**STAT** genes display differential evolutionary rates that correlate with their roles in the endocrine and immune system

Marnix Gorissen¹, Erik de Vrieze¹, Gert Flik¹ and Mark O Huisin¹,²

¹Department of Organismal Animal Physiology, Faculty of Science, Radboud University Nijmegen, Heyendaalseweg 135, 6525AJ Nijmegen, The Netherlands
²The Clayton Foundation Laboratories for Peptide Biology, Peptide Biology Laboratories, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

(Correspondence should be addressed to M Gorissen; Email: m.gorissen@science.ru.nl)

---

**Abstract**

We identified orthologues of all mammalian Janus kinase (JAK) and signal transducer and activator of transcription (STAT) genes in teleostean fishes, indicating that these protein families were already largely complete before the teleost tetrapod split, 450 million years ago. In mammals, the STAT repertoire consists of seven genes (STAT1, -2, -3, -4, -5α, -5β, and -6). Our phylogenetic analyses show that STAT proteins that are recruited downstream of endocrine hormones (STAT3 and STAT5α and -5β) show a markedly higher primary sequence conservation compared with STATs that convey immune signals (STAT1-2, STAT4, and STAT6). A similar dichotomy in evolutionary conservation is observed for the JAK family of protein kinases, which activate STATs. The ligands to activate the JAK/STAT-signalling pathway include hormones and cytokines such as GH, prolactin, interleukin 6 (IL6) and IL12. In this paper, we examine the evolutionary forces that have acted on JAK/STAT signalling in the endocrine and immune systems and discuss the reasons why the JAK/STAT cascade that conveys classical immune signals has diverged much faster compared with endocrine JAK/STAT paralogues.

*Journal of Endocrinology* (2011) **209**, 175–184

---

**Introduction**

Class-I helical cytokines constitute a monophyletic group of proteins consisting of molecules that convey signals of the endocrine system (e.g. GH, prolactin, and erythropoietin (EPO)), as well as signalling molecules that coordinate host defence (e.g. interleukins; Huisin et al. 2006). All class-I helical cytokines fold in a typical four α-helix barrel structure and signal through a group of related receptors (Liongue & Ward 2007). These cytokines activate the Janus kinase/signal transducer and activator of transcription (JAK/STAT) pathway, ultimately leading to changes in gene expression. Since its discovery as a regulator of interferon (IFN) responses in the immune system (Schindler et al. 1992, Darnell et al. 1994), JAK/STAT molecules were shown to represent a common signalling pathway shared by many cytokines (Shuai & Liu 2003). The binding of a cytokine to its receptor typically leads to dimerisation of the receptor and subsequently to phosphorylation of recruited JAK molecules (Chen et al. 2004). The phosphorylated JAKs in turn phosphorylate several key tyrosines in the intracellular domain of the receptor, which then serve as a docking site for STAT proteins (Gadina et al. 2001). These STATs are phosphorylated on a single tyrosine residue, after which they form homo- or heterodimers with other phosphorylated STAT proteins. These dimers detach from the receptor and are then translocated to the nucleus, where they promote or inhibit gene expression (Darnell 1997, Levy & Darnell 2002, O’Shea et al. 2002; Fig. 1).

In mammals, the JAK family consists of four distinct genes (JAK1–3 and TYK2), whereas the STAT repertoire consists of seven distinct STAT genes, (STAT1, -2, -3, -4, -5α, -5β, and -6; Darnell 1997). Some promiscuity exists among STAT proteins regarding the ligands and cytokine receptors that can activate particular STAT members, with sometimes multiple STATs that can be activated downstream of the same cytokine receptor (e.g. leptin’s main actions are exerted through STAT3 homodimers, but STAT1/STAT3 heterodimers also serve in leptin signalling (Bendinelli et al. 2000)). Besides the well-established roles of STAT1-2, 4, and 6 in the immune system, in recent years, STAT3 and STAT5a/b have emerged as regulators of T regulatory (Treg) and T helper 17 (Th17) cell development, differentiation, and maintenance (Wei et al. 2008). Despite these contributions to vital aspects of the mammalian immune system, a very interesting demarcation exists between STAT3 and STAT5 and the other STAT family members, as most of the classical hormones within the class-I helical cytokine family, such as GH, PRL, and EPO, signal predominantly via STAT3 and STAT5, whereas the other members of the STAT family serve predominantly in the immune response (Horvath et al. 1995). A similar differentiation exists between JAK1 and JAK2, which serve both in the...
signalling of immune cytokines and endocrine cytokines, and JAK3 and Tyk2, which serve in immunology only.

