*Project Reports, Comparative Genomics, Fall –04*

# Sex Determination in Primates: the SRY Locus.

## *Sex with your SOX on.*

Steven M. Thompson, stevet@bio.fsu.edu

## Abstract

Sex determination in all mammals is initiated by the *Sry* locus on the Y chromosome. *Sry* is a member of the large *Sox* HMG gene family postulated to have evolved from an ancestor of its paralogue on the Y chromosome, *Sox3*. This study investigates the relationship of *Sry* to the *Sox* genes of other animals, rooted against the HMG protein genes of a number of protists. It also looks at the evolution of *Sry* itself within the primates to ascertain whether *Sry* has differentially evolved amongst the widely varying lifestyles of the group. Two alignments are required for the analyses: just the HMG box of representative SOX proteins and the HMG outgroup, and a full length alignment of *Sry* from primates, rooted on *Sry* from the gray seal. Results corroborated but did not prove that *Sry* and *Sox3* have a most recent common ancestor. *Sox30* clustered with *Sry* in our tree as a paraphyletic group, and it had two *Danio* orthologues. Non-mammalian *Sox*-like sequences were absent only in the *Sry*, the *Sox15*, and the *Sox7*/*Sox17*/*Sox18* clades. *Sry* appeared to have an intermediate rate of evolution compared to the *Sox* sequences we compared it to. The full length *Sry* phylogeny generally agreed with conventional primate trees. Two rapid bursts of evolution are associated with the diversification of the Old World apes and monkeys from the New World monkeys. Ka/Ks ratios were all over the place, presenting little

logical order and demand further analyses. In particular the Old World monkeys had both the highest, 3.4, and lowest, 0.0 ratios.

## Introduction

The undergraduate Experimental Biology Lab course (BSC3402L) has a long history at Florida State University. It has been a required portion of the Biology Bachelor of Science degree for several years, teaching the methodology of biological research through a wet-lab and/or fieldwork based approach. Fredrik Ronquist and Steve Thompson, of the Florida State University School of Computational Sciences and Information Technology and the Biology Department, were asked to develop and teach a computationally based version of the course entitled Comparative Genomics. We wanted to teach the course using a practical, project-oriented pedagogy, and considered many model systems to use as an illustrative example. Ultimately we decided that questions related to sex and gender determination in primates by the *Sry* (Sex determining Region on the Y chromosome) locus would be an interesting candidate.

*Sry*, previously known as TDF (Testis-Determining Factor), is the single gene that initiates a long cascade toward becoming male in all mammals. The SRY protein is a transcription factor, a SOX (Sry-like bOX) HMG (High Mobility Group) box type. All SOX proteins have a single HMG box approximately 80 residues long. There are twenty two *Sox* genes in the human genome, including one pseudogene and one remote homologue, bobby sox, *Bbx* , spread over more than half the human chromosomes, all involved in the regulation of embryonic development and in the determination of the cell fate (NCBI Entrez Gene, 2004). All transcription factors interact with DNA to affect transcription (that is the generation of message RNA from the DNA template), either negatively or positively. SRY isn't a typical transcription factor, though, and all the players in the mammalian sex determination pathway are not understood. Brennan and Capel (2004) and Knower et al. (2003) review the current state of knowledge, stating that WT1+KTS, GATA4, and FOG2 are all implicated in the regulation of *Sry*, that *Sox9* is a downstream target of SRY, and that *Amh*, *Fgf9*, and *Dax1* may be subsequent downstream targets involved in the cascade toward testis development.

What is clearly known is the SRY/SOX HMG box transcription factors preferentially bind, through partial intercalation, bent DNA, especially palindromes and cruciforms, in the minor groove at the consensus sequence A/TAACAAT/A (Harley et al., 1994), though they tolerate lots of variation, particularly SRY. That binding further bends the DNA to almost $90^o$, raising speculation that they work by affecting chromatin architecture, bringing non-adjacent portions of DNA together into a transcriptional complex. An iMol (Rotkiewicz, 2003) snapshot of human SRY bound to DNA as solved through Nuclear Magnetic Resonance spectroscopy by Murphy et al. (2001) is shown in Figure 1. The model is deposited in the Protein Data Bank database (Berman et al., 2000) under accession code 1J46 (e.g. see http://molbio.info.nih.gov/cgi-bin/moldraw?1J46).

A role in pre-mRNA splicing has also been postulated (Lalli et al., 2003, and Ohe et al., 2002). They propose that (one of) SRY's regulatory action(s) is achieved through post-transcriptional repression of testis inhibiting pathways via regulation by splicing control. A possible mechanism for this would be for SRY to bind to specific pre-mRNA sites adjacent to splice sites such that splicing of key players would be inhibited.

The group research project will address two broad evolutionary questions: 1) what factors and what particular *Sox* genes were involved in the evolution and specialization of *Sry* from the rest of the *Sox* family in only the Mammalia lineage, and 2) were there differential factors in *Sry*'s evolution and hence the evolution of sexual determination in different branches of the primate tree? Many other questions relate to the above two: How did sex determination evolve, in animals, in chordates, in primates, in humans? How does SRY differ from the other HMG box, and especially SOX proteins? How and why did those genes evolve, and can they be used to determine species phylogenies? If so, do these species phylogenies agree with commonly held views, if not, why not? A recent paper by Koopman et al. (2004) addresses many of these issues. However, we feel the questions are worth reassessment in light of explosive growth in genomics data — GenBank doubles almost every year (NCBI, 2004).

Two different datasets were required to address the two main questions. A HMG box only protein dataset from a broad spectrum of life was used to ascertain how *Sry* had evolved from the rest of the *Sox* genes, and a full length *Sry* primate DNA dataset was used to investigate *Sry* evolution in primates.

Previous reports, as summarized by Graves (2002) have suggested that *Sox3* is the immediate ancestor of *Sry*. This makes sense, as *Sox3* lies on the X chromosome and the Y chromosome is thought to have evolved from the Y (see excellent reviews from Bachtrog and Charlesworth, 2001, and Jobling and Tyler-Smith, 2003), and is consistent with the fact that birds and reptiles have neither a *Sry* gene nor a Y chromosome. Marsupials have the *Sry* gene, but monotremes appear not to (Graves, 2002), which suggests, but certainly doesn't prove, that *Sry* evolved after the divergence of monotremes from the mammalian lineage. The first part of our study will investigate aspects of the assertion that *Sry* evolved from *Sox3*. The second portion of the study will examine *Sry*'s evolution in just the primate lineage. Our literature review found very little recent comprehensive coverage of this second question other than the assertion that *Sry* evolves very quickly and erratically in primates (see e.g. Wang et al., 2002).

## Data and Methods

We decided to use the human SRY protein as a starting point for this semester's Experimental Biology Comparative Genomics laboratory experience after searching traditional literature sources and Entrez at NCBI (2004) for genes involved in human sex determination. The sequence was obtained from the Swiss-Prot database (Boeckmann et al., 2003) with the ID SRY_Human. A combination of text-based, LookUp (based on SRS, Etzold and Argos, 1993), and similarity-based, BLAST (Altschul et al., 1990 and 1997) and Fast (Pearson, 1998, and Pearson and Lipman, 1988), searches, all through the Accelrys Genetics Computer Group's Wisconsin Package v.10.3 (GCG, 1982–2004) was used to assemble two different datasets to address the posed questions. Table I and II lists these datasets and identifies each sequence used in the respective analyses.

Three GCG LookUp list files were used to prepare the first dataset — all SwissProt rel.44.2 entries that are 1) eukaryotic, but are not plants, animals, or fungi, i.e. the so-called 'primitive' eukaryotes; 2) all chordates that are not mammalian; and 3) all primates. These files facilitated research by not having to sort through all SRY/SOX orthologues and paralogues for the entire Swiss-Prot database. FastA was then used to screen each LookUp list for similarity to human SRY. Furthermore, BLAST was used to screen the NRL_3D rel.28 database (Pattabiraman, et al., 1990) of protein sequences whose three-dimensional structure has been solved, and to scan the *Danio* zebrafish genome v.4 (Sanger Institute, 2004). Each similarity search produced output files with logical, easily seen breaks in their Expectation Value distributions that corresponded to other SRY orthologous sequences (when they were in the search set), the SRY paralogous SOX sequences, non-SOX HMG homologues, and sequences with little or no detectable homology to SRY. Results from all searches were loaded into SeqLab, GCG's Graphical User Interface (based on GDE, Smith et al., 1994), sequences not clearly homologous to HMG were rejected, the *Danio* genome sequences were translated, MEME profiles (Bailey and Elkan, 1994, and Bailey and Gribskov, 1998) were built from the sequences and overlaid on the dataset to facilitate manual alignment fine-tuning, and a multiple sequence alignment was prepared of the full length of the primate SRY/SOX protein sequences using GCG's PileUp (Feng and Doolittle, 1987) with a BLOSUM30 substitution matrix (Henikoff and Henikoff, 1992). Attempts were made to improve the alignment using GCG's insitu realignment option, and by hand; however, it became readily apparent that little or no homology existed outside the HMG box (see Figure 2 from GCG's PlotSimilarity program). Therefore, the alignment was truncated to include only the box region.

