Accelerated molecular evolution of insect orthologues of *ERG28*/*C14orf1*: a link with ecdysteroid metabolism?

REINER A. VEITIA[1]* and LAURENCE D. HURST[2]†

[1]*Unité d'Immunogénétique Humaine/Université Denis Diderot – Paris VII, Institut Pasteur,*
25 rue du Dr Roux, 75015 Paris, France
[2]*Department of Biology and Biochemistry, University of Bath, Claverton Down,*
Bath, Somerset BA2 7AY, UK

Abstract

We have analysed the evolution of *ERG28*/*C14orf1*, a gene coding for a protein involved in sterol biosynthesis. While primary sequence of the protein is well conserved in all organisms able to synthesize sterols *de novo*, strong divergence is noticed in insects, which are cholesterol auxotrophs. In spite of this virtual acceleration, our analysis suggests that the insect orthologues are evolving today at rates similar to those of the remaining members of the family. A plausible way to explain this acceleration and subsequent stabilization is that Erg28 plays a role in at least two different pathways. Discontinuation of the cholesterogenesis pathway in insects allowed the protein to evolve as much as the function in the other pathway was not compromised.

[Veitia R. A. and Hurst L. D. 2001 Accelerated molecular evolution of insect orthologues of *ERG28*/*C14orf1*: a link with ecdysteroid metabolism? *J. Genet.* **80**, 17–21]

Introduction

In a previous work we have cloned a human gene that codes for a basic membrane protein that is conserved across eukarya. The *C14orf1*, which maps to chromosome 14q24.3, was shown to be the orthologue of budding yeast *Yer044c* (Veitia *et al*. 1999; Ottolenghi *et al*. 2000). A systematic global transcription profiling using microarrays has shown that the profile generated by the yeast mutant for the gene (*Yer044c*) was profoundly close to those of several mutants in the ergosterol biosynthesis pathway (*erg*). This genomic approach suggested a functional link between *Yer044c* and sterol biosynthesis in yeast (Hughes *et al*. 2000). Further work in the characterization of the effect of the mutation on sterol biosynthesis has shown that the mutant accumulates 3-keto and 4-carboxyl sterol intermediates (Gachotte *et al*. 2001). This clearly suggests the existence of a defect in sterol C-4 demethylation. Removal of C-4 methyl groups is a complex reaction involving the products of genes *ERG25*, *ERG26* and *ERG27*. First, Erg25 (the C-4 methyloxidase) converts the C-4 methyl group to a carboxylate. Formation of a keto group after oxidation of the C-3 OH destabilizes the structure and leads to decarboxylation (Erg26: dehydrogenase/decarboxylating). Then, Erg27 (reductase) restores the C-3 OH by reducing the C-3 carbonyl group. Since the mutant yeast is able to synthesize ergosterol, the end product of the pathway, it is clear that Erg28 plays an indirect role in the demethylation reaction. It has been proposed that Erg28 might anchor the proteins Erg26 and Erg27 to the endoplasmic reticulum membranes because they apparently lack obvious transmembrane domains (Gachotte *et al*. 2001).

A comparison of the genomic structure of several *ERG28* orthologues allowed us to detect introns with strictly conserved positions in *Schizosaccharomyces pombe*, *Arabidopsis thaliana* and mammals (Ottolenghi *et al*. 2000). We have considered that primary sequence and intron position conservation through an extremely

*For correspondence. E-mail: rveitia@pasteur.fr.
†l.d.hurst@bath.ac.uk.

Keywords. ergosterol; cholesterol; steroids; ecdysteroids; Erg28; C14orf1.

wide evolutionary scale is enough to suggest common ancestry for all genes. Further, demonstration of orthology and conservation of function has been obtained by complementation of the yeast *erg28* mutant by the human homologue (Gachotte *et al*. 2001).

