Accelerated Evolution and Loss of a Domain of the Sperm–Egg–Binding Protein SED1 in Ancestral Primates

Ondrej Podlaha, David M. Webb, and Jianzhi Zhang

Department of Ecology and Evolutionary Biology, University of Michigan

Proteins involved in sperm–egg binding have been shown to evolve rapidly in several groups of invertebrates and vertebrates. Mammalian SED1 (secreted protein containing N-terminal Notch-like type II epidermal growth factor (EGF) repeats and C-terminal discoidin/F5/8 C domains) is a recently identified sperm surface protein that binds the egg zona pellucida and facilitates sperm–egg adhesion. SED1-null male mice are subfertile. Here we examine the SED1 gene from 11 mammalian species and provide evidence that it underwent accelerated evolution in ancestral primates, most likely driven by positive selection. Specifically, the intensity of the positive selection across various protein domains of SED1 was heterogeneous. Although one of the 2 Notch-like EGF domains, which mediate protein–protein binding, was lost in primate SED1, the second EGF domain evolved under strong positive selection favoring polar to nonpolar amino acid replacements. By contrast, the 2 discoidin/F5/8 type C domains, which are involved in protein–cell membrane binding, do not show definite signs of positive selection. The structural modification and occurrence of directional selection in ancestral primates but not any other lineage suggest that the function of SED1 may have changed during primate evolution. These results reveal a different evolutionary pattern of SED1 from that of many other sperm–egg–binding proteins, which often show diversifying selection occurring in multiple lineages.

Genes involved in sperm–egg binding often show a rapid pace of evolution driven by positive diversifying selection (reviewed in Swanson and Vacquier 2002). SED1 (*s*ecreted protein containing N-terminal Notch-like type II *E*GF repeats and C-terminal *d*iscoidin/F5/8 C domains) is a recently identified sperm protein from the mouse *Mus musculus* (Ensslin and Shur 2003). It is expressed in spermatogenic cells and secreted by the initial segment of the caput epididymis, resulting in its localization on the sperm plasma membrane overlying the acrosome. SED1 binds to the zona pellucida, specifically the glycoproteins ZP2 and ZP3, of unfertilized oocytes but not to the zona of fertilized eggs. The fertility of SED1-null male mice is about one-third that of wild-type mice, and SED1-null sperm cannot bind to the egg coat in vitro (Ensslin and Shur 2003). Here we report the finding of a structural change and positive directional selection of SED1 in an ancestral primate lineage and suggest that the SED1 function may have changed in primates.

Mouse SED1 is a short splice form of the well-known milk fat globule-EGF factor 8 (MFGE8) gene. SED1 contains 2 Notch-like EGF domains responsible for protein–protein interaction and 2 discoidin/F5/8 type C domains involved in protein–cell membrane interaction (fig. 1). To examine the evolutionary pattern of SED1, we downloaded 8 mammalian SED1 sequences from GenBank, including those from the human, mouse, rat, pig, cow, horse, dog, and opossum. Sequence alignment revealed conserved domain organization of SED1 across all examined taxa with the exception of the human, which lacks the N-terminal EGF domain (fig. 1). One structural–functional model of SED1 is that it forms a homodimer through the interaction of EGF domains (Ensslin and Shur 2003), and it is believed that at least 2 EGF domains are required for successful protein–protein binding (Lawrence et al. 2000; Balzar et al. 2001). If correct, loss of one of these domains would reduce or prohibit the dimerization of SED1, suggesting a potential functional change of SED1 in primates. To narrow down the time when the EGF-like domain was lost, we sequenced 5 additional species that represent major lineages of higher primates (chimpanzee, gorilla, orangutan, rhesus monkey, and spider monkey). The gene tree of the SED1 sequences from 6 primates and 5 representatives of nonprimate mammals showed a topology consistent with the well-established species tree (Murphy et al. 2004; Goodman et al. 2005), indicating that the SED1 gene sequences under investigation are orthologous. We did not use the dog and opossum sequences because they were from draft genome sequences that may contain errors. The SED1 of all 6 primates have the same domain structure (fig. 1), indicating that the first EGF domain was lost after the separation of primates from rodents but before the divergence of platyrrhines (New World monkeys) and catarrhines (Old World monkeys, apes, and humans). For convenience, we will refer to this period of time as an ancestral lineage of primates.

