafish_GDF11A	AF411599.2	GDF11		
Zebrafish_GDF11	BC134028.1	GDF11		
Human_GDF11	NM_005811.3	GDF11		
Mouse_GDF11	NM_010272.1	GDF11		
Rat_GDF11	XM_001071574.1	GDF11		

Rhesus Monkey_GDF 11	XM_001096135.1	GDF11		
Boar_GDF11	XM_001927555.1	GDF11		
Rat_GDF11A	XM_343148.3	GDF11		
Chimpanzee_ GDF11	XM_509122.2	GDF11		
Human_GDF 15	AF019770.1	GDF15		
Mouse_GDF1 5	AF159571.1	GDF15		
Human_GDF 15A	AK291530.1	GDF15		
Human_GDF 15B	BC000529.2	GDF15		
Human_GDF 15C	BC008962.2	GDF15		
Mouse_GDF1 5A	BC067248.1	GDF15		
Mouse_GDF1 5B	NM_011819.2	GDF15		
Rat_GDF15	NM_019216.2	GDF15		
Rhesus Monkey_GDF 15	XM_001114375.1	GDF15		
Chimpanzee_ GDF15	XM_524157.2	GDF15		
Mouse_GDF2	AF156890.1	GDF2		
Human_GDF 2	AK314956.1	GDF2		
Human_GDF 2B	BC069643.1	GDF2		
Human_GDF 2A	BC074921.2	GDF2		
Mouse_GDF2 D	BC103625.1	GDF2		
Mouse_GDF2 C	BC103679.1	GDF2		
Mouse_GDF2 A	BC103680.1	GDF2		
Mouse_GDF2 B	BC103681.1	GDF2		

rat_GDF2	NM_001106096.1	GDF2		
Human_GDF2c	NM_016204.1	GDF2		
Mouse_GDF2E	NM_019506.4	GDF2		
junglefowl_GDF2	NM_205432.1	GDF2		
rhesus monkey_GDF2	XM_001109523.1	GDF2		
horse_GDF2	XM_001500654.1	GDF2		
Chimpanzee_GDF2	XM_507775.2	GDF2		
cattle_GDF2	XM_593677.3	GDF2		
dog_GDF2	XM_848793.1	GDF2		
Human_GDF3	BC030959.1	GDF3		
Mouse_GDF3B	BC101963.1	GDF3		
Mouse_GDF3C	BC101964.1	GDF3		
Mouse_GDF3D	BC103565.1	GDF3		
rat_GDF3A	DQ372084.1	GDF3		
Mouse_GDF3E	L06443.1	GDF3		
rat_GDF3	NM_001109671.1	GDF3		
Mouse_GDF3A	NM_008108.4	GDF3		
Human_GDF3A	NM_020634.1	GDF3		
Mouse_GDF3	S52658.1	GDF3		
cattle_GDF3	XM_001254180.1	GDF3		
Chimpanzee_GDF3	XM_508988.2	GDF3		
sea anemone_GDF5	AY391717.1	GDF5		

sea anemone_GDF5a	AY496945.1	GDF5		
Human_GDF5A	BC032495.1	GDF5		
Human_GDF5	NM_000557.2	GDF5		
horse_GDF5	NM_001082520.1	GDF5		
Mouse_GDF5	U08337.1	GDF5		
rat_GDF5	XM_001066344.1	GDF5		
rhesus monkey_GDF5	XM_001099702.1	GDF5		
rhesus monkey_GDF5A	XM_001099806.1	GDF5		
Chimpanzee_GDF5A	XM_001164592.1	GDF5		
boar_GDF5	XM_001929405.1	GDF5		
Chimpanzee_GDF5	XM_530287.2	GDF5		
dog_GDF5	XM_542974.2	GDF5		
cattle_GDF5	XM_588072.3	GDF5		
Human_GDF6	AJ537424.1	GDF6		
Mouse_GDF6A	BC141339.1	GDF6		
Mouse_GDF6	BC141340.1	GDF6		
Human_GDF6A	NM_001001557.2	GDF6		
rat_GDF6	NM_001013038.1	GDF6		
XT_GDF6	NM_001016077.2	GDF6		
X_GDF6	NM_001090364.1	GDF6		