A Hidden Markov Model (HMM) profile (Eddy, 1996 and 1998, based on Gribskov et al., 1987 and 1989) was then built of the primate HMG SRY/SOX box, which was used to align the rest of the datasets to the existing primate alignment with HMMerAlign (Eddy, 1996 and 1998). The combined HMG box protein alignment dataset was the end result of this procedure (see Table I).

The second dataset was comprised of the full length of all primate *Sry* genes, and one of the most similar, non-primate outgroups available. This dataset began as a LookUp list of all primate sequences from both Swiss-Prot and Tremble (Boeckmann et al., 2003). FastA was then used to search and order that list for similarity to SRY_Human. All true SRY protein sequences from the FastA output were then put into SeqLab and aligned with PileUp and the BLOSUM30 matrix. Redundant sequences were eliminated. Each protein's corresponding DNA sequence (as listed in its database annotation) was next loaded into SeqLab such that the resulting DNA alignment was based on the previous protein alignment, and then the alignment was manually refined to reconcile gap placement with full codons. Meanwhile a BLAST search identified the most similar, non-primate SRY sequence from Swiss-Prot and Tremble as SRY from the gray seal, and its corresponding DNA sequence was identified from its database annotation. Finally, a DNA HMM profile was created from the full length primate alignment and HMMerAlign was used to align the gray seal *Sry* DNA sequence to the existing primate *Sry* DNA alignment. This dataset is listed in Table II. The overall similarity of this dataset was very high (see Figure 4 from GCG's PlotSimilarity program).

Both alignments were then subjected to phylogenetic analysis. An evolutionary tree was inferred from the protein HMG box only alignment using the Kimura protein distance correction model (Kimura, 1983) and neighbor-joining method (originally from Saitou and Nei, 1987) with GCG's Distance and GrowTree programs.

The full length primate *Sry* DNA alignment was first converted to a NEXUS file with the GCG PAUPSearch NoRun routine and then analyzed with PAUP*'s v.40b10 (Swofford, 2004) maximum likelihood implementation, using an HKY+I+G (Hasegawa-Kishino-Yano, plus percent invariant sites and Gamma distributed rates) model

(Hasegawa, 1985, and Swofford, 2004) with the following parameters, optimized from an initial neighbor-joining tree using ModelTest v.3.5 (Posada and Crandall, 1998):

> Base frequencies A=0.25, C= 0.30, G= 0.26, T=0.18; transition/transversion ratio=0.93; Gamma shape parameter=1.41 with four categories; and percent invariant positions=0.08.

A heuristic search with ten random additions and tree-bisection-reconnection branch swapping found one best tree with a maximum likelihood score of 3332.9.

Finally the GCG program Diverge (as modified from Li, 1993) was used to calculate all pairwise Ka/Ks ratios within the primate *Sry* DNA alignment, that is the rate of nonsynonymous substitutions per nonsynonymous site over the rate of synonymous substitutions per synonymous site.

## Results and Discussion

### *Part I: SRY among the SOX proteins.*

Our data collection rationale enabled us to assemble a very informative dataset. Since we were interested in how and where SRY came from, not in the evolution of SRY among all mammals, we were able to excluded most of the spectrum of mammals in our searches. Not only does the reduced database size increase the sensitivity and speed of similarity searches, it also facilitated our study by not confounding issues with all the mammalian Swiss-Prot proteins orthologous and paralogous to SRY. We retained primates since they were the focus of the second portion of the study, and we included all relevant NRL_3D sequences in order to infer secondary structure across our final alignment.

All literature surveyed claims that there is little or no homology outside the HMG box among the SOX proteins. In fact several papers (see e.g. Lalli, et al., 2003) assert that even SRY itself has little homology between species outside of the HMG box. We found this to definitely be the case with our combined dataset of SOX homologues from a broad spectrum of animal life, but not to be the case between *Sry* sequences of primates (see Part II). Figure 2 illustrates the running similarity of our full length SOX homologue dataset using GCG's PlotSimilarity program and a window size of ten.

The HMG box is clearly homologous between the sequences, yet, as previously reported, there is minimal similarity either upstream or downstream from it. Therefore, we trimmed the alignment down to just the HMG box portion. Interestingly, the signature motif for the HMG box, [FI]-S-[KR]-K-C-x-[EK]-R-W-K-T-M (PROSITE entry PS00353, Bairoch, 1992), does not occur anywhere within this alignment; the Pfam (A database of protein domain family alignments and HMMs, 1996-2004, The Pfam Consortium) HMM profile is found on all of the sequences. The alignment is shown in Figure 3 as represented by GCG's SeqLab at the 25% consensus level using the BLOSUM30 matrix and a threshold score of 3.

Our results, presented in an evolutionary tree (Figure 4), were consistent with previous assertions (see e.g. Graves, 2002) that a *Sox3*-like gene is *Sry*'s immediate ancestor. Though, SOX3 does cluster with SOX1, SOX2, SOX14, and SOX21 in our analysis, so it would be difficult to exclude any of this group. This entire clade is SRY's closest relative on our tree. However, only *Sox3* is found on the X chromosome, which is theorized to be the ancestor of the Y chromosome (see e.g. Bachtrog and Charlesworth, 2001), so the most parsimonious argument is for its ancestry. A surprising result of our initial analysis is that SRY has a 'sister' protein among the family — SOX30. We have not seen this reported in the literature and will need further analysis. In fact, some reviews (e.g. Nagai, 2000) assert the human SOX30 protein lies outside the entire SOX clade, other papers (Koopman, 2004) put SOX30 in the SOX clade basal to SRY near SOX15. Regardless, SOX30's tentative role in male germ cell differentiation (Osaki, 1999) is consistent with the notion that SRY and SOX30 are closely related.

Another interesting result was the placement of the several *Danio* SOX homologues all over the tree, including two sequences closely related to human SOX30. *Danio* sequences were also found in the SOX4/SOX11/SOX12 clade, and the SOX1/SOX2/SOX3/SOX14/SOX21 clade. The SOX protein family was clearly quite diversified early on in vertebrate evolution. Koopman et al. (2004) found a similar distribution of SOX proteins in the *Fugu* genome. Non-mammalian SOX like sequences were absent only in the SRY clade, and the SOX15, and SOX7/SOX17/SOX18 clades (where, surprisingly, no other primate orthologues were present in the Swiss-Prot database). Absence could just be the fact that the sequence is not in the database, rather than not in the organism though, and therefore, should not be taken as proof. The SOX8/SOX9/SOX10 clade has no other non-mammalian Swiss-Prot members besides chicken, though a quick NCBI Entrez search shows that both alligator and gecko have a SOX9 protein. Perhaps this reflects the evolution of downstream sex determination properties associated with SOX9 before the split of reptiles/birds from the lineage that led to mammals. If this is the case, the downstream sex determination properties of SOX9 predated the evolution of its potential activator, the SRY protein.

In addition to the apparent full diversification of the SOX family in vertebrates, SOX appears to be a metazoan invention, at least as far as our analysis of Swiss-Prot suggests, and is consistent with other reports. This was obvious even without the inclusion of non-chordate metazoans in our study. No eukaryote sequences that were not metazoan could be found in any of the SOX clades — those protist sequences discovered by the search strategy all clearly fell into a well-defined HMG clade along with HMG from chickens, trout, and rodents. This HMG clade is the furthest diverged from the rest of the dataset and forms a natural outgroup. Combined with the inclusion of several protist sequences, the HMG clade's ancestor is likely the ancestor to all of the SOX proteins. Nagei (2001) corroborates this conclusion; however, we do not see the greatly accelerated rate of evolution he speaks of in the SRY branch compared to the other SOX lineages. SRY diverged from the rest of the SOX proteins at an intermediate rate in our analysis, not the fastest, nor the slowest.

The extremely short sequence region available for the HMG box severely limits the resolution power of their analysis. Subsequent analyses will use more robust methods over neighbor-joining, such as protein maximum likelihood as implemented by Felsenstein (2004) in ProML. Furthermore, we will create a corresponding DNA alignment so that we can use the full suite of DNA phylogenetic inference tools available, e.g. PAUP* (Swofford, 2004) and MrBayes (Ronquist and Huelsenbeck, 2003). In any case bootstrap confidence assessment needs to be performed.