All STATs share a highly conserved domain structure, including an SRC2 homology (SH2) domain – involved in the formation of STAT dimers (Shuai et al. 1994), a DNA-binding domain (Horvath et al. 1995), and a transactivation domain (TAD; Shuai et al. 1993). The latter domain shows the highest degree of variability among STATs at the primary sequence level and the gene structure (Supplementary Figure 1, see section on supplementary data given at the end of this article) and enables them to interact with different cofactors required for activation of a STAT-selective transcriptional profile (Levy & Darnell 2002).

Almost all of our knowledge on intracellular signalling of class-I helical cytokines is based on rodent and primate models. In recent years, genomes of sufficient and sufficiently diverse vertebrate species have been elucidated to initiate a comprehensive study on the phylogeny and evolution of this key family of proteins. In the present study, we compare the JAK and STAT repertoires of mammals with those of key distantly related vertebrate species, including teleostean fishes. This approach gives us unique tools to reconstruct an evolutionary history, which is surprisingly dynamic and features multiple gene duplications and subsequent deletions. Moreover, our phylogenetic analyses reveal differential evolutionary rates for the immune and endocrine members of the JAK and STAT protein family.

Materials and Methods

Identification of JAK and STAT orthologues from databases

We retrieved JAK and STAT sequences from the NCBI protein (www.ncbi.nlm.nih.gov/protein) and swissprot (www.expasy.org/sprot/) databases. To complete the JAK/STAT repertoire of key vertebrate species, we conducted an extensive BLAST (Altschul et al. 1997) search in the publicly available genome databases (www.ensembl.org; Hubbard et al. 2007). Because of the (as of yet) incomplete annotation of several genomes, it is inevitable that some BLAST searches yielded JAK/STAT orthologues that were overtly incomplete. These incomplete annotations were corrected manually, by
searching for the correct intron–exon splice sites and coding sequences in the genome. In our phylogenetic analysis, only complete coding sequences of JAK and STAT genes were used.

Reconstruction of phylogeny

Multiple sequence alignments were constructed with ClustalW (www.ebi.ac.uk/Tools/clustalw2/index.html) and uploaded into MEGA 3.0 (Kumar et al. 2004). Phylogeny was constructed on the basis of amino acid differences (P distance) using the neighbour-joining algorithm. Phylogenetic trees were constructed using both pairwise and complete deletion parameters, which rendered trees with similar topology to one another. Only the phylogenetic analyses using pairwise deletion are shown. Reliability of the trees was assessed by bootstrapping (1000 replications).

Characterisation of the nature of selective force: pN/pS ratios

We calculated the ratio between the proportion of non-synonymous (pN) and synonymous (pS) substitutions for all stats. In order to do this, the coding region of each stat parologue of zebrafish and pufferfish was aligned pairwise to its human orthologue. We corrected these nucleotide alignments manually for overt mismatches, guided by the corresponding amino acid alignments. Then, the number of (non)synonymous sites and (non)synonymous substitutions were determined using MEGA 3.0, according to the Nei & Gojobori (1986) method. To test if the level of purifying selection (i.e. pN/pS < 1) is statistically different from neutral selection (i.e. pN/pS = 1), we conducted a Z-test on the pN/pS ratios.

Results

Phylogeny of vertebrate JAK/STATs

By scrutinizing available genomes and protein databases, we identified teleostean orthologues for all mammalian JAK and STAT family members. This indicates that the contemporary JAK/STAT repertoire was already complete before the teleost–tetrapod split (~450 Mya), with the exception of the mammalian and teleostean STAT5 paralogues (as discussed later in this paper). Although STATs have been found in several invertebrates, a repertoire of STAT proteins reminiscent of the vertebrate STAT family has not been identified in any non-vertebrate. In the sea squirt (Ciona intestinalis), a key species in chordate evolution as it represents one of the closest non-vertebrate relatives to the vertebrate subphylum, only one jak and two stat genes (stat-a and stat-b) have been identified (Hino et al. 2003). This suggests that both the JAK and the STAT repertoire radiated early during vertebrate evolution, after the urochordate–vertebrate bifurcation but before the teleost–tetrapod split.