An indicator of a protein functional shift is the occurrence of positive directional selection, which can be tested by comparing the rates of synonymous and nonsynonymous nucleotide substitutions for the tree branch in which the functional shift is suspected (Nei and Kumar 2000). We used 3 different methods to conduct such a test for the ancestral primate branch in which the N-terminal EGF domain of SED1 was lost. First, we estimated branch lengths in the SED1 gene tree, in terms of the number of synonymous substitutions per synonymous site (b_S) and the number of nonsynonymous substitutions per nonsynonymous site (b_N), using the least-squares method (Zhang et al. 1998) (fig. 2). The ancestral primate branch (bolded in fig. 2) exhibits a distinct pattern of $b_N > b_S$. Using a 2-tail z-test, we found b_N (0.183 ± 0.020) to be significantly greater than b_S (0.059 ± 0.020) ($P < 10^{-4}$). This large-sample test (Zhang et al. 1997) is appropriate here because the inferred numbers of synonymous and nonsynonymous substitutions are both greater than 10 for the concerned branch (see below). Second, we inferred the ancestral gene sequences at all interior nodes of the SED1 tree by a Bayesian method (Yang et al. 1995). There were $n = 120.5$ nonsynonymous

Key words: SED1, MFGE8, positive selection, primates, sperm–egg binding, fertilization.

E-mail: jianzhi@umich.edu.

Mol. Biol. Evol. 23(10):1828–1831. 2006
doi:10.1093/molbev/msl066
Advance Access publication July 24, 2006