***Part II: Sry within the primates.***

The full length primate *Sry* DNA alignment was unequivocal — it had few gaps and the overall identity of the sequences was over 83% — there were no regions of homoplasy. Figures 5 and 6 attest to the high conservation of the data as seen with GCG's PlotSimilarity graph and SeqLab display. So, in spite of claims to the contrary (Lalli, et al., 2003), between species conservation is very high between different *Sry* orthologues, at least within the primates.

The maximum likelihood *Sry* phylogeny (Figure 7) generally agreed with conventional wisdom as per placement of the primate groups within the tree. Two distinct clades, each with rapid bursts of *Sry* evolution before the diversification of each distinct clade, separate the great apes and Old World monkeys from the New World monkeys. The cause of these bursts is open to speculation. The lineages within the New World monkeys segregate as expected. Of some interest is the extremely rapid evolution of *Saimiri Sry* gene in comparison to all other sequences in the study, though this may be an artifact of its truncated length in the dataset.

Within the Catarrhini, that is the apes and Old World monkeys, a few discrepancies were noted. Specifically within Hominidae (humans, gorilla, chimpanzee, bonobo, and orangutan) orangutan was excluded and grouped paraphyletic to gibbons. Gibbons also had one of the fasting evolving *Sry*'s of the entire dataset, whether this attributed to a 'long-branch-attraction' problem with orangutan remains to be seen, and orangutan is the most basal of the Hominidae in most accepted trees, so it is the most closely related of the group to the gibbon. Furthermore, the rapid diversification of *Sry* in the gibbon line may have something to do with the gibbon's extreme arboreal brachiating lifestyle, or even its monogamous lifestyle, unique among most primates, but this remains an open question. Also within the Hominidae, it is of note that chimpanzees and humans have the slowest evolving *Sry*. Why this may be is certainly an interesting question, especially given the contrasting sexual strategies of the two groups, that is, generally monogamy in humans, and polygynandry in bonobos and chimpanzees.

The Ka/Ks ratio (reviewed by Hurst, 2002) has often been used as an indicator of selection. Ratios much higher than one usually indicate positive, adaptive natural selection and ratios much smaller than one tend to indicate negative, purifying natural selection. Ratios close to one may indicate neutral selection or merely be a balance of the other two. In our analyses most pairwise ratios were about 0.5, however, a few extraordinary values appeared. A few fairly high ratios were found in the apes: human/orangutan 1.8, chimpanzee/bonobo 1.7, bonobo/gorilla 1.6, and especially bonobo/orangutan 2.6. A few of the New World monkeys were also quite high. Values between different species within the same genera of *Cebus* ranged from 2.3 to 3.2. However, some of the highest Ka/Ks ratios were within the Old World monkeys. In particular *Allenopithecus* compared to *Erythroipithecus* had a ratio 2.7, against *Papio* the value was 3.4, and compared against several of the *Cercopithecus* and *Macaca*s species the ratio values ranged up to 3.0. *Erythropithecus* also had a high ratio when calculated against *Theropithecus*, 3.2. The *Cercopithecus* genus was especially prone to high Ka/Ks ratios; several within genera between species ratios were in the 2.0 to 3.1 range. In contrast many *Macaca* comparisons generated values of zero and near zero.

The *Macaca* values are troubling though, because several other between species values within the genus were well above 1.0. Perhaps the very low values just reflect the extremely high similarity seen between some of these sequences. The interpretation of the very high Ka/Ks ratio values within many of the Old World monkeys is a mystery and demands further analysis. Whether it can be correlated with life styles or specialized niche utilization remains to be seen.

Subsequent analyses will utilize MrBayes (Ronquist and Huelsenbeck, 2003) for phylogenetic inference, and will use bootstrap confidence testing with maximum likelihood. Furthermore, the ‘counting style’ Ka/Ks ratios estimated by Diverge are quite primitive. Much more accurate estimates are available using maximum likelihood approaches as implemented in PAML (Phylogenetic Analysis by Maximum Likelihood, Yang, 1997). Likelihood ratio tests can then be used to ascertain the relevance of the results. Future work in this area may prove extremely enlightening as the evolution of *Sry* and sex determination among primates could and should vary between species in some manner that correlates with the widely varying lifestyles and sexual behaviors of the organisms.

## References Cited

Altschul, S.F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*. **215**:403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. **25**:3389–3402.

Bachtrog, D. and Charlesworth, B. (2001) Towards a complete sequence of the human Y chromosome. *Genome Biology*. **2**:1016.1–1016.5.

Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California. pp. 28–36.

Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*. **14**:48–54.

Bairoch A. (1992) PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Research*. **20**:2013–2018.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*. **28**:235–242.

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*. **31**:365–370.

Brennan, J. and Capel, B. (2004) One tissue, two fates: molecular genetic events that underlie testis versus ovary development. *Nature Reviews*. **5**:509–521.

Eddy, S.R. 1996. Hidden Markov models. *Current Opinions in Structural Biology*. **6**:361–365.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics*. **14**:755–763.

Etzold, T. and Argos, P. (1993) SRS — an indexing and retrieval tool for flat file data libraries. *Computer Applications in the Biological Sciences*. **9**:49–57.

Felsenstein, J. (1980–2004) PHYLIP (Phylogeny Inference Package), version 3.6. public domain software distributed by the author. Found on the WWW at http://evolution.genetics.washington.edu/phylip.html. Department of Genetics, University of Washington, Seattle, Washington.

Feng, D.F. and Doolittle, R. F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*. **25**:351–360.

Genetics Computer Group (GCG®). (1982–2004) *Program Manual for the Wisconsin Package®*, version 10.3. Found on the WWW at http://www.accelrys.com/products/gcg_wisconsin_package/index.html. Accelrys, a wholly owned subsidiary of Pharmacopeia Inc., San Diego, California.

Graves, J.A.M. (2002) The rise and fall of SRY. *TRENDS in Genetics*. **18**:259–264.

Gribskov M., McLachlan M., Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences, U.S.A.* **84**:4355–4358.

Gribskov, M., Luethy, R., and Eisenberg, D. (1989) Profile analysis. In: *Methods in Enzymology* 183. R.F. Doolittle, ed. Academic Press, San Diego, California. pp. 146–159.

Harley, V.R., Lovell-Badge, R. and Goodfellow, P.N. (1994). Definition of a consensus DNA binding site for SRY. *Nucleic Acids Research*. **22**:1500–1501.

Hasegawa, M., Kishino, H., and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitchondrial DNA. *Journal of Molecular Evolution*. **22**:160–174.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences, U.S.A.* **89**:10915–10919.

Hurst, L.D. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*. **18**:486–487.

Jobling, M.A., and Tyler-Smith, C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews*. **4**:598–612.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, United Kingdom.

Knower, K.C., Kelly, S., and Harley, V.R. (2003) Turning on the male—SRY, SOX9 and sex determination in mammals. *Cytogenetics and Genomic Research*. **101**:185–198.

Koopman, P., Schepers, G., Brenner, S., and Venkatesh, B. (2004) Origin and diversity of the *Sox* transcription factor gene family: genome-wide analysis in *Fugu rubripes*. Gene. **328**:177–186.

Lalli, E., Ohe, K., Latorre, E., Bianchi, M.E., and Sassone-Corsi, P. (2003) Sexy splicing: regulatory interplays governing sex determination from *Drosophila* to mammals. *Journal of Cell Science*. **116**:441–445.

Li, W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution*. **36**:96–99.

Murphy, E.C., Zhurkin, V.B., Louis, J.M., Cornilescu, G., and Clore, G.M. (2001) Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation. *Journal of Molecular Biology*. **312**:481–499.

Nagei, K. (2001) Molecular evolution of Sry and Sox gene. *Gene*. **270**:1:161-169.

National Center for Biotechnology Information (NCBI) Entrez. (2004) Found on the WWW at http://www.ncbi.nlm.nih.gov/Entrez/. National Library of Medicine, National Institutes of Health, Bethesda, Maryland.

Ohe, K., Lalli, E., and Sassone-Corsi, P. (2002) A direct role of SRY and SOX proteins in pre-mRNA splicing. *Proceedings of the National Academy of Sciences, U.S.A.* **99**:1146–1151.

Online Mendelian Inheritance in Man, OMIM™ (2000) McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University Baltimore, Maryland, U.S.A and the National Center for Biotechnology Information, National Library of Medicine Bethesda, Maryland, U.S.A. Found on the WWW at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.

Osaki, E., Nishina, Y., Inazawa, J., Copeland, N.G., Gilbert, D.J., Jenkins, N.A., Ohsugi, M., Tezuka, T., Yoshida, M., and Semba, K. (1999) Identification of a novel Sry-related gene and its germ cell-specific expression. *Nucleic Acids Research*. **27**:2503–2510.