Zebrafish_GDF6	NM_001159994.1	GDF6		
Mouse_GDF6B	NM_013526.1	GDF6		
horse_GDF6	XM_001915579.1	GDF6		
cattle_GDF6	XM_867875.3	GDF6		
Human_GDF7	AB158468.1	GDF7		
Human_GDF7A	AF522369.1	GDF7		
Mouse_GDF7	AF525752.1	GDF7		
Mouse_GDF7A	NM_013527.1	GDF7		
Human_GDF7B	NM_182828.2	GDF7		
Rat_GDF7C	XM_001063724.1	GDF7		
Rat_GDF7B	XM_001067529.1	GDF7		
Rat_GDF7	XM_001067581.1	GDF7		
Rhesus Monkey_GDF7	XM_001096970.1	GDF7		
Rat_GDF7A	XM_345646.3	GDF7		
Cattle_GDF7	XM_616701.3	GDF7		
Cattle_GDF9	AB058416.1	GDF9		
Rat_GDF9A	AF099912.1	GDF9		
Cattle_GDF9A	AF307092.2	GDF9		
boar_GDF9	AY649763.1	GDF9		
Zebrafish_GDF9	AY833104.1	GDF9		
Mouse_GDF9	BC052667.1	GDF9		
Human_GDF9B	BC096228.3	GDF9		
Human_GDF9A	BC096229.3	GDF9		
Human_GDF9	BC096230.3	GDF9		

Human_GDF9C	BC096231.1	GDF9		
Zebrafish_GDF9B	BC108013.1	GDF9		
buffalo_GDF9A	EF202171.2	GDF9		
Yak_GDF9	EU267798.1	GDF9		
Sheep_GDF9	FJ429111.1	GDF9		
buffalo_GDF9	FJ529501.1	GDF9		
Cat_GDF9	GQ294481.1	GDF9		
Mouse_GDF9A	L06444.1	GDF9		
boar_GDF9A	NM_001001909.1	GDF9		
Zebrafish_GDF9A	NM_001012383.1	GDF9		
Sheep_GDF9A	NM_001142888.1	GDF9		
Cat_GDF9A	NM_001165900.1	GDF9		
Human_GDF9D	NM_005260.3	GDF9		
Mouse_GDF9B	NM_008110.2	GDF9		
Rat_GDF9B	NM_021672.1	GDF9		
Cattle_GDF9B	NM_174681.2	GDF9		
Rat_GDF9	X81899.1	GDF9		
Chimpanzee_GDF9	XM_527008.2	GDF9		
Human_GDNF	AF053748.1	GDNF		
Junglefowl_GDNF	AF176017.1	GDNF		
Junglefowl_GDNF2	AF176018.1	GDNF		
rat_GDNF	AF205713.1	GDNF		
rat_GDNF2	AF205714.1	GDNF		

rat_GDNF3	AF205715.1	GDNF		
Zebrafish_GDNF	AF329853.1	GDNF		
rat_GDNF4	AF497634.1	GDNF		
Human_GDNF2	AY052832.1	GDNF		
rhesus monkey_GDNF	AY288835.1	GDNF		
cattle_GDNF	AY382559.1	GDNF		
carp_GDNF	AY646353.1	GDNF		
Human_GDNF3	AY893733.1	GDNF		
Human_GDNF4	BC069119.1	GDNF		
Human_GDNF5	BC069369.1	GDNF		
Mouse_GDNF	BC119031.1	GDNF		
Human_GDNF6	BC128108.1	GDNF		
Human_GDNF7	BC128109.1	GDNF		
Zebrafish_GDNF2	BC150163.1	GDNF		
X_GDNF	BC169813.1	GDNF		
Mouse_GDNF2	D49921.1	GDNF		
Human_GDNF8	DQ235474.1	GDNF		
Rat_GDNF5	EU068467.1	GDNF		
Rat_GDNF6	EU068468.1	GDNF		
Rat_GDNF7	EU068469.1	GDNF		
Rat_GDNF8	EU068470.1	GDNF		
Rat_GDNF9	EU068471.1	GDNF		
Rat_GDNF10	EU068472.1	GDNF		
X_GDNF2	EU732590.1	GDNF		
X_GDNF3	EU732591.1	GDNF		
Human_GDNF9	NM_000514.2	GDNF		