Our phylogenetic analyses show that JAK1 and JAK2 (Fig. 2), which act downstream of immune and endocrine signals, display a noticeably higher primary sequence conservation than JAK3 and Tyk2, which are restricted to signalling in immune pathways, as they cluster more compactly. Similarly, STAT3 and STAT5 (Fig. 3), which both convey endocrine signals, display noticeably higher primary sequence conservation than the STAT proteins that serve in immunity. To quantify this observation, we calculated the ratio of non-synonymous to synonymous substitutions (pN/pS ratios) of the stat repertoire of zebrafish (Danio rerio) and Japanese pufferfish (Takifugu rubripes) in comparison to the human repertoire of STAT genes (Table 1). The ratio between the pN and the pS substitutions provides us with insight into the type and strength of the selective pressure that has acted on a protein sequence in a given evolutionary time frame. In addition, pS values provide information regarding divergence time between two sequences. As synonymous substitutions generally experience no selection, orthologous or paralogous genes that have separated earlier in evolution, in general will have acquired more synonymous substitutions – and therefore carry a higher pS value – than genes that separated more recently, and therefore the pS value is an indicator of the divergence time between two sequences. By definition, pN/pS values < 1 indicate purifying selection, which is aimed at maintaining an amino acid sequence constant. The counterpart of purifying selection is positive selection, which favours amino acids changes and is characterised by a pN/pS ratio > 1. Neutral selection is assumed if neither purifying nor positive selection is demonstrated.

For all stat family members, pN/pS ratios < 1 are observed, which indicates that all stats have been subjected to some degree of purifying selection over the examined time frame. Stat3 and stat5 show markedly lower pN/pS values compared with the other stat genes, and this corroborates our earlier observation concerning the noticeably more compact clustering of these stats in phylogenetic analysis (Fig. 2) and indicates that stat3 and stat5 experienced stronger purifying selection over the course of vertebrate evolution which led to better conservation of their primary amino acid sequences in comparison to the other stat family members. In addition to the pN value for each protein, we also examined the distribution of the non-synonymous substitutions (between the STAT repertoire of zebrafish and that of human) within each member of the STAT family (Fig. 4). As already indicated by the phylogenetic analysis and pN/pS values, STAT1–2, 4, and 6 show markedly more non-synonymous substitutions. Interestingly, the domain that displays most non-synonymous substitutions is the TAD. The complete lack of studies addressing the properties of teleostean TADs precludes speculation on the consequences of the widely variable and poorly conserved TAD domains; from mammalian studies, we know that the TAD is involved in the binding of different co-factors required for STAT-induced transcription of target genes.
A model for the genesis of contemporary vertebrate STAT repertoires

Alternating views exist on early key formative events that shape contemporary vertebrate genomes. The 2R hypothesis postulates that two successive rounds of whole genome duplication occurred before the teleost–tetrapod split, accounting for the presence of many genes, or gene clusters found on four paralogous loci (Sharman & Holland 1996, Sidow 1996, Meyer & Van de Peer 2005). Others have pointed out that series of tandem duplications of large genomic segments could also account for the distribution of ancestral genes across paralogous loci (Hughes & Friedman 2003, 2004). Regardless, both hypotheses agree on the occurrence of large-scale genomic duplication events in the formative stages of the ancestral vertebrate genome. Following these large-scale rearrangement events, many of the newly acquired duplicates were lost in order to arrive at the present-day genomic distribution of many gene families, including the STATs.