Pattabiraman, N., Namboodiri, K., Lowrey, A., and Gaber, B.P. (1990) NRL_3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment. *Protein Sequences & Data Analysis*. **3**:387–405.

Pearson, W.B. (1998) Empirical statistical estimates for sequence similarity searches. *Journal of Molecular Biology*. **276**:71–84.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence analysis. *Proceedings of the National Academy of Sciences, U.S.A*. **85**:2444–2448.

Rotkiewicz, Piotr (2003) iMol, a free Mac OS X molecular visualization tool from PIRX. Found on the WWW at http://www.pirx.com/iMol/.

Posada, D. and Crandall, K.A. (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics*. **14**:917–818.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. **19**:1572–1574.

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

Sanger Institute (2004) "These sequence data were produced by the *Danio rerio* Sequencing Group at the Sanger Institute and can be obtained through http://www.sanger.ac.uk/Projects/D_rerio/." Hinxton, Cambridge, United Kingdom.

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 Biological Sciences*. **10**:671–675.

Swofford, D.L. (1989–2004) PAUP* (Phylogenetic Analysis Using Parsimony and other methods), version 4.0+. Florida State University, Tallahassee, Florida. http://paup.csit.fsu.edu/. Distributed through Sinaeur Associates, Inc. at http://www.sinauer.com/ Sunderland, Massachusetts.

Wang, X., Zhang, J., and Zhang, Y.P. (2002) Erratic evolution of SRY in higher primates. *Molecular Biology and Evolution*. **19**:582–584.

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*. **15**:555–556.

## General References for Further Reading

*BioInteractive: the Meaning of Sex, Genes and Gender* (2004) Howard Hughes Medical Institute, Chevy Chase, Maryland, U.S.A. Found on the WWW at http://www.hhmi.org/biointeractive/gender/index.html.

*Rediscovering Biology: Molecular to Global Perspectives* (2003) Annenberg/CPB, a unit of The Annenberg Foundation, located at the Corporation for Public Broadcasting, Washington D.C., U.S.A. Found on the WWW at http://www.learner.org/channel/courses/biology/index.html.

Wolfe, J. (2002) Sex Determination WWW Lecture. University College London, London, U.K. Found at http://www.ucl.ac.uk/~ucbhjow/b250/sex_determination.html.

*The Y Chromosome* (2003) a Nature Publishing Group Web Focus gateway site. Found on the WWW at http://www.nature.com/nature/focus/ychromosome/.

## Appendix One: Tables (Datasets in rough phylogenetic groupings)

**Table I.** Combined dataset of HMG box region from most similar primate, non-mammalian Chordata, and 'primitive' eukaryote SOX-like Swiss-Prot, all NRL_3D, and *Danio* genome protein sequences.

| Name in tree | ID | Accession | Database |
|---|---|---|---|
| hmg_tetrahymena | HMG_TETPY | P40625 | Swiss-Prot |
| hmgc_tetrahymena | HMGC_TETTH | P11873 | Swiss-Prot |
| hmg6_yeast | 1CG7A | 1CG7A | NRL_3D |
| hmgd_fruitfly | 1HMA | 1HMA | NRL_3D |
| ssrp_chicken | SSRP_CHICK | Q04678 | Swiss-Prot |
| hmg4_chicken | HMG4_CHICK | P40618 | Swiss-Prot |
| hmg1_chicken | HMG1_CHICK | P36194 | Swiss-Prot |
| hmg1_hamster | 1HSM | 1HSM | NRL_3D |
| hmg1_rat | 1HME | 1HME | NRL_3D |
| hmg2_chicken | HMG2_CHICK | P26584 | Swiss-Prot |
| hmgt_trout | HMGT_ONCMY | P07746 | Swiss-Prot |
| hmglike_trypanosome | HMGL_TRYBR | P26586 | Swiss-Prot |
| hmglike_babesia | NHP1_BABBO | P40632 | Swiss-Prot |
| sox30_human | SX30_HUMAN | O94993 | Swiss-Prot |
| 1617_3_danio | | | Sanger |
| 10216_danio | | | Sanger |
| srylikeaes4_alligator | AES4_ALLMI | P40639 | Swiss-Prot |
| 6939_1_danio | | | Sanger |
| srylikeaes1_alligator | AES1_ALLMI | P40637 | Swiss-Prot |
| srylikeaes2_alligator | AES2_ALLMI | P40638 | Swiss-Prot |
| srylikeaes6_alligator | AES6_ALLMI | P40640 | Swiss-Prot |
| sox4_human | SOX4_HUMAN | Q06945 | Swiss-Prot |
| 1534_20_danio | | | Sanger |
| 6939_danio | | | Sanger |
| sox11_human | SX11_HUMAN | P35716 | Swiss-Prot |
| sox13_xenopus | SX13_XENLA | P40650 | Swiss-Prot |
| 1004_2_danio | | | Sanger |
| 1703_34_danio | | | Sanger |
| 975_02_danio | | | Sanger |
| sox11_xenopus | SX11_XENLA | Q91731 | Swiss-Prot |
| sox11_chicken | SX11_CHICK | P48435 | Swiss-Prot |
| sox12_human | SX12_HUMAN | O15370 | Swiss-Prot |
| srylikemg43_gecko | MG43_TARMA | P40652 | Swiss-Prot |
| srylikemg42_gecko | MG42_TARMA | P40651 | Swiss-Prot |
| srylikeadw5_alligator | ADW5_ALLMI | P40636 | Swiss-Prot |
| srylikeadw4_alligator | ADW4_ALLMI | P40635 | Swiss-Prot |
| srylikeadw2_alligator | ADW2_ALLMI | P40634 | Swiss-Prot |
| srylikemg44_gecko | MG44_TARMA | P40653 | Swiss-Prot |
| srylikelg27_gecko | LG27_EUBMA | P40654 | Swiss-Prot |
| srylikelg28_gecko | LG28_EUBMA | P40655 | Swiss-Prot |
| sry_chimpanzee | SRY_PANTR | Q28798 | Swiss-Prot |
| sry_bonobo | SRY_PANPA | Q28778 | Swiss-Prot |
| sry_human | SRY_HUMAN | Q05066 | Swiss-Prot |
| sry_gorilla | SRY_GORGO | P48046 | Swiss-Prot |
| sry_orangutan | SRY_PONPY | Q28783 | Swiss-Prot |
| sry_gibbon | SRY_HYLLA | Q28447 | Swiss-Prot |

|  |  |  |  |
|---|---|---|---|
| sry_macaque | SRY_MACFA | P36391 | Swiss-Prot |
| sry_marmoset | SRY_CALJA | P51501 | Swiss-Prot |
| sox15_human | SX15_HUMAN | O60248 | Swiss-Prot |
| sox6_human | SOX6_HUMAN | P35712 | Swiss-Prot |
| sox5_human | SOX5_HUMAN | P35711 | Swiss-Prot |
| sox5_xenopus | SOX5_XENLA | P40647 | Swiss-Prot |
| sox13_human | SX13_HUMAN | Q9UN79 | Swiss-Prot |
| sox12_xenopus | SX12_XENLA | P40649 | Swiss-Prot |
| sox10_human | SX10_HUMAN | P56693 | Swiss-Prot |
| sox10_chicken | SX10_CHICK | Q9W757 | Swiss-Prot |
| sox9_orangutan | SOX9_PONPY | P61754 | Swiss-Prot |
| sox9_chimpanzee | SOX9_PANTR | Q9BG89 | Swiss-Prot |
| sox9_human | SOX9_HUMAN | P48436 | Swiss-Prot |
| sox9_macmaque | SOX9_MACMU | P61753 | Swiss-Prot |
| sox9_chicken | SOX9_CHICK | P48434 | Swiss-Prot |
| sox8_human | SOX8_HUMAN | P57073 | Swiss-Prot |
| sox8_chicken | SOX8_CHICK | P57074 | Swiss-Prot |
| sox17_human | SX17_HUMAN | Q9H6I2 | Swiss-Prot |
| sox18_human | SX18_HUMAN | P35713 | Swiss-Prot |
| sox7_human | SOX7_HUMAN | Q9BT81 | Swiss-Prot |
| srylikech2_chicken | CH02_CHICK | P40666 | Swiss-Prot |
| srylikeama2_alligator | AMA2_ALLMI | P40642 | Swiss-Prot |
| srylikech1_chicken | CH01_CHICK | P40665 | Swiss-Prot |
| srylikech32_chicken | CH32_CHICK | P40671 | Swiss-Prot |
| hmgb_tetrahymena | HMGB_TETTH | P40626 | Swiss-Prot |
| srylikeama3_alligator | AMA3_ALLMI | P40643 | Swiss-Prot |
| srylikech7_chicken | CH07_CHICK | P40669 | Swiss-Prot |
| srylikech31_chicken | CH31_CHICK | P40670 | Swiss-Prot |
| srylikech4_chicken | CH04_CHICK | P40668 | Swiss-Prot |
| srylikeama1_alligator | AMA1_ALLMI | P40641 | Swiss-Prot |
| sox14_human | SX14_HUMAN | O95416 | Swiss-Prot |
| sox14_macaque | SX14_MACFA | P61259 | Swiss-Prot |
| sox14_chicken | SX14_CHICK | Q9W7R6 | Swiss-Prot |
| 2132_6_danio |  |  | Sanger |
| sox21_human | SX21_HUMAN | Q9Y651 | Swiss-Prot |
| sox21_chicken | SX21_CHICK | Q9W7R5 | Swiss-Prot |
| 1617_2_danio |  |  | Sanger |
| sox3_human | SOX3_HUMAN | P41225 | Swiss-Prot |
| sox3_chicken | SOX3_CHICK | P48433 | Swiss-Prot |
| sox3_xenopus | SOX3_XENLA | P55863 | Swiss-Prot |
| sox1_newt | SOX1_PLEWA | P37839 | Swiss-Prot |
| 1251_08_danio |  |  | Sanger |
| 473_4_danio |  |  | Sanger |
| sox19_zebrafish | SX19_BRARE | P47792 | Swiss-Prot |
| 602_danio |  |  | Sanger |
| 373_08_danio |  |  | Sanger |
| sox2_chicken | SOX2_CHICK | P48430 | Swiss-Prot |
| sox2_xenopus | SOX2_XENLA | O42569 | Swiss-Prot |
| 603_danio |  |  | Sanger |
| sox2_human | SOX2_HUMAN | P48431 | Swiss-Prot |
| sox1_human | SOX1_HUMAN | O00570 | Swiss-Prot |
| 801_danio |  |  | Sanger |
| 85_02_danio |  |  | Sanger |