Results and discussion

We have extensively searched for prokaryotic homologues of the protein. However, we have failed to find any sequence bearing significant similarity to Erg28. We have therefore claimed that the whole protein or at least its central portion (most of which is encoded by the conserved exons) could be considered as an ancient eukaryote-specific conserved region (Ottolenghi *et al*. 2000). This makes sense, since sterol synthesis is specific to eukaryotes. Considering that the proteins have been conserved from yeast to man in sequence and function, we have been troubled by the fact that neither *Drosophila melanogaster* nor *Caenorhabditis elegans* have clear homologous proteins (Ottolenghi *et al*. 2000). Again the implication of *ERG28* in sterol biosynthesis explains its 'virtual' absence in these organisms, known to be unable to synthesize cholesterol *de novo*. While the homologue in *C. elegans* is still elusive we could detect the orthologue in *D. melanogaster*. We have compared *ERG28* genes representative of yeasts, plants and mammals against the nonredundant division of GenBank using the program PSI-BLAST. When using a protein sequence from the higher eukaryotes, a protein from *D. melanogaster* with a similar length was consistently retrieved though marginally detected (CG17270; BLAST's S ~ 40 bits, E ~ 0.05). This was also the case in any further iteration, unless the *Drosophila* sequence itself was included to build the PSI-BLAST model. In the backward search, the sequence from *Drosophila* weakly detected the orthologues of Erg28. In spite of the poor statistical significance of the alignments we decided to further analyse the gene. Analysis of genomic structure showed that the open reading frame is split into four exons as in mammals and that the second and third exons have the same lengths and symmetry in both types of organisms (figure 1). The analysis of the hydrophobic profiles (Kite and Doolittle) of the proteins also revealed a structural connection between all of them (including similar topology in the membrane). At this point we considered that CG17270 is a divergent orthologue of Erg28. It is noteworthy that CG17270 was the only sequence to be consistently and reproducibly detected in our searches. This rules out the possibility of CG17270 being a divergent *Drosophila* paralogue.

All Erg28 orthologues that we have detected so far are basic proteins, even the divergent *Drosophila* protein. Yeast Erg28 has a calculated *p*I = 9.5 (ratio HKR/ED = 2.13) while the *p*I for the human protein is 10.4 (HKR/ED = 2.83). For the rest of the proteins HKR/ED > 1.85. However, only two basic sites (R in the RTFG motif and H in ϕHFϕ(S/T)E, ϕ = hydrophobic amino acid) are almost universally conserved (R universal, H absent in *Drosophila*). From this, it is clear that basic character is required for protein function and that this property has been maintained through a wide evolutionary span and even in insects.

To further study the evolution of Erg28 along eukarya, we undertook tests of the variation in the rate of evolution between lineages. The amino acid sequences were aligned using the program ClustalX (Thompson *et al*. 1997). The nucleotide alignment was reconstructed using MRTRANS (written by Bill Pearson). From this, a neighbour-joining tree was constructed (figure 2A).

The topology of the tree and the estimated branch lengths were used to initialize a maximum-likelihood analysis of the mode of evolution of the sequences using the program PAML (Yang 1997). Most particularly, we wished to examine the ratio of the number of nonsynonymous substitutions per nonsynonymous site (K_a) and the number of synonymous substitutions per synonymous site (K_s). The ratio ($w = K_a/K_s$) is indicative of the mode of evolution operating on the sequences. If selection is dominantly purifying then we expect few nonsynonymous changes per background synonymous change and hence a low ratio. If selection is absent then a ratio of unity is expected. To examine whether there are differences in the rate of evolution between branches, two models were compared.