In mammals, STAT genes are distributed over three independent chromosomal regions (Fig. 5), with each locus carrying two genes, with the exception of the region that contains STAT3 and both STAT5 paralogues. This suggests that the ancestral stat gene was duplicated by tandem duplication. Indeed, the stat repertoire of C. intestinalis, the sea squirt, consists of two stat genes (stat-a and stat-b) that reside on separate loci (Hino et al. 2003). One of these loci may be the representative of the ancestral stat gene that gave rise to the contemporary vertebrate stat repertoire, while the other sea squirt stat gene may have originated independently of the mechanisms that gave rise to the vertebrate stat family, or was lost in the course of vertebrate evolution. After the first tandem duplication, the ancestral locus, carrying two tandem copies of stat, was subsequently distributed over three independent loci by two large-scale genome duplication events, possibly involving the genome duplications that constitute the 2R hypothesis (Copeland et al. 1995).

The mammalian STAT5a/5b duplication is the result of a much more recent tandem duplication event that took place...
Figure 3 Neighbour-joining phylogenetic analysis of vertebrate STATs, performed under pairwise deletion and P distance conditions in MEGA3 (Kumar et al. 2004). For mammals (red), marsupials (purple), birds (yellow), amphibians (green), and teleostean fish (blue), key species are included in the phylogenetic analysis. Bootstrap values of all main branches are indicated by the size of the dots. Protein accession codes are listed in Supplementary Table 1, see section on supplementary data given at the end of this article.
in the mammalian lineage after the teleost–tetrapod split, as is discussed in the next section.

In contrast to mammalian STATs that are nicely arranged in tandem repeats, teleostean stat genes (with the exception of the stat3/stat5.1 pair) are no longer distributed in tandem pairs. Although it is clear, based on our phylogenetic analysis, that teleostean stats are orthologs of the mammalian STATs, the genes that encode them have somehow been scattered over their genome. To understand the events that underlie this distribution, we compared the synteny of the teleostean and mammalian stat genes to arrive at the following scenario for the distribution of stat genes in teleostean fish.

Early in the teleostean lineage, an additional large-scale gene duplication occurred (Wittbrodt et al. 1998, Jaillon et al. 2004), also known as the 'fish-specific genome duplication' or '3R hypothesis' and it appears that all teleostean loci that carry stat genes were duplicated in this event. In order for both duplicate copies to be maintained, each paralogue must acquire a distinct function that is subject to selection (in some cases, gene dosage may result in the maintenance of both paralogues (Kondrashov et al. 2002). However, as both paralogues will initially act fully redundantly, a failure to acquire distinct function, spatial or temporal expression patterns will usually lead to one member of each pair disappearing through genetic drift. In general, it appears that the majority of these duplicated genes in the 3R event was subsequently lost, as the total estimated gene number of teleostean species does not greatly exceed the number of genes in the human or mouse genome (Aparicio et al. 2002). This is also true for the stat gene family. After duplication, one of the duplicated genes was subsequently lost in a manner that left the contemporary teleostean stat apparently isolated at their respective loci (Fig. 5). Human STAT1 and STAT4, for example, are located on chromosome 2. In zebrafish, stat4 is located on chromosome 9, but stat1 is positioned on chromosome 22 (Fig. 4A). However, both zebrafish chromosomes carry neighboring genes that are orthologous to the neighboring genes on the human STAT1/STAT4 locus; zebrafish chromosome 9 and 22 each shares three genes in synten with the human 2q32.2/3 locus that carries human STAT1 and STAT4. Interestingly, on zebrafish chromosome 9, which carries zebrafish stat4, a remnant of another stat-like gene can be found (Stein et al. 2007). It consists only the first 16 of the 25 exons that typically encode a full STAT protein, and is present in the zebrafish EST databases (EH485578.1), indicating that this truncated stat-like gene is expressed in zebrafish and therefore does not constitute a pseudogene. This STAT-like protein does not have a TAD and may possess regulatory properties such as the mammalian truncated STAT3β, which is considered a dominant negative regulator of STAT signalling (Maritano et al. 2004). More importantly, this observation underpins our hypothesis that stat genes in the teleostean ancestor resided in tandem pairs prior to their duplication. One of the tandem copies on most loci disappeared subsequently; an hypothesis further strengthened by the fact that, in zebrafish, stat3 and stat5.1 are located in tandem, whereas stat5.2 is on a separate locus.