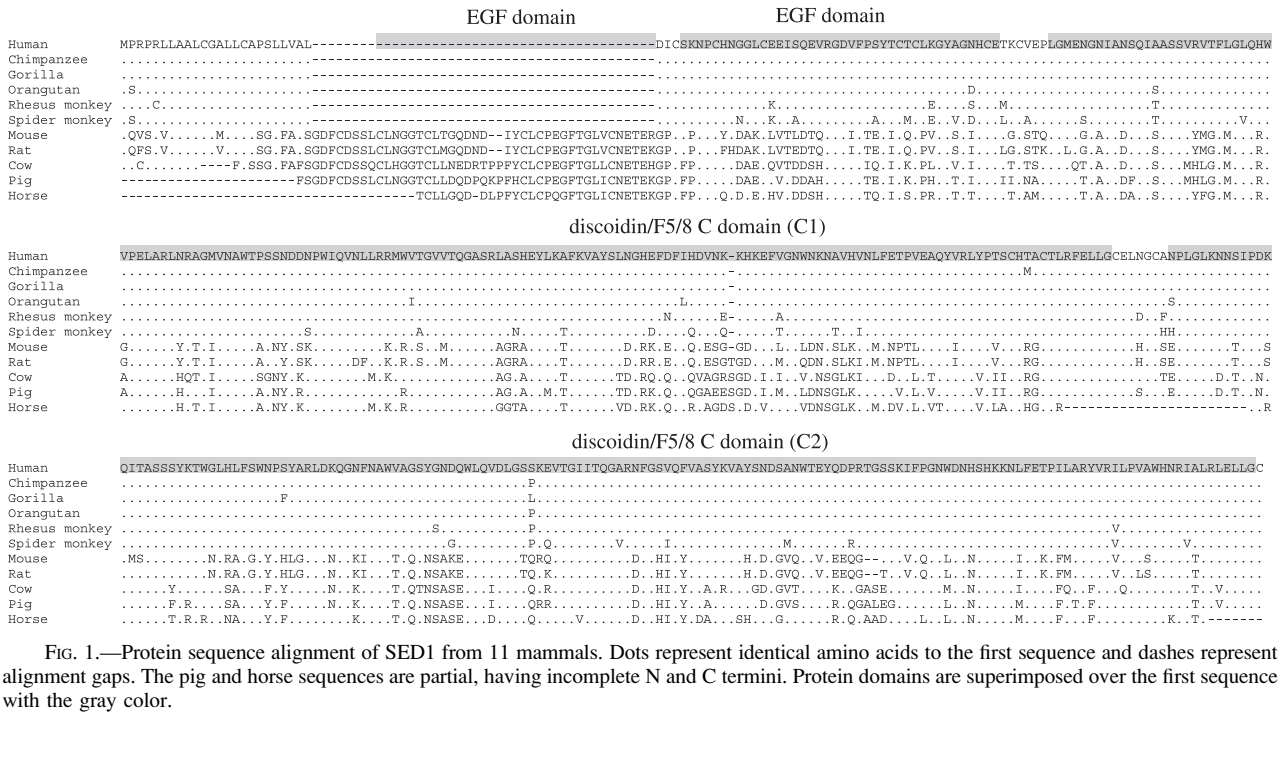

FIG. 1.—Protein sequence alignment of SED1 from 11 mammals. Dots represent identical amino acids to the first sequence and dashes represent alignment gaps. The pig and horse sequences are partial, having incomplete N and C termini. Protein domains are superimposed over the first sequence with the gray color.

and $s = 34.5$ synonymous differences between the 2 nodes that are at the ends of the ancestral primate branch. The potential numbers of nonsynonymous and synonymous sites in SED1 are $N = 688.1$ and $S = 304.9$, respectively. Thus, the n/s ratio is 3.49, significantly greater than its neutral expectation ($N/S = 2.26$) ($P < 0.01$, Fisher's exact test). Finally, we used a likelihood method known as the branch-site test of positive selection or test 2 (Zhang et al. 2005), which compares the likelihood of a null model that does not invoke positive selection with that of an alternative model that invokes positive selection in a predetermined tree branch (i.e., the bolded branch in fig. 2). For SED1, the null neutral model is rejected in favor of the alternative model ($\chi^2 = 7.3$, df = 1, $P < 0.007$), with $p_2 = 20\%$ of sites being estimated to be under positive selection in the ancestral primate branch (nonsynonymous/synonymous rate ratio $\omega_2 = 5.42$ for these sites). Thus, the distance-, parsimony-, and likelihood-based methods all show a significantly higher substitution rate at nonsynonymous sites than synonymous sites in the ancestral primate branch of the SED1 tree, strongly suggesting the operation of positive selection.