In model 1 all branches have the same K_a/K_s ratio. The log likelihood observed using this model was – 3598.186, and the mean K_a/K_s ratio was 0.16. This suggests that the sequence is predominantly under purifying selection. The tree is described in figure 2B. Note the especially long branch (36.068) leading to the two insect species. Under model 2 (free ratio model) we assume a different K_a/K_s ratio for each branch of the tree. The likelihood of this model is – 3594.819. There are 13 branches (hence 13 K_a/K_s ratios), so in the comparison there are 12 degrees of freedom for the chi-squared statistic, which is two times

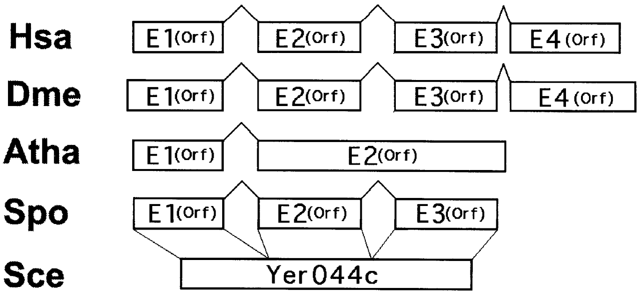

Figure 1. Schematic representation of the genomic structures of *ERG28* orthologues. Hsa, *Homo sapiens*; Dme, *Drosophila melanogaster*; Atha, *Arabidopsis thaliana*; Spo, *Schizosaccharomyces pombe*; Sce, *Saccharomyces cerevisiae*.

the difference in the two likelihoods, which equals 6.73, which is not significant ($P \gg 0.05$). Therefore the model suggesting that there is variation in all branches in K_a/K_s does not fit the data better.

Nonetheless, examination of the branch lengths indicates, by visual inspection, that the one leading to the insects might be unusual. We can test whether this is different by allowing all branches except this one to have the same rate (model 3). Applying this method resolves to a likelihood of – 3597.922, with the branch in question having a K_a/K_s ratio now of 1.144, all others having a ratio of 0.16. The difference between this and model 1 is not significant ($P \gg 0.05$). We conclude that there is no support for the proposition that the K_a/K_s ratio of this branch is different from that of the others. This suggests that the mode of evolution between branches is probably not different. However, as estimation of K_s over distant comparisons is of dubious value, these results should be interpreted with some caution.

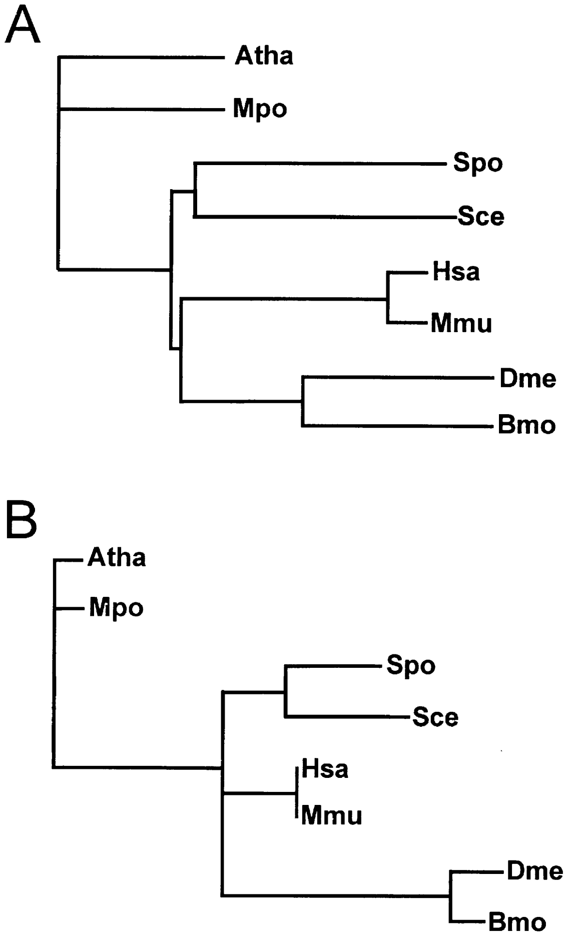

Figure 2. A, Neighbour-joining tree (amino acid sequences aligned using ClustalX and nucleotide alignment reconstructed using MRTRANS). B, Tree from a maximum-likelihood analysis of sequences (model 1, see text). Accession identifiers: Atha, *Arabidopsis thaliana* (AAC34343); Mpo, *Marchantia polymorpha* (C95915); Spo, *Schizosaccharomyces pombe* (CAA21279); Sce, *Saccharomyces cerevisiae* (P40030); Hsa, *Homo sapiens* (AF134159); Mmu, *Mus musculus* (AF270646); Dme, *Drosophila melanogaster* (consensus: AE003732/AA735889/AI23922); Bmo, *Bombyx mori* (AU006388). (Sequence alignments and statistical details of tree construction available from authors.)