X_GDNF4	NM_001096727.1	GDNF		
Human_GDNF10	NM_001165038.1	GDNF		
Human_GDNF11	NM_001165039.1	GDNF		
Human_GDNF12	NM_001495.4	GDNF		
Mouse_GDNF3	NM_010275.2	GDNF		
Rat_GDNF11	NM_019139.1	GDNF		
Human_GDNF13	NM_199231.1	GDNF		
Human_GDNF14	NM_199234.1	GDNF		
Mouse_GDNF4	U37459.1	GDNF		
Mouse_GDNF5	U66196.1	GDNF		
Rat_GDNF12	X92495.1	GDNF		
rhesus monkey_GDNF2	XM_001094714.1	GDNF		
Junglefowl_GDNF3	XM_425018.2	GDNF		
dog_GDNF	XM_546342.2	GDNF		
cattle_GDNF2	XM_615361.4	GDNF		
rhesus monkey_IHNA2	AY574369.1	inhibin alpha		
Human-Synthetic construct_IHNA	AY889895.1	inhibin alpha		
Human_IHNA	BC006391.2	inhibin alpha		
cattle_IHNA	BC109837.1	inhibin alpha		
Human_IHNA4	BT006954.1	inhibin alpha		

buffalo_IHNA	EU884446.1	inhibin alpha		
Human_INHA	M13144.1	inhibin alpha		
Bovine_IHNA	M13273.1	inhibin alpha		
pig_IHNA	M13980.1	inhibin alpha		
Human_INHA2	M13981.1	inhibin alpha		
rhesus monkey_IHNA	NM_001032955.1	inhibin alpha		
Human_INHA3	NM_002191.2	inhibin alpha		
rat_IHNA	NM_012590.2	inhibin alpha		
cattle_IHNA2	NM_174094.3	inhibin alpha		
boar_IHNA	NM_214189.1	inhibin alpha		
Porcine_IHNA	X03265.1	inhibin alpha		
Chimpanzee_IHNA	XM_001148064.1	inhibin alpha		
mouse_IHNB3	BC026140.1	Inhibin-beta C		
rat_IHNB2	BC089799.1	Inhibin-beta C		
rat_IHNB3	NM_022614.2	Inhibin-beta C		
rhesus monkey_IHNB	XM_001115940.1	Inhibin-beta C		
mouse_IHNB4	NM_010565.3	Inhibin-beta C		
horse_IHNB	XM_001488583.1	Inhibin-beta C precursor		
cattle_IHNB	XM_609262.3	Inhibin-beta C precursor		
dog_IHNB	XM_844076.1	Inhibin-beta C precursor		
human_IHNB2	AK075285.1	Inhibin-beta E		
human_IHNB	BC005161.2	Inhibin-beta E		

mouse_IHNB	BC010404.1	Inhibin-beta E		
mouse_IHNB2	NM_008382.2	Inhibin-beta E		
human_IHNB3	NM_031479.3	Inhibin-beta E		
rat_IHNB	NM_031815.2	Inhibin-beta E		
Junglefowl_LEFTY	AB031398.1	Lefty		
flounder_LEFTY	AB232902.1	Lefty		
human_LEFTY	AF081512.1	Lefty		
zebrafish_LEFTY	AF132444.1	Lefty		
Junglefowl_LEFTY11	AF179483.1	Lefty		
X_LEFTY	AF209744.1	Lefty		
X_Lefty11	AF283563.1	Lefty		
mouse_lefty	AJ000082.1	Lefty		
human_lefty11	AK129605.1	Lefty		
human_lefty22	AK222714.1	Lefty		
human_lefty33	AK313115.1	Lefty		
sea urchin_Lefty	AY442296.1	Lefty		
human_Lefty44	BC027883.1	Lefty		
mouse_lefty11	BC050221.1	Lefty		
X_Lefty22	BC169650.1	Lefty		
mouse_Lefty22	D83921.1	Lefty		
rabbit_Lefty	EF112476.1	Lefty		
catshark_Lefty	EF174301.1	Lefty		
Japanese killifish_Lefty	EF206722.1	Lefty		
sea urchin_Lefty11	EU307282.1	Lefty		