**Table II.** Nonredundant primate *Sry* DNA sequence dataset rooted with gray seal *Sry*.

| Name in tree | Locus ID | Accession | Common name |
|---|---|---|---|
| halichoerus_grypus | AY424660 | AY424660 | gray seal |
| homo_sapiens | HSSRY | X53772 | human |
| pan_troglodytes | PTDNASRY | X86380 | chimpanzee |
| pan_paniscus | PPDNASRY | X86381 | bonobo |
| gorilla_gor | GGDNASRY | X86382 | gorilla |
| pongo_pygmaeus | PPSRYDNA | X86383 | orangutan |
| hylobates_lar | HSDNASRY | X86384 | common gibbon |
| saimiri_sciur | AF151695 | AF151695 | squirrel monkey |
| cebus_apella_ap | AF338387 | AF338387 | |
| cebus_apella_xan | AF338390 | AF338390 | |
| cebus_capucinus | AF338388 | AF338388 | white-faced sapajou |
| cebus_albifrons | AF338385 | AF338385 | white-fronted capuchin |
| cebus_olivaceus | AF338389 | AF338389 | weeper capuchin |
| aotus_azarai | AF338375 | AF338375 | Azara's night monkey |
| aotus_lemurinus | AF338374 | AF338374 | lemurine night monkey |
| leontopithecus_ros | AF338373 | AF338373 | golden lion tamarin |
| saguinus_midas | AF338391 | AF338391 | |
| callimico_goeldii | AF338383 | AF338383 | Goeldi's marmoset |
| callithrix_aurita | AF338392 | AF338392 | white-eared marmoset |
| callithrix_pygmaea | AF338382 | AF338382 | pygmy marmoset |
| callithrix_jacchus | AF338379 | AF338379 | white-tufted-ear marmoset |
| presbytis_mela | AF284282 | AF284282 | mitred leaf monkey |
| pygathrix_bieti | AF454966 | AF454966 | black snub-nosed monkey |
| pygathrix_nemaeus | AF454972 | AF454972 | Douc langur |
| trachypithecus_cris | AF284283 | AF284283 | silvered leaf monkey |
| trachypithecus_pha | AF454971 | AF454971 | Phayre's leaf monkey |
| trachypithecus_fran | AF454965 | AF454965 | Francois's leaf monkey |
| erythrocebus_patas | AY048073 | AY048073 | red guenon |
| allenopithecus_nig | AF284331 | AF284331 | Allen's swamp monkey |
| miopithecus_talapoin | AY048076 | AY048076 | talapoin |
| cercopithecus_mit | AY450885 | AY450885 | blue monkey |
| cercopithecus_mon | AF284332 | AF284332 | Mona monkey |
| cercopithecus_aet | AY450882 | AY450882 | African green monkey |
| cercopithecus_dia | AY450890 | AY450890 | Diana monkey |
| cercopithecus_wol | AY450889 | AY450889 | Wolf's monkey |
| cercopithecus_sol | AY450888 | AY450888 | |
| cercopithecus_nic | AY450887 | AY450887 | white-nosed guenon |
| cercopithecus_neg | AY450886 | AY450886 | De Brazza's monkey |
| cercopithecus_ham | AY450884 | AY450884 | owl-faced monkey |
| cercopithecus_cep | AY450883 | AY450883 | moustached monkey |
| theropithecus_gel | AF284329 | AF284329 | gelada baboon |
| papio_hamadryas | AF284328 | AF284328 | hamadryas baboon |
| macaca_nemestrina | AY224239 | AY224239 | pig-tailed macaque |
| macaca_hecki | AF284307 | AF284307 | |
| macaca_maura | AF284308 | AF284308 | moor macaque |
| macaca_tonkeana | AF284286 | AF284286 | Tonkean macaque |
| macaca_ochreata | AF284320 | AF284320 | booted macaque |
| macaca_nigra | AF284318 | AF284318 | Celebes crested macaque |
| macaca_silenus | AF284288 | AF284288 | liontail macaque |
| macaca_cyclopis | AF425289 | AF425289 | Taiwan macaque |
| macaca_sinica | AF284284 | AF284284 | toque macaque |
| macaca_sylvanus | AF425296 | AF425296 | Barbary ape |
| macaca_assam | AY224238 | AY224238 | Assam macaque |
| macaca_thibetana | AY224240 | AY224240 | Pere David's macaque |
| mandrillus_leuco | AF454968 | AF454968 | drill |
| mandrillus_sphinx | AF284330 | AF284330 | mandrill |