While the mode of evolution may not be different, it does not follow that the insect branch is not unusually long. To examine this issue, we can apply a relative rate test. For example, taking *S. pombe* as the outgroup, we can compare human and *Drosophila*. This reveals a highly significant deviation from null expectations: using Tajima's 2D method (Tajima 1993) we get 71.138 ($P < 0.001$) for the sequence overall. Using *Bombyx* instead of *Drosophila* gives the same conclusions (Tajima 2D 61.557, $P < 0.001$). This appears to be most profound at coding sites: human/*Bombyx* – four-fold redundant silent sites, Tajima 2D = 4.172, ns, versus 21.728, $P < 0.001$ and 27.200, $P < 0.001$ for first and second sites respectively; human/*Drosophila*, four-fold redundant silent sites, Tajima 2D = 3.933, ns, versus 25.200, $P < 0.001$ and 39.497, $P < 0.001$ for first and second sites respectively.

By contrast, taking human and mouse with *S. pombe* as the outgroup, we find no evidence for anything other than clock-like behaviour (the Tajima 2D test gives 0.835, $P \gg 0.05$). Comparing *Drosophila* and *Bombyx*, with humans as an outgroup, suggests a weak deviation from clock-like behaviour (2D test = 7.806, $P < 0.05$). Here, however, there is no deviation at protein coding sites, only at silent sites (first and second sites: 1.143, ns, and 1.667, ns, four-fold redundant sites, 9.500, $P < 0.01$). In sum, this suggests that the non-clock-like behaviour is unique to the lineage leading to insects and possibly persisting within it, at least at silent sites.

Overall, therefore, we have (a) evidence for an acceleration in rate of evolution in the lineage leading to insects, but (b) no evidence that the K_a/K_s ratio changed during this phase.

We have taken advantage of the fact that the full genomic sequence of *Drosophila* is, in principle, available and searched for sequences homologous to the enzymes involved in some steps of sterol biosynthesis. Notably, squalene synthase (Erg9), which catalyses the first committed step in cholesterogenesis, was absent. This was expected taking into account biochemical evidence (Svoboda and Feldlaufer 1991). Squalene epoxidase (Erg1) and lanosterol synthase (Erg7), which follow in the classical pathway, also went undetected. Similarity hits with Cyp51, the C-14 demethylase (Erg11), were obtained (CG2397, CG10247 and other Cyps). However, careful analysis of CG2397 and CG10247 showed that they are probably Cyp6a proteins. This suggests that Cyp51 is probably missing. A C-14 reductase (Erg24) homologue was detected (CG17952) by BLAST and the same sequence was retrieved in a PSI-BLAST search using Erg4 (the C-24 reductase) as the query. Similarity between the two reductases (Erg4 and Erg24) has already been noticed (Lees *et al*. 1995). Homologous sequences for Erg25 (C-4 methyloxidase) and Erg26 (C-3 dehydrogenase) were found (CG11162/CG1998 and CG7724/CG12030, respectively). In the case of CG7724 similarity with yeast Erg26 extends over the first 250 amino acids while the remaining 150 amino acids are less conserved. Although CG12030 was also retrieved, it is more likely to be an epimerase/dehydratase. We failed to find any clear homologue of Erg27 (C-3 ketoreductase), although several oxidoreductases were detected. C-8 isomerase homologues (Erg2) were missing and proteins similar to Erg3 (C-5 desaturase) were found. From this information it is clear that *Drosophila* is far from being able to fully synthesize cholesterol from the isoprenoid precursors. It is, however, intriguing that sequence homologues of Erg25, Erg26 and Erg28 are present.