sea squirt_LEFTY	NM_001078529.1	Lefty		
X_Lefty33	NM_001088574.1	Lefty		
rat_lefty	NM_001109080.1	Lefty		
sea urchin_Lefty22	NM_001129809.1	Lefty		
XT_Lefty	NM_001130253.1	Lefty		
rabbit_lefty11	NM_001163090.1	Lefty		
mouse_Lefty33	NM_010094.3	Lefty		
human_Lefty55	NM_020997.2	Lefty		
zebrafish_Lefty11	NM_130960.1	Lefty		
rhesus monkey	XM_001090030.1	Lefty		
rhesus monkey_LEFTY	XM_001092988.1	Lefty		
chimpanzee_Lefty	XM_001138066.1	Lefty		
chimpanzee_Lefty11	XM_001138156.1	Lefty		
cattle_Lefty	XM_001253685.1	Lefty		
horse_Lefty	XM_001915014.1	Lefty		
horse_Lefty11	XM_001915019.1	Lefty		
dog_Lefty	XM_547508.2	Lefty		
dog_Lefty11	XM_849632.1	Lefty		
fruit fly_MAV	AF252386.1	maverick		

fruit fly_MAV2	NM_001014690.1	maverick		
fruit fly_MAV3	NM_001144384.1	maverick		
fruit fly_MAV4	NM_079887.2	maverick		
bee_MAV	XM_001122118.1	maverick		
beetle_MAV	XM_001811382.1	maverick		
oyster_MGDF1	AJ130967.1	MGDF1		
oyster_MGDF2	AJ544883.1	MGDF2		
oyster_MGDF3	AJ544884.1	MGDF3		
oyster_MGDF4	AJ544885.1	MGDF4		
fruit fly_myoglianin1	AF132814.1	myoglianin		
fruit fly_myoglianin2	NM_079888.4	myoglianin		
fruit fly_myoglianin3	NM_166786.1	myoglianin		
fruit fly_myoglianin4	NM_166787.1	myoglianin		
fruit fly_myoglianin5	NM_166788.1	myoglianin		
human_MSTN	AF104922.1	Myostatin		
tilapia_MSTN	AF197193.3	Myostatin		
salmon_MSTN	AJ344158.3	Myostatin		
fugu_MSTN	AY445321.1	Myostatin		
bass_MSTN	DQ666527.3	Myostatin		
Scallop_MSTN	EU563852.2	Myostatin		
cattle_MSTN	NM_001001525.2	Myostatin		

salmon_MSTN2	NM_001123549.1	Myostatin		
salmon_MSTN3	NM_001123634.1	Myostatin		
human_MSTN2	NM_005259.2	Myostatin		
mouse_MSTN	NM_010834.2	Myostatin		
mouse_MSTN2	U84005.1	Myostatin		
fruit fly_SCREW	NM_080124.3	screw		
fruit fly_SCREW2	U17573.1	screw		
seabream_TGFB1	AF424703.1	TGF-beta 1		
carp_TGFB1	EU099588.1	TGF-beta 1		
grouper_TGFB1	GQ503351.1	TGF-beta 1		
zebrafish_TGFB1B	XM_001923618.1	TGF-beta 1		
zebrafish_TGFB1A	XM_001923622.1	TGF-beta 1		
zebrafish_TGFB1	XM_687246.2	TGF-beta 1		
nematode_TGFB2	AF104016.1	TGF-beta 2		
zebrafish_TGFB2B	AY338730.1	TGF-beta 2		
hookworm_TGFB2	AY942844.1	TGF-beta 2		
sea squirt_TGFB2	NM_001078370.1	TGF-beta 2		
zebrafish_TGFB2	NM_194385.1	TGF-beta 2		
zebrafish_TGFB2A	XM_683088.1	TGF-beta 2		

zebrafish_TGFB3C	AY338731.1	TGF-beta 3		
zebrafish_TGFB3B	AY614705.1	TGF-beta 3		
zebrafish_TGFB3A	BC081579.1	TGF-beta 3		
zebrafish_TGFB3	NM_194386.2	TGF-beta 3		
Junglefowl_TGFB3	NM_205454.1	TGF-beta 3		
platypus_TGFB3	XM_001506359.1	TGF-beta 3		
Finch_TGFB3	XM_002199958.1	TGF-beta 3		
dog_TGFB3A	XM_547918.2	TGF-beta 3		
dog_TGFB3	XM_849026.1	TGF-beta 3		
dog_TGFB3E	XM_863106.1	TGF-beta 3		
dog_TGFB3D	XM_863109.1	TGF-beta 3		
dog_TGFB3C	XM_863112.1	TGF-beta 3		
dog_TGFB3B	XM_863118.1	TGF-beta 3		
human_TGFB4C	AF081513.1	TGF-beta 4		
zebrafish_TGFB4A	AF132445.1	TGF-beta 4		
X_TGFB4C	AF283562.1	TGF-beta 4		
human_TGFB4A	AK027520.1	TGF-beta 4		
human_TGFB4	AK304549.1	TGF-beta 4		
human_TGFB4D	BC035718.1	TGF-beta 4		
mouse_TGFB4A	BC066224.1	TGF-beta 4		
X_TGFB4	BC169590.1	TGF-beta 4		
X_TGFB4A	BC169594.1	TGF-beta 4		