## Appendix Two: Figures.

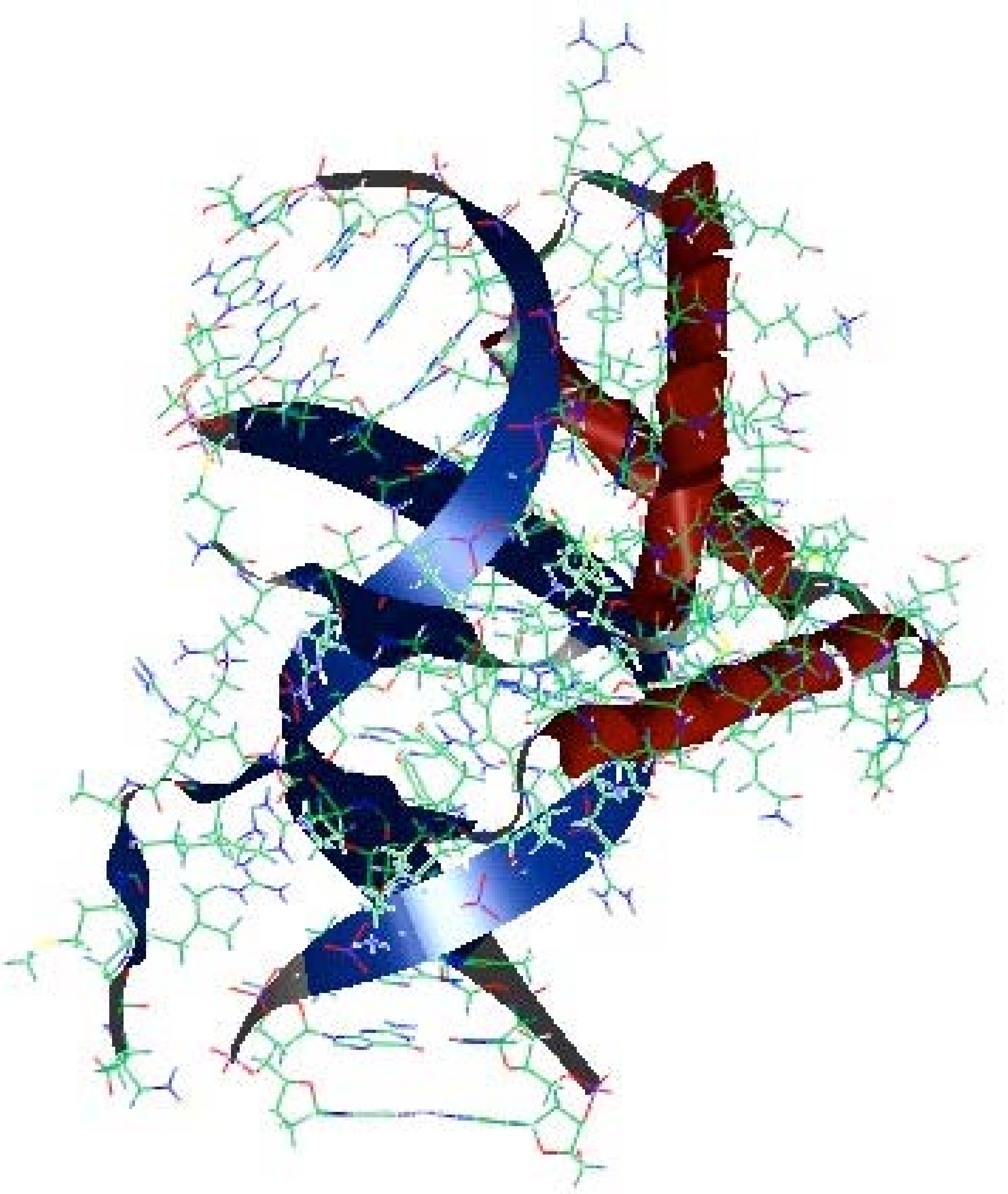
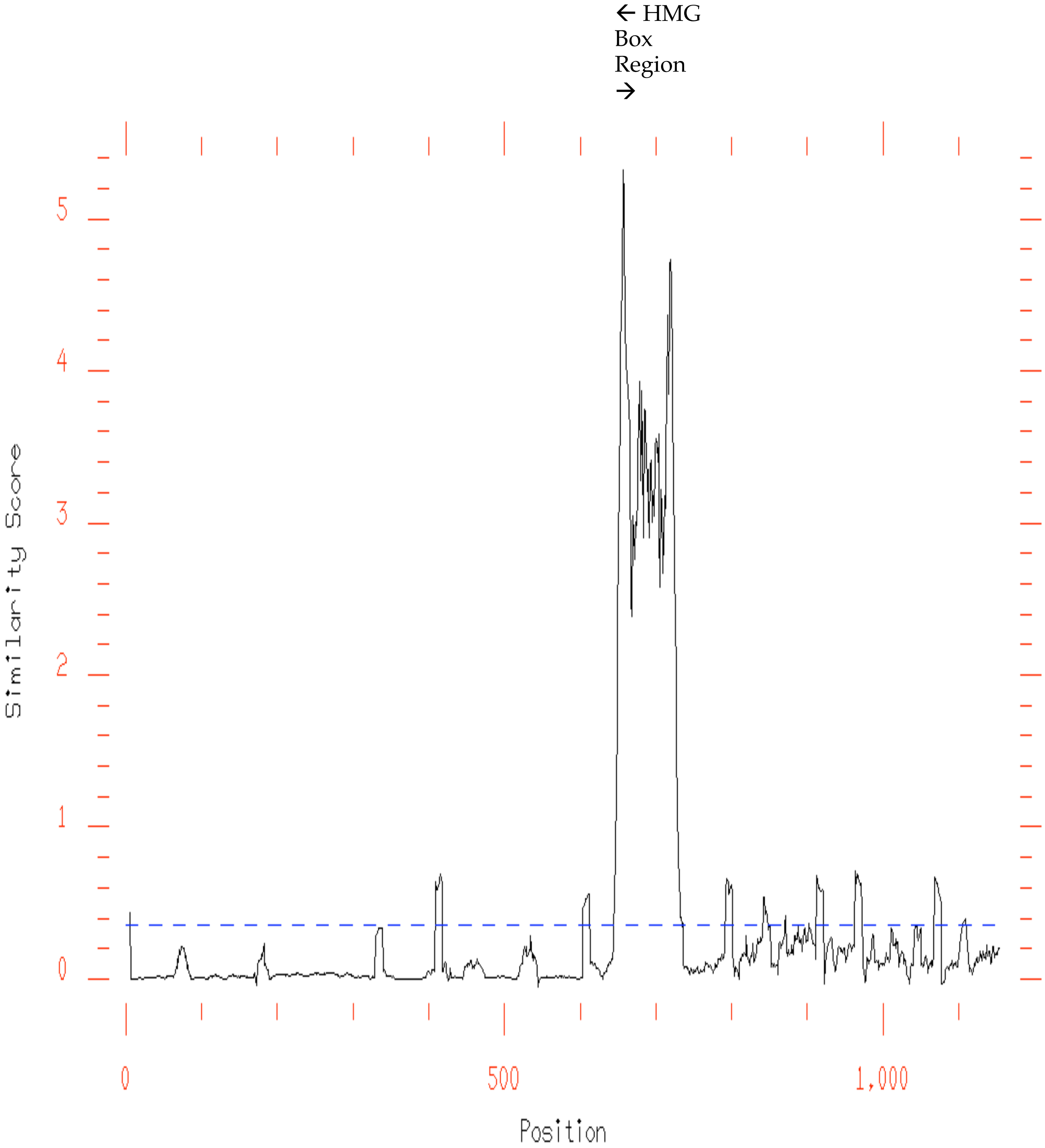


**Figure 1.** Human wild-type SRY complexed with DNA, 1J46 in PDB, as visualized with iMol set to display a “Flat Ribbon” “Backbone Style” and “Wire” “Atom Style.”



← HMG Box Region →

**Figure 2.** Overall similarity of the original full length combined dataset was extremely low, with a mean of about 0.2 on the BLOSUM30 scale, and a pronounced peak, with a BOSUM30 score up to 5.0, at the single HMG box, as visualized with GCG's PlotSimilarity and the default window size of ten residues.