Gachotte *et al*. (2001) have noticed that the *erg28* mutant yeast strain accumulates 3-keto and 4-carboxylic acid sterols and this is probably because the reaction is not well streamlined. Several scenarios can be proposed that may explain the accumulation of oxidized intermediates (detectable in the steady state in the mutant but not in the wild type). If we follow the idea of Erg28 being an anchor, once the carboxylic intermediate is produced in the mutant, it would have to wait for Erg26 to approach the endoplasmic reticulum (ER) membranes to be dehydrogenated at position 3 (lower local concentration of the enzyme because of the absence of the 'anchor'). The same is valid for Erg27. Notice that while accommodation in the membrane of keto sterols seems easy, this is less obvious for a carboxylic derivative at position 4. In the latter case the amphipathic behaviour of cholesterol is 'blurred' by the introduction of an ionizable carboxyl group in a region of the molecule that contributes to hydrophobicity. Handling this intermediate (which might not insert readily in the membrane!) can be difficult unless there is physical proximity between Erg25 (the methyloxidase), Erg26 and Erg27. From this, Erg28, in addition to anchoring Erg26 and Erg27 to the ER membranes, could promote their interaction with Erg25. As discussed above, Erg28 proteins are all basic. They could anchor Erg26 and Erg27 and mediate their interaction with Erg25 via electrostatic interactions with acidic domains. Erg28 itself could also help handle the carboxylic and 3-keto intermediates through interaction between the basic side-chains and the sterol polar groups. Whatever the precise molecular effect of Erg28 is, it can be considered as a channelling factor.

Speculations

We have shown here that Erg28 homologues in the insects have undergone a process of acceleration in their evolution. However, it seems that these sequences are now evolving at rates similar to those seen in the rest of the eukaryotes, at least at protein coding sites. Since insects are unable to synthesize cholesterol *de novo*, it is likely that the selective constraints operating on Erg28 do not result from the utilization of a segment of this pathway. We cannot exclude the possibility that the protein was recruited to perform a novel function. However, an attractive way to explain the acceleration and subsequent stabilization in Erg28 evolution in insects is that this protein plays a role in at least two different pathways. Removal of the putative interacting partners in cholesterol synthesis in insects allowed the protein to evolve as much as the function in the other pathway was not compromised. We asked what the other pathway could be, taking into account that Erg28 somehow interacts with Erg26/Erg27. The most obvious possibility is its implication in ecdysteroid metabolism (the pathway for biosynthesis of the main hormones in arthropods).

Erg26 is responsible for the dehydrogenation of the C-3 hydroxyl group in a NAD^+-dependent way. Recent work has shown the existence of a membrane-bound NAD^+-dependent 3*b*-hydroxysteroid dehydrogenase (3*b*-HSD, EC 1.1.1.145) involved in ecdysteroid biosynthesis (Dauphin-Villemant *et al*. 1997). Erg26 and 3*b*HSD are structurally related and might have similar catalytic mechanisms. So it is tempting to speculate that CG7724, the only Erg26 homologue in *Drosophila*, is involved in ecdysteroid metabolism and somewhat interacts with Erg28/CG17270. In line with this hypothesis is the fact that the short isoform of the human *ERG28* transcript is highly and specifically expressed in the testis (Veitia *et al*. 1999). This gene could act at two levels: (i) synthesis of cholesterol, which will be further transformed by the steroidogenic cells, and (ii) directly in steroid synthesis by interacting with the 3*b*HSD which is highly expressed in steroidogenic cells.