rat_TGFB4	NM_001007556.1	TGF-beta 4		
X_TGFB4B	NM_001085745.1	TGF-beta 4		
human_TGFB4B	NM_003240.2	TGF-beta 4		
zebrafish_TGFB4	NM_130961.1	TGF-beta 4		
mouse_TGFB4	NM_177099.3	TGF-beta 4		
cattle_TGFB4	XM_613627.3	TGF-beta 4		
X_TGFB5	BC129720.1	TGF-beta 5		
X_TGFB5B	J05180.1	TGF-beta 5		
X_TGFB5A	NM_001087861.1	TGF-beta 5		
nematode_UNC-129	AF029887.1	UNC-129		
nematode_UNC-129A	NM_069165.4	UNC-129		
X_VG1	AF041844.1	Vg1		
Chirping Frog_VG1	AF248497.1	Vg1		
Chirping Frog_VG1A	AY251032.1	Vg1		
X_VG1A	AY838794.1	Vg1		
X_VG1B	BC090232.1	Vg1		
lancelet_VG1	EU670255.1	Vg1		
X_VG1C	NM_001095591.1	Vg1		
Junglefowl_VG1	U73003.1	Vg1		

Table 5.1 Sequences Downloaded

APPENDIX 2. COMMAND LINES

Neighbor-Joining tree using PAUP

Step1:

Run the first time ML without gamma distribution to make a sample tree

```
#nexus
begin paup;
set autoclose=yes warntree=no warnreset=no;
log start file=*.GTR.paupout;
execute *.PAUP;
set criterion=distance;
dset distance=ml;
dset ?;
lset nst=6 basefreq=estimate rmatrix=estimate rates=equal
pinvar=0;
lset ?;
nj;
end;
```

Step2:

Repeat ML

```
##Repeat
likelihoods /basefreq=estimate rmatrix=estimate rates=equal
pinvar=0;
lset nst=6 basefreq=previous rmatrix=previous rates=equal
```

pinvar=0;

lset ?

nj;

##

##Repeat

likelihoods /basefreq=estimate rmatrix=estimate rates=gamma

shape=estimate ncat=16 pinvar=estimate;

lset nst=6 basefreq=previous rmatrix=previous rates=gamma

shape=previous ncat=16 pinvar=previous;

lset ?

nj;

##

Until the -ln L score remain the same

likelihoods /basefreq=estimate rmatrix=estimate rates=gamma

shape=estimate ncat=16 pinvar=estimate;

Step3:

Make the tree

#nexus

begin paup;

lset nst=6 basefreq=previous rmatrix=previous rates=gamma

shape=previous ncat=16 pinvar=previous;

lset ?

nj brlens=yes;

savetrees /fmt=phylip brlens=yes file=*.phy;

savetrees /fmt=nexus brlens=yes file=*.nex;

showdist;

savedist /format=onecolumn file=*.distances.ml.model.1col;

basefreq;

bootstrap nreps=1000 method=nj keepall=yes file=*.treefile;

end;

Neighbor-Joining tree using Phylip

Build Nj tree

Protdist→*.prodist.outfile

neighbor→*.nj.outtree, *.nj.outfile

Settings of each command:

Protdist.exe

Categories model JTT

Gamma distribution of rates among positions No

One category of substitution rate Yes

Use weights for positions No

Analyze multiple data sets No

Input sequences interleaved Yes

Terminal type IBM PC

Print out the data at start of run No

Print indications of progress of run Yes

neighbor.exe

Neighbor-joining or UPGMA

Outgroup root No, use as outgroup species 1

Lower-triangular data matrix No

Upper-triangular data matrix No

Subreplicates No

Randomize input order of species No. Use input order

Analyze multiple data sets No

Terminal type IBM PC

Print out the data at start of run No

Print indications of progress of run Yes

Print out tree Yes

Write out trees onto tree file Yes

Build Bootstraped NJ tree

seqboot→*.boot.outfile

protdist→*.boot.prodist.outfile

neighbor→*.boot.nj.outtree, *.boot.nj.outfile

consence→*.boot.nj.consense.outfile, *.boot.nj.consense.outtree

Settings of each command:

seqboot.exe

Sequence, Morph, Rest., Gene Freqs Molecular sequences

Bootstrap, Jackknife, Permute, Rewrite Bootstrap

Regular or altered sampling fraction Regular

Block size for block-bootstrapping 1<regular bootstrap>

How many replicates 1000

Read weights of characters No

Read categories of sites No

Write out data sets or just weight Data sets

Input sequences interleaved Yes

Terminal type IBM PC

Print out the data at start of run No

Print indications of progress of run Yes

Protdist.exe

Categories model JTT

Gamma distribution of rates among positions No

One category of substitution rate Yes

Use weights for positions No

Analyze multiple data sets Yes

Multiple data sets or multiple weights D (data sets)

How many data sets 1000

Input sequences interleaved Yes

Terminal type IBM PC

Print out the data at start of run No

Print indications of progress of run Yes

neighbor.exe

Neighbor-joining or UPGMA

Outgroup root No, use as outgroup species 1

Lower-triangular data matrix No

Upper-triangular data matrix No

Subreplicates No

Randomize input order of species No. Use input order

Analyze multiple data sets Yes

How many data sets 1000

Terminal type IBM PC

Print out the data at start of run No

Print indications of progress of run Yes

Print out tree Yes

Write out trees onto tree file Yes

consense.exe

Consensus type Majority rull <extended>

Outgroup root No, use as outgroup species 1

Trees to be treated as Rooted No

Terminal type IBM PC

Print out the sets of species Yes

Print indications of progress of run Yes

Print out tree Yes

Write out trees onto tree file Yes

Likelihood tree using PhyML

Nucleotide sequences:

Data type DNA

Input sequences interleaved

Analyze multiple data sets no

Run IDnone

Model of Nucleotide/Amino-acid substitution GTR

Optimise equilibrium frequencies model

Proportion of invariable sites estimated

One category of substitution rate no

Number if substitution rate categories 16

Gamma distribution parameter estimated

Middle of each rate class mean

Optimise tree topology Yes

Starting tree BioNJ

Tree topology search operations NNI

Non parametric bootstrap analysis Yes

Number of replicates 100

Approximate likelihood ratio test no

Amino Acid sequences:

Data type AA

Input sequences interleaved

Analyze multiple data sets no

Run IDnone

Model of Nucleotide/Amino-acid substitution MtREV

Amino acid frequencies model

Proportion of invariable sites estimated

One category of substitution rate no

Number if substitution rate categories 16

Gamma distribution parameter estimated

Middle' of each rate class mean

Optimise tree topology Yes

Starting tree BioNJ

Tree topology search operations NNI

Non parametric bootstrap analysis Yes

Number of replicates 100

Approximate likelihood ratio test no

Likelihood tree using MrBayes

log start filename=(filename).(Temperature).mbout

execute (datasile).paup

lset nst=6 rates=invgamma ngammacat=16

help lset

mcmcp ngen=1000000 nruns=2 nchains=4 temp=(Temperature)

help mcmcp

mcmc

##Temperature is default set to 0.2, usually it is tried around 0.2, 0.1, 0.05, 0.02 and so on. In this project, the temperature is usually about 0.02 to 0.03.

##When mcmc stopped, check the last 100,000. if the average standard deviation of split frequencies are lower than 0.01 and the chains are still swapping, it can be stopped. Type "y" to agree to stop running.

##Check the numbers shown in the table, if they are all between 0.1 and 0.7, following steps can be took place.

help sump

sump burnin=X

help sump

help sumt

sumt burnin=X

help sumt

##X is the number that need to be deleted. X=(number of tree)+1-(number want to keep). (number of tree)=(total ngen)/100. 1 is the begining tree. (number want to keep), in this project it is 1000.

##The tree is saved as *.con.tre file.

APPENDIX 3. TIPS

On a Mac computer, MacGDE can change all formats needed. On a PC, Geneious can change PAUP format into other formats while BioEdit can change other formats to Fasta or Phylip.

BioEdit Sequence Alignment Editor Version 7.0.5.3 (10/28/05), MacGDE and Geneious were used to change the file format for different programs. On a Mac computer, MacGDE can be used to change different formats needed. On a PC, Geneious can be used to change the PAUP format into other formats whereas BioEdit can be used to change other formats to Fasta or Phylip.