To examine which domain of the ancestral primate SED1 was under positive selection, we examined each domain separately by the branch-site likelihood method. The null neutral hypothesis is strongly rejected for the retained EGF domain ($\chi^2 = 16.8$, df = 1, $P < 0.001$) but is only marginally rejected or not rejected for the 2 discoidin/F5/8 type C domains (abbreviated C1 and C2; C1, $\chi^2 = 4.24$, df = 1, $P = 0.039$; C2, $\chi^2 = 3.56$, df = 1, $P = 0.059$). The ω_2 values for the EGF, C1, and C2 domains are estimated to be 999 (i.e., exceeding the largest ω_2 examined by the program) ($p_2 = 0.38$), 5.35 ($p_2 = 0.15$), and 3.4 ($p_2 = 0.03$), respectively. Interestingly, the identified positively selected sites (with posterior probabilities >0.95) in the C domains are enriched in spike regions, which are involved in membrane binding (Shur et al. 2004). When substitution rates are calculated from ancestral sequences, only the EGF domain had significantly higher n/s over N/S ($P = 0.0023$, Fisher's exact test). We also investigated whether any particular type of nonsynonymous substitutions were preferentially fixed in the ancestral primate branch by comparing the rates of conservative and radical nonsynonymous substitutions (Zhang 2000). When amino acid polarity is considered, the number of radical nonsynonymous substitutions per radical nonsynonymous site (d_R) is significantly greater than the number of conservative nonsynonymous substitutions per conservative nonsynonymous site (d_C) in the EGF domain ($d_R/d_C = 4.5$, $P < 0.001$, Fisher's exact test) but not in the other domains. More specifically, 8 of the 10 polarity-altering amino acid replacements in EGF were from polar to nonpolar residues, suggesting that nonpolar amino acids may have been selectively favored in the ancestral primate branch.