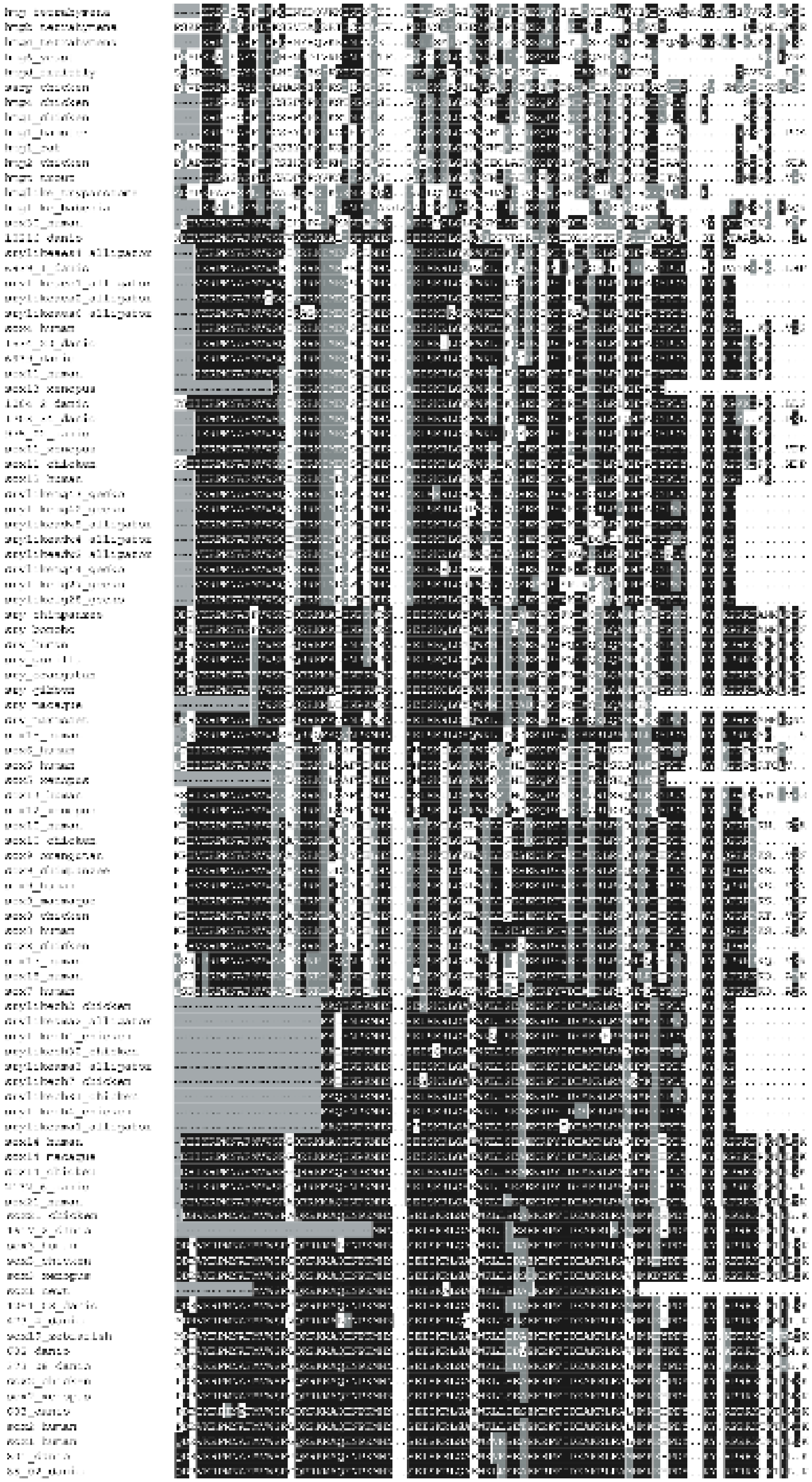


**Figure 3.** 25% consensus GCG SeqLab alignment of combined HMG box dataset.

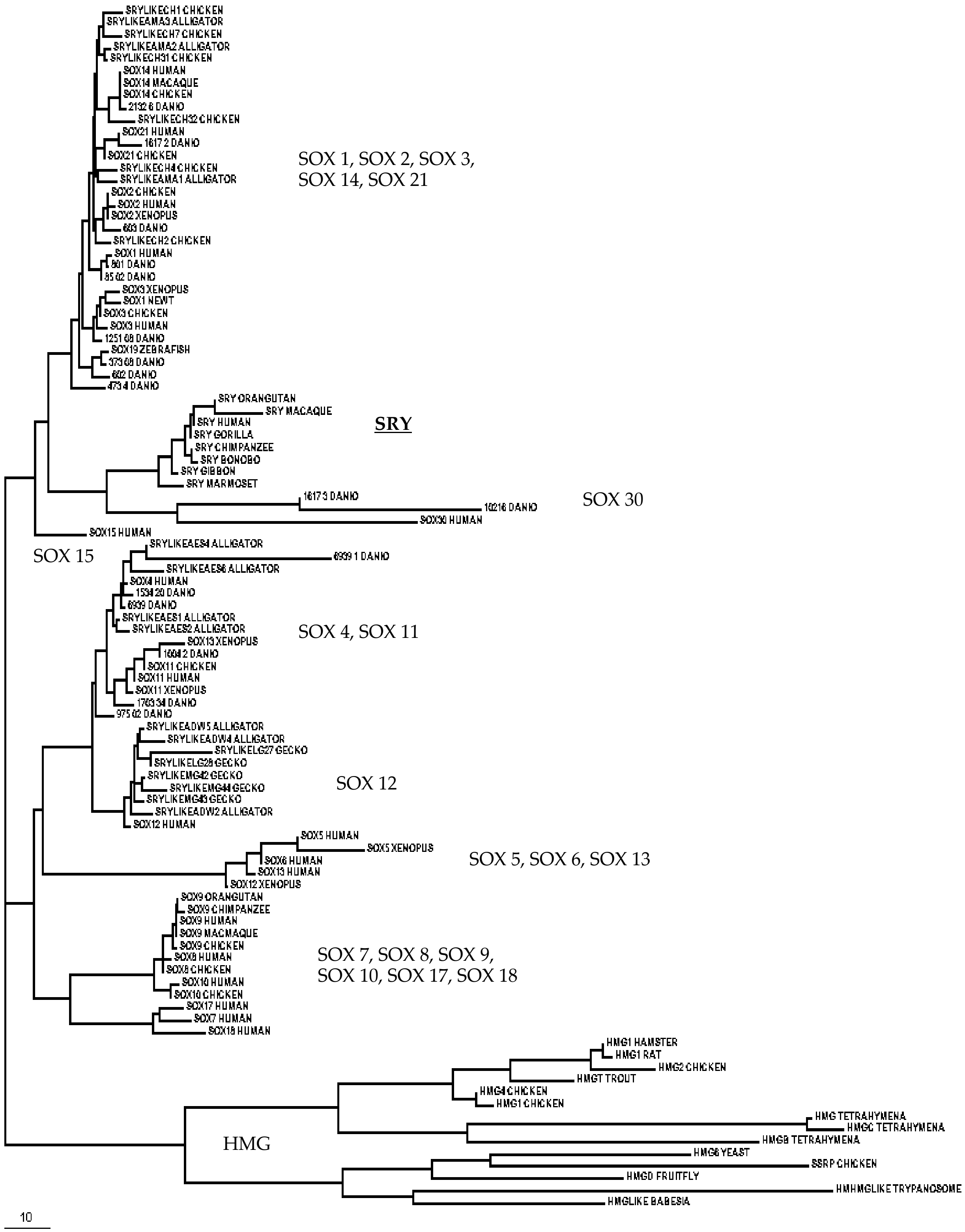


**Figure 4.** Neighbor-Joining tree from combined protein dataset of HMG box regions

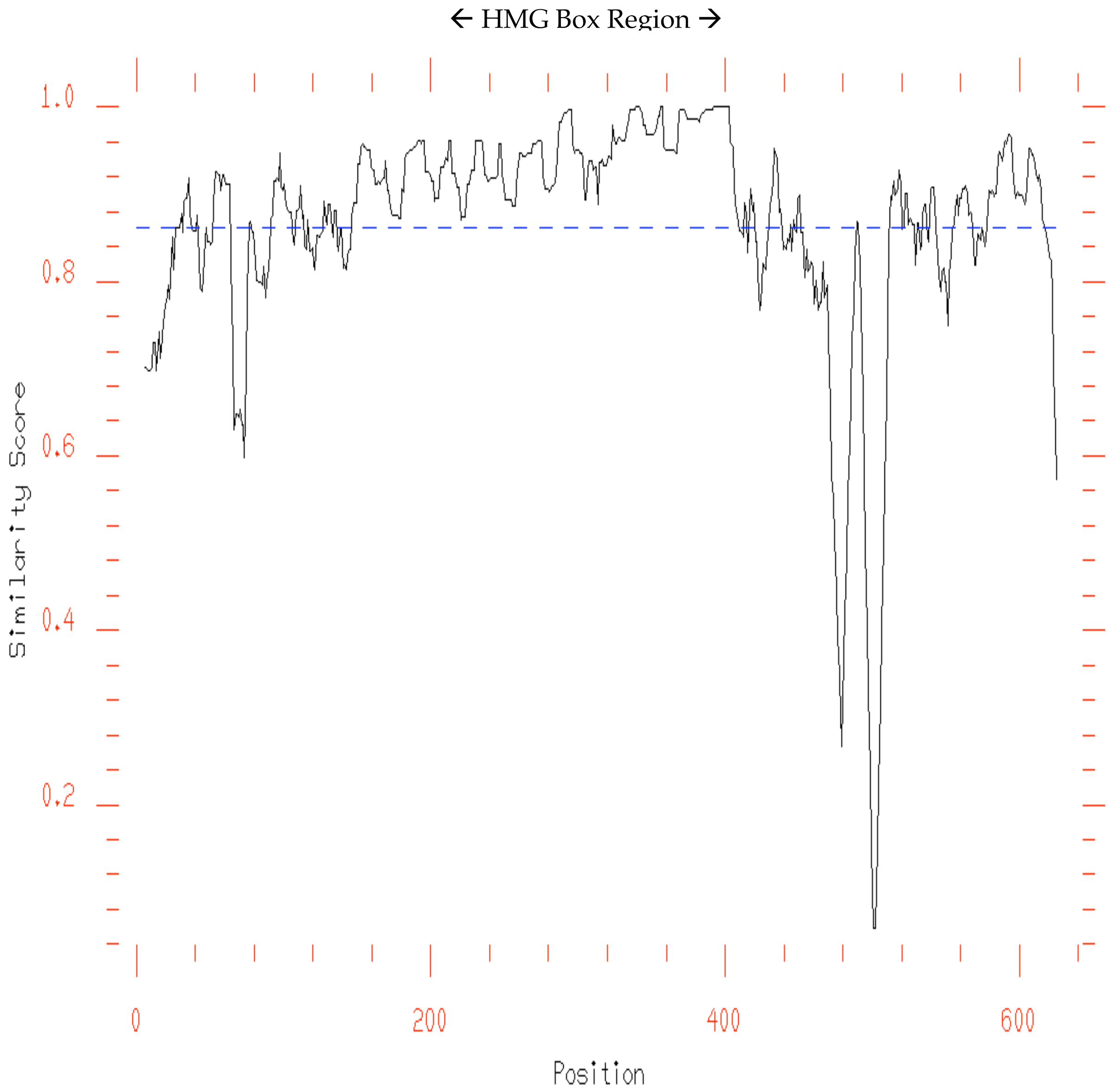


**Figure 5.** GCG plotsimilarity graph of primate (+outgroup) SRY dataset. Note extremely high conservation along the full length of the alignment, with a mean similarity of over 83% and several regions of identity.