![Table 1](https://example.com/table1.png)

Table 1 pN/pS ratios for human versus teleostean fish STATs and representatives of the signals they convey. pN and pS are calculated with MEGA3 software (Kumar et al. 2004). For zebrafish, values for the duplicated stat5 genes are indicated with 5.1 and 5.2 in brackets.

We employed a Z-test based on the pN/pS ratios; all pN/pS ratios were proven significant with P<0.001.

Figure 4 Localisation of non-synonymous substitutions between zebrafish and human STAT genes. For each domain of the STAT proteins, the primary sequence similarity is indicated as a percentage, i.e. the lower the percentage, the higher the amount of non-synonymous substitutions.
The same phenomenon can be witnessed for the other zebrafish chromosomes containing stat genes (Fig. 5): neighbouring genes of the zebrafish stat paralogues have maintained synteny with their human orthologues. Some duplicate genes are conserved (i.e. neither of the two paralogues is discarded), and are present in synteny on both zebrafish loci. It is apparent that the teleost’s additional genome duplication resulted in the scattering of stat genes over more loci than in tetrapods (Fig. 6).

Stat5 underwent two independent tandem duplications during vertebrate evolution

Although the framework for the STAT protein family was largely complete before the teleost–tetrapod split, two additional gene duplications have occurred thereafter. In both mammals and teleostean fish, but not in birds and amphibians, duplicate stat5 genes are found. Where mammals have STAT5a and STAT5b paralogues, the teleostean duplicate stat5 genes have been named STAT5a and STAT5b (Lewis & Ward 2004). Teleostean stat5 duplicates are present in both zebrafish (D. rerio) and Japanese medaka (Oryzias latipes), but appear to be absent in pufferfishes (Tetraodon nigroviridis, Takifugu rubripes). The pN/pS ratios for STAT5a and STAT5b, compared with either STAT5a or -5b, are similar and lower than the ratios for the ‘immune’ stats (Table 1), indicating that relatively strong purifying selection has acted on both stat5 duplicates in teleosts and tetrapods alike.

The presence of STAT5.1 and STAT5.2 genes in zebrafish and medaka indicates that these two genes have arisen before the estimated divergence time of these species (~300 Mya), early in teleostean evolution. The pufferfish lineage arose ~180 Mya (Muffato & Roest-Crollius 2008) and would therefore be expected to have duplicate stat5 paralogues as well. However, the genomic landscape of the puffers is substantially different from most vertebrate genomes as it is very condensed and contains relatively little non-protein-coding DNA (Aparicio et al. 2002). In light of these profound changes in genetic makeup that were experienced in the pufferfish lineage, it is plausible that pufferfishes lost one of their stat5 paralogues after the teleostean genome duplication in the course of evolution, although we cannot conclusively rule out that we are unable to retrieve a second stat5 gene in both pufferfish species, as their respective genomes may not be entirely covered. The mammalian and teleostean stat5 duplicates do not form a uniform clade in our phylogenetic tree, as teleostean stat5 paralogues form a clade with other teleostean stat5s. This further supports our assertion that the teleostean stat5 duplication occurred independently from the mammalian duplication. The pS values calculated for both human and zebrafish stat5 paralogues provide an estimate of when the duplications of teleostean and mammalian stat5

Figure 5  Conserved synteny of human and zebrafish stat1/stat4 (A), stat3/stat5 (B), and stat2/stat6 (C) loci. The partial zebrafish stat1 orthologue is named p(stat1). In Fig. 3C, we omitted several genes that are not of interest regarding the conserved synteny of the stat loci.
paralogues may have occurred. Synonymous mutations occur and are fixed in a population at a relatively constant rate, since there is generally no selective pressure acting on these nucleotide positions. Instead, selective pressure acts on amino acid sequences, and those are not affected by synonymous substitutions. The pS value for STAT5a versus STAT5b is 0.377, whereas for the teleostean stat5.1 versus stat5.2 paralogues, the pS value is 0.766. Under the assumption of constant nucleotide substitution rate that is equal in both lineages, these numbers indicate that the teleostean paralogues arose independently from mammalian STAT5a and STAT5b and approximately twice as early in evolution. As the STAT5 duplication events in mammals and teleostean fishes occurred independently, the fact that the duplicated stat5s still exist in contemporary mammals and fish suggests that the presence of two stat5 genes presented evolutionary advantages to mammals and bony fish that has led to their maintenance in both lineages (with the aforementioned exception of the pufferfishes).