Our evolutionary analyses of the mammalian SED1 sequences provide strong evidence that SED1 was subject to positive selection in ancestral primates. The selection appears to target the EGF domain and favors changes of amino acid polarity. A difficulty in deciphering the selective agent on SED1 is that it is a short splicing variant of the *MFGE8* gene. Ensslin and Shur (2003) named the short variant SED1 to distinguish it from the previously known long splicing variant (MFGE8) that is found in the milk. MFGE8 participates in many different physiological functions including vascular endothelial growth factor–dependent neovascularization (Silvestre et al. 2005), inhibition of blood-coagulating enzymes (Shi and Gilbert 2003), inhibition of rotaviral infections (Kvistgaard et al. 2004), and somatic cell-to-cell interaction (Ishii et al. 2005). In mice, SED1 has a broader expression than MFGE8. SED1 expression was detected through reverse transcriptase–polymerase

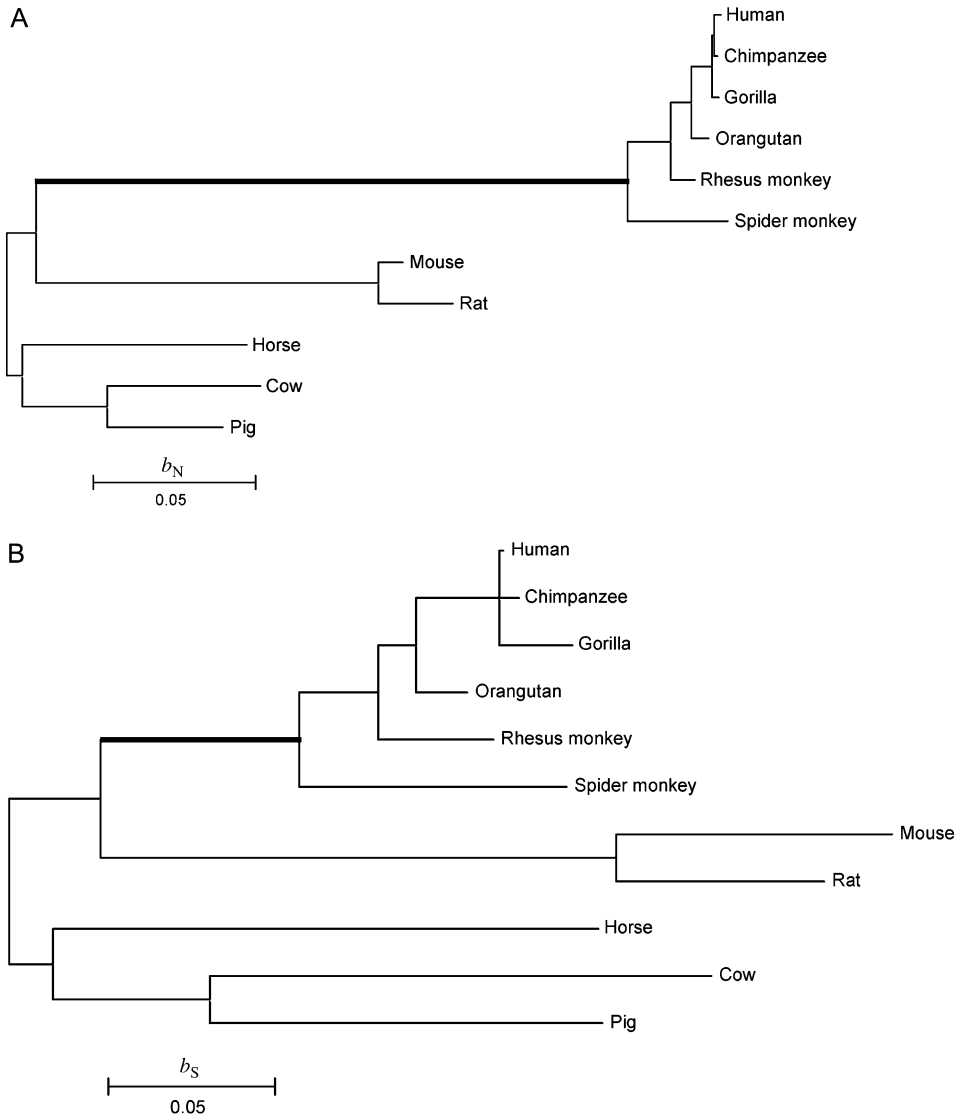

FIG. 2.—The SED1 gene tree, with branch lengths measured by (*A*) the number of nonsynonymous substitutions per nonsynonymous site and (*B*) the number of synonymous substitutions per synonymous site. The modified Nei–Gojobori method was used to estimate the synonymous and nonsynonymous distances between pairs of extant species, and the least-squares method was used to estimate the branch lengths.

chain reaction in liver, intestine, kidney, skin, stomach, heart, testis, brain, spleen, mammary glands, and lung tissues, whereas MFGE8 was detected primarily in the mammary glands and at much lower levels in skin, stomach, testis, spleen, and lung (Watanabe et al. 2005). Primate SED1, however, is less well studied. Information of expressed sequence tags show that human SED1/MFGE8 is expressed in many tissues, including the testis (http://www.ncbi.nlm.nih.gov/UniGene/ESTProfileViewer.cgi?uglist=Hs.3745).

Many sperm–egg–binding proteins show rapid evolution driven by diversifying selection occurring in multiple evolutionary lineages (reviewed in Swanson and Vacquier 2002). Such diversifying selection may alter the binding efficiency or specificity but is unlikely to change the basic function of the protein. In the case of mammalian SED1, however, positive selection was identified in a single lineage (ancestral primates), with additional characteristics of directional selection and functional shifts such as the loss of an EGF domain and increase of nonpolar residues in the other EGF domain. Although the exact selective agent on primate SED1 is unknown, in part due to the complication of multifunctionality and alternative splicing, the discrepancy in evolutionary pattern between SED1 and other sperm–egg–binding proteins is intriguing. A functional change of a sperm–egg–binding protein during primate evolution, if proven, would have significant evolutionary, physiological, and medical implications. Our results call for a thorough functional assay of primate SED1.

Methods

The SED1 gene was amplified via polymerase chain reaction from the genomic DNAs of the chimpanzee *Pan troglodytes*, gorilla *Gorilla gorilla*, orangutan *Pongo*