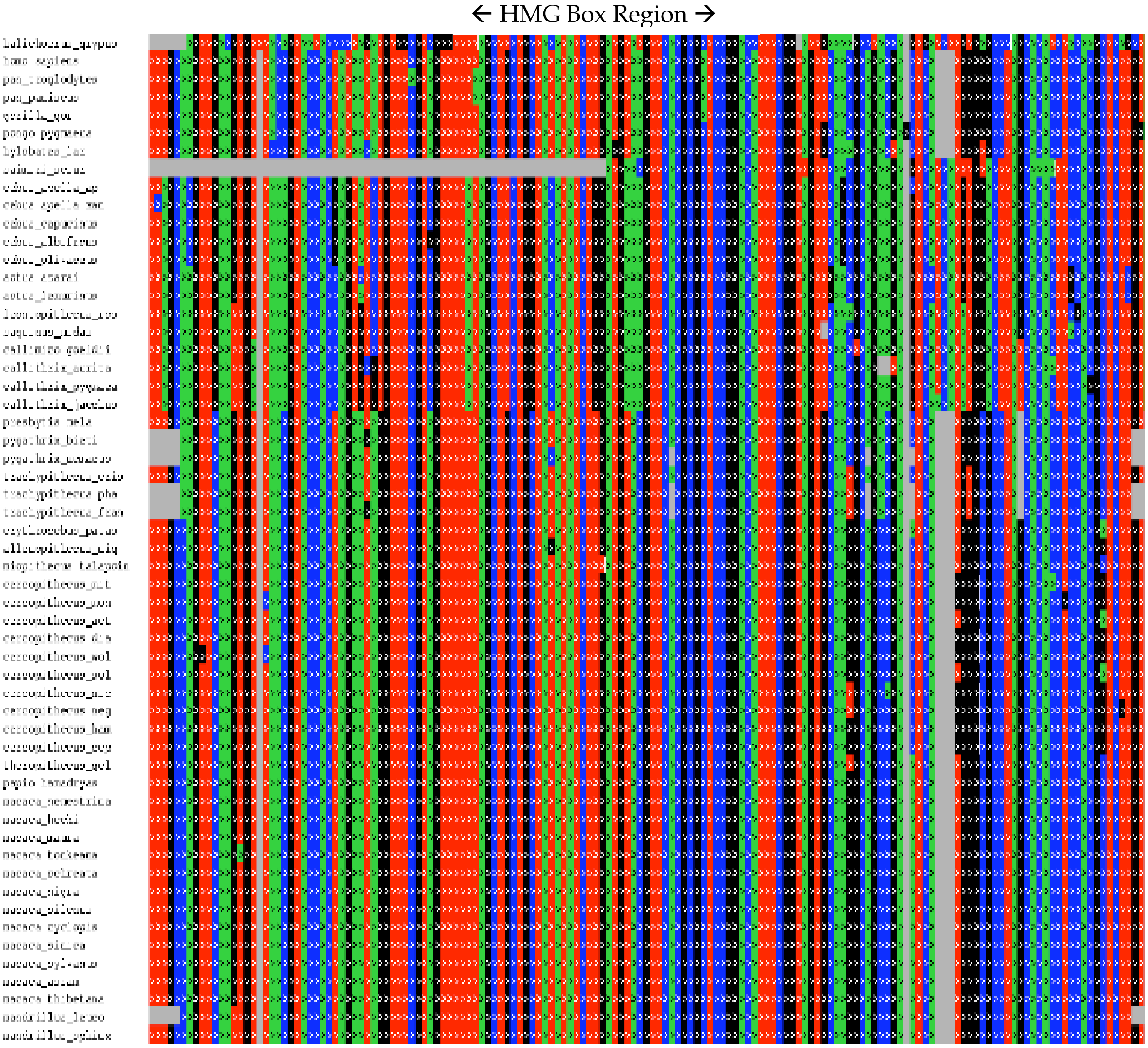


**Figure 6.** Full length primate (+outgroup) SRY alignment as portrayed by GCG's SeqLab at a four to one 'zoom' ratio. A's are red, C's are blue, G's are black, and T's are green in this representation (in the online PDF file). The HMG box region is indicated. Even without being able to read individual bases the high conservation and minimum gap placement of the alignment is obvious.

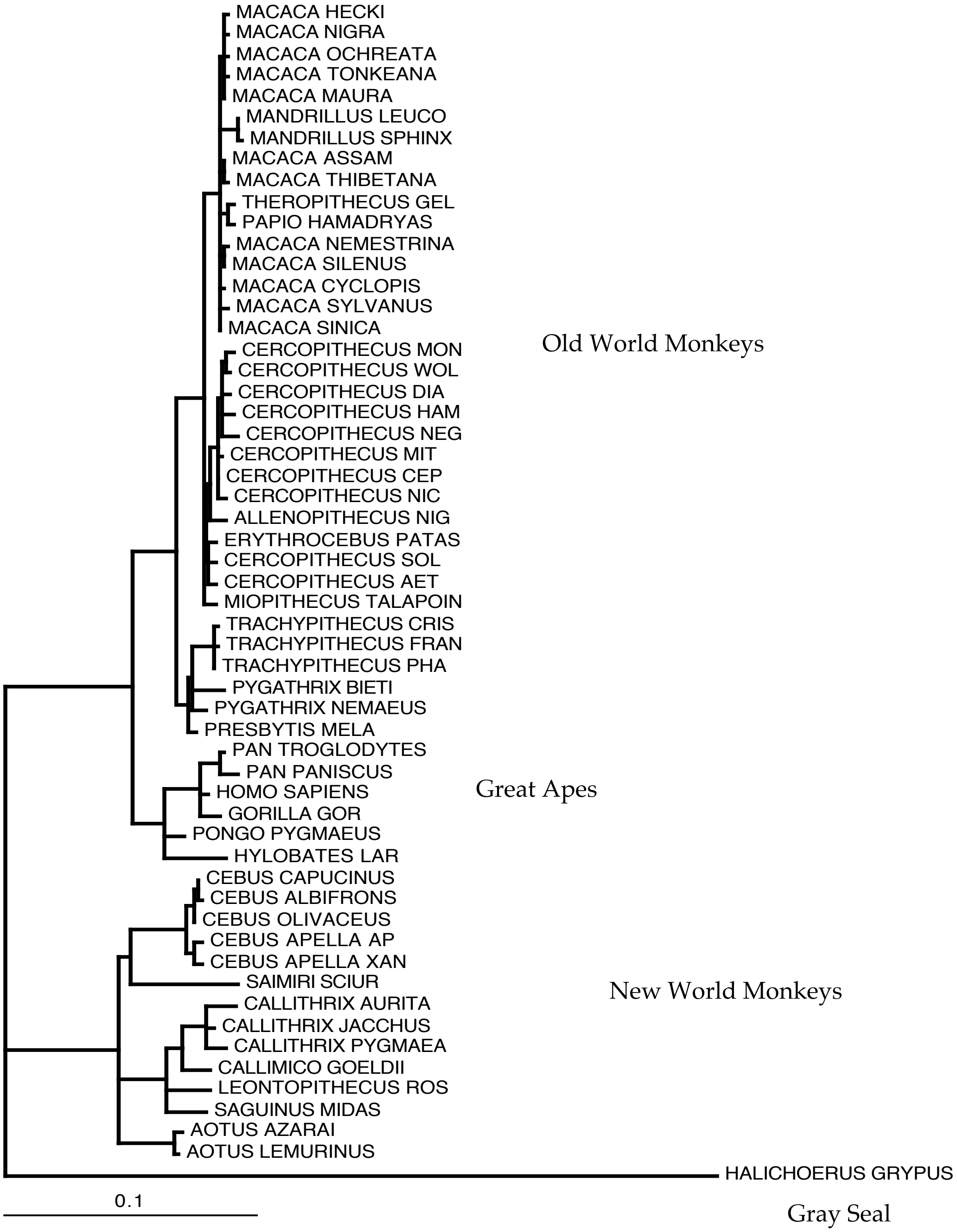


**Figure 7.** Maximum likelihood tree of complete primate SRY gene sequences, rooted with the most similar non-primate SRY sequence available (from Gray Seal). Tree inference from PAUP*, HKY+I+G model with ten random additions and TBR branch swapping.