We know that the mammalian STAT5 paralogues, while highly similar in primary sequence, acquired partially independent functions: STAT5a serves in prolactin signalling, STAT5b acts downstream of GH (Schindler 2002). This is illustrated by the observations from genetic models, which revealed that Stat5a knock-out mice are deficient in prolactin signalling, while Stat5b knock-out mice display sexually dimorphic growth retardation. Although both single knock-outs are viable, mice that lack functional copies of both Stat5a and -5b die a few weeks after birth (O’Shea et al. 2002), suggesting that some redundancy still exists between these paralogues. For the teleostean stat5 paralogues, it is not known if they exert identical functions (and thus act fully redundantly) or if they have acquired different functions during the course of teleostean evolution. Nevertheless, given the fact that the primary sequences of stat5.1 and -5.2 share less identity with each other than mammalian STAT5a and STAT5b, it is tempting to speculate that the teleostean stat5 genes too have adopted at least partially unique functions.

Discussion

We have seen a dynamic evolution of the STAT transcription factors family. Just as for the class-I helical cytokines (Huising et al. 2006) and JAKs (Fig. 2), a differential primary sequence conservation for the endocrine and immune STATs is observed, with the endocrine signals being better conserved.

![Diagram of STAT evolution](https://example.com/stat_evolution_diagram.png)
than the immune signals. In mammals, it is now clear that STAT3 and STAT5a/b serve in the balance of Treg and Th17 cells (Wei et al. 2008). Our understanding of the early vertebrate immune system is not sufficient enough to proclaim that Treg and/or Th17 cells are common aspects of vertebrate immunity or constitute evolutionary recent additions to the mammalian immune systems. Regardless, the strong (endocrine driven) purifying selection that acted on STAT3 and STAT5 may mask the additional weak, immune-driven, purifying selection, resulting from additional roles of STAT3 and STAT5 in immunity.

The continuous threat of invasion by a large array of potential pathogens may have stimulated the evolutionary rate of immune signalling cascades. For example, members of the paramyxovirus family target STAT1 and STAT2 proteins in an attempt to evade the immune response. Some viruses prevent STAT tyrosine phosphorylation, and thus activation; others express STAT ubiquitin ligases, which results in the degradation of STAT proteins (Horvath 2004). STAT1 and STAT2, involved in the anti-viral response downstream of IFNs, display higher pN/pS ratios, indicating a faster rate of evolution, compared with STAT3 and STAT5a/5b. As can be seen in Fig. 2, the family of JAK proteins displays a similar dichotomy in primary sequence conservation. JAK1 and JAK2 serve in the immune system and in the endocrine system, whereas JAK3 and Tyk2 are restricted to immune system. Indeed, JAK1 and JAK2 display shorter branch lengths in phylogenetic analysis, reflecting higher primary sequence conservation.

As the challenges for the immune system are ever-changing, molecular adaptation may provide an answer to these threats. This loop of continuous adaptations of virus and host finds a remarkable homology in the ‘Red Queen hypothesis’ (Van Valen 1973). This hypothesis states that prey and predator co-evolve as one adapts to the changes of the other in a continuous loop. The endocrine system evolved under relatively constant conditions, as the communication principles in the endocrine system changed relatively little over time. On the other hand, vertebrates are under continuous threat of invasion by a large number of different and continuously evolving potential pathogens, and this may have culminated in an evolutionary arms race between the immune system and the plethora of ever-changing pathogens. Under these conditions, it may have proven advantageous for the vertebrate hosts to relax the constraints of purifying selection sufficiently to enable those STATs that concern themselves with host defence to adapt to the constantly changing playing field of pathogenic insults and thereby contribute to lasting homeostatic equilibrium and reproductive success.

**Declaration of interest**

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

**Funding**

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

**Acknowledgments**

The authors want to thank Prof. Dr Tom Gerats for constructive comments on the manuscript.

**References**


Received in final form 24 January 2011
Accepted 17 February 2011
Made available online as an Accepted Preprint 17 February 2011