pygmaeus, rhesus monkey *Macaca mulatta*, and spider monkey *Ateles geoffroyi*, purified, and sequenced by automatic DNA sequencing. Attempts to amplify prosimian SED1 genes were unsuccessful. The SED1 sequences from the human *Homo sapiens*, cow *Bos taurus*, pig *Sus scrofa*, horse *Equus caballus*, mouse *M. musculus*, and rat *Rattus norvegicus* were obtained from GenBank. The DNA sequences were aligned following a protein alignment by ClustalX (Thompson et al. 1997). Several methods were used to compare rates of synonymous and nonsynonymous substitutions in SED1. First, we used the modified Nei–Gojobori method (Zhang et al. 1998) to estimate pairwise synonymous and nonsynonymous distances between extant sequences and the least-squares method to estimate synonymous and nonsynonymous branch lengths of a given tree (Zhang et al. 1998). Second, we used a Bayesian method (Yang et al. 1995) to infer ancestral SED1 gene sequences at all interior nodes of the SED1 gene tree and compare the numbers of synonymous and nonsynonymous substitutions for individual tree branches using Fisher's exact test (Zhang et al. 1997). Finally, we used an improved branch-site likelihood method (Zhang et al. 2005) implemented in PAML (Yang 1997) to test positive selection. Rates of conservative and radical nonsynonymous substitutions were estimated using the method of Zhang (2000). MEGA3.1 (Kumar et al. 2004) was used for phylogenetic analysis. The domain structure of SED1 was also examined for the preliminary sequences identified from the dog and opossum draft genome sequences.

Acknowledgments

We thank Daniel Green for experimental assistance and 3 anonymous reviewers for valuable comments. This work was supported by research grants from the University of Michigan and the National Institutes of Health to J.Z.

Literature Cited

Balzar M, Briaire-de Bruijn IH, Rees-Bakker HA, et al. (11 co-authors). 2001. Epidermal growth factor-like repeats mediate lateral and reciprocal interactions of Ep-CAM molecules in homophilic adhesions. Mol Cell Biol 21:2570–80.

Ensslin MA, Shur BD. 2003. Identification of mouse sperm SED1, a bimotif EGF repeat and discoidin-domain protein involved in sperm-egg binding. Cell 114:405–17.

Goodman M, Grossman LI, Wildman DE. 2005. Moving primate genomics beyond the chimpanzee genome. Trends Genet 21:511–7.

Ishii M, Kanai Y, Kanai-Azuma M, Tajima Y, Wei TT, Kidokoro T, Sanai Y, Kurohmaru M, Hayashi Y. 2005. Adhesion activity of fetal gonadal cells to EGF and discoidin domains of milk fat globule-EGF factor 8 (MFG-E8), a secreted integrin-binding protein which is transiently expressed in mouse early gonadogenesis. Anat Embryol (Berl) 209:485–94.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150–63.

Kvistgaard AS, Pallesen LT, Arias CF, Lopez S, Petersen TE, Heegaard CW, Rasmussen JT. 2004. Inhibitory effects of human and bovine milk constituents on rotavirus infections. J Dairy Sci 87:4088–96.

Lawrence N, Klein T, Brennan K, Martinez Arias A. 2000. Structural requirements for notch signalling with delta and serrate during the development and patterning of the wing disc of Drosophila. Development 127:3185–95.

Murphy WJ, Pevzner PA, O'Brien SJ. 2004. Mammalian phylogenomics comes of age. Trends Genet 20:631–9.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Shi J, Gilbert GE. 2003. Lactadherin inhibits enzyme complexes of blood coagulation by competing for phospholipid-binding sites. Blood 101:2628–36.

Shur BD, Ensslin MA, Rodeheffer C. 2004. SED1 function during mammalian sperm-egg adhesion. Curr Opin Cell Biol 16: 477–85.

Silvestre JS, Thery C, Hamard G, et al. (17 co-authors). 2005. Lactadherin promotes VEGF-dependent neovascularization. Nat Med 11:499–506.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet 3:137–44.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–82.

Watanabe T, Totsuka R, Miyatani S, Kurata S, Sato S, Katoh I, Kobayashi S, Ikawa Y. 2005. Production of the long and short forms of MFG-E8 by epidermal keratinocytes. Cell Tissue Res 321:185–93.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–6.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–50.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50:56–68.

Zhang J, Kumar S, Nei M. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. Mol Biol Evol 14:1335–8.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–9.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95:3708–13.

Sudhir Kumar, Associate Editor

Accepted July 18, 2006