When performing a BLAST search using CG7724 most sequences retrieved corresponded to members of the 3*b*HSD family which are also $\Delta^{5,4}$ isomerases. Interestingly, the 3-keto intermediates discovered so far in ecdysteroid metabolism are also Δ^4 which presupposes the existence of a $\Delta^{5,4}$ isomerase activity (Blais *et al*. 1996). The direct mammalian homologue of Erg26 seems to be NSDHL (Konig *et al*. 2000), and comparison of CG7724 with the latter and other members of the 3*b*HSD family shows that the gene in *Drosophila* is closer to the classical dehydrogenases/isomerases. The latter are membrane proteins of the ER and mitochondria (at least the types I and II, Thomas *et al*. 1999). So the hypothetical interaction between CG7724/3*b*HSD and CG17270/Erg28 could have a regulatory role or be involved in some type of channelling as suggested above. Ecdysteroids are also detected in *C. elegans*, which almost certainly lacks Erg28. However, their relevance or even their origin (phytoecdysteroids or endogenously modified) are not well established in the current literature.

Acknowledgements

The authors wish to thank C. Dauphin-Villemant and R. Lafont from the Unité d'Endocrinologie Moléculaire et Evolution (Université Paris VI) for reading the manuscript and for helpful discussions.

References

Blais C., Dauphin-Villemant C., Kovganko N., Girault J. P., Descoins C. and Lafont R. 1996 Evidence for the involvement of 3-oxo-D4 intermediates in ecdysteroid biosynthesis. *Biochem. J.* **320**, 413–419.

Dauphin-Villemant C., Bocking D., Blais C., Toullec J. Y. and Lafont R. 1997 Involvement of a 3*b*-hydroxysteroid dehydrogenase activity in ecdysteroid biosynthesis. *Mol. Cell. Endocrinol.* **128**, 139–149.

Gachotte D., Eckstein J., Barbuch R., Hughes T., Roberts C., and Bard M. 2001 A novel gene conserved from yeast to humans is involved in sterol biosynthesis. *J. Lipid Res.* **42**, 150–154.

Hughes T., Marton M. J., Jones A. R., Roberts C. J., Stoughton R., Armour C. D., Bennett H. A., Coffey E., Dai H., He Y. D., Kidd M. J., King A. M., Meyer M. R., Slade D., Lum P. Y., Stepaniants S. B., Shoemaker D. D., Gachotte D., Chakraburtty K., Simon J., Bard M. and Friend S. H. 2000 Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126.

Konig A., Happle R., Bornholdt D., Engel H. and Grzeschik K. H. 2000 Mutations in the NSDHL gene, encoding a 3*b*-hydroxysteroid dehydrogenase, cause CHILD syndrome. *Am. J. Med. Genet.* **90**, 339–346.

Lees N. D., Skaggs B., Kirsch D. R. and Bard M. 1995 Cloning of the late genes in ergosterol biosynthetic pathway of *Saccharomyces cerevisiae*—a review. *Lipids* **30**, 221–226.

Ottolenghi C., Daizadeh I., Ju A., Kossida S., Renault G., Jacquet M., Fellous A., Gilbert W. and Veitia R. 2000 The genomic structure of C14orf1 is conserved across eukarya. *Mamm. Genom.* **11**, 786–788.

Svoboda J. A. and Feldlaufer M. F. 1991 Neutral sterol metabolism in insects. *Lipids* **26**, 614–618.

Tajima F. 1993 Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**, 599–607.

Thomas J. L., Evans B. W., Blanco G., Mason J. I. and Stricker R. C. 1999 Creation of a fully active, cytosolic form of type I 3-beta-hydroxysteroid dehydrogenase/isomerase by deletion of a membrane spanning domain. *J. Mol. Endocrinol*. **23**, 231–239.

Thompson J. D., Gibson T. J., Plewniak F., Jeanmougin F. and Higgins D. G. 1997 The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* **25**, 4876–4882.

Veitia R., Ottolenghi C., Bissery M. C. and Fellous A. 1999 A novel human gene, encoding a potential membrane protein conserved from yeast to man, is strongly expressed in testis and cancer cell lines. *Cytogenet. Cell Genet.* **85**, 217–220.

Yang Z. 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13**, 555–556.

Received 28 March 2001