

Structural bioinformatics

Lateral gene transfer of a dermonecrotic toxin between spiders and bacteria

Matthew H. J. Cordes^{1,*} and Greta J. Binford^{2,*}¹Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, AZ 85721, USA and ²Department of Biology, Lewis and Clark College, Portland, OR 97219, USA

Received on September 25, 2005; revised on November 23, 2005; accepted on December 1, 2005

Advance Access publication December 6, 2005

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Spiders in the genus *Loxosceles*, including the notoriously toxic brown recluse, cause severe necrotic skin lesions owing to the presence of a venom enzyme called sphingomyelinase D (SMaseD). This enzyme activity is unknown elsewhere in the animal kingdom but is shared with strains of pathogenic *Corynebacteria* that cause various illnesses in farm animals. The presence of the same toxic activity only in distantly related organisms poses an interesting and medically important question in molecular evolution.

Results: We use superpositions of recently determined structures and sequence comparisons to infer that both bacterial and spider SMaseDs originated from a common, broadly conserved domain family, the glycerophosphoryl diester phosphodiesterases. We also identify a unique sequence/structure motif present in both SMaseDs but not in the ancestral family, supporting SMaseD origin through a single divergence event in either bacteria or spiders, followed by lateral gene transfer from one lineage to the other.

Contact: cordes@email.arizona.edu; binford@lclark.edu

INTRODUCTION

Venoms from *Loxosceles* spiders, also called brown or violin spiders, cause severe dermonecrosis in humans. The single venom toxin sphingomyelinase D (SMaseD) causes the complete dermonecrotic syndrome in animal models (Fernandes-Pedrosa *et al.*, 2002; Ramos-Cerillo *et al.*, 2004; Tambourgi *et al.*, 2004). SMaseD activity has not been found elsewhere in the animal kingdom or in any other organisms except bacterial pathogens in the genus *Corynebacteria* (Soucek *et al.*, 1967). Spider and bacterial SMaseDs are similar in molecular weight and isoelectric focusing point (Bernheimer *et al.*, 1985; Truett and King, 1993) as well as enzyme kinetics and substrate specificities (van Meeteren *et al.*, 2004) hinting that they may be related by common ancestry. Unfortunately, simple comparisons of the amino acid sequences of the spider and bacterial proteins were unable to confirm this homology because the sequences are too dissimilar (Binford *et al.*, 2005). However, the structure of the spider SMaseD was recently reported (Murakami *et al.*, 2005) and an active site proposed. On the basis of sequence alignments the key putative active-site residues were suggested to be identical in bacterial SMase, despite the overall dissimilarity of the sequences. The apparent presence of a conserved catalytic core

in the two toxins, with a conserved order in the sequence, lends further support for their distant homology.

The present study relates to the following mystery: how does it come to pass that two similar, medically important protein toxins, putatively sharing a common evolutionary origin, are found in two very dissimilar types of organism but nowhere else? The presence of homologous SMaseDs only in select, very distantly related taxa could be explained either by independent divergence from the same broadly conserved protein family or by a single divergence event followed by lateral gene transfer (LGT) between spiders and bacteria. Although LGT is a major mechanism in the evolution of bacterial genomes (Brown, 2003), and toxin-encoding genes are particularly prone to lateral mobility (Hacker *et al.*, 2004), examples of origin of novel gene function through lateral transfer between eukaryotes and bacteria are rare. Here, through superposition of recently determined structures, we first confirm that the SMaseDs both originated by divergence from a domain family known as GDPDs (glycerophosphoryl diester phosphodiesterases), which has representatives in all major classes of organisms. We then identify a strongly conserved but highly unusual structural motif in the spider and bacterial SMaseDs that is not present in GDPDs, implying evolution of both toxins through the same divergence event. This supports LGT as an explanation for the presence of these medically important toxins in both bacteria and spiders.

MATERIALS AND METHODS

Coordinates for *Loxosceles* SMaseD (1XX1) and various GDPDs (1YDY, 1T8Q, 1V8E and 1O1Z) were obtained from the Protein Data Bank (PDB, available at www.resb.org/pdb). Three of these are structures from structural genomics projects which have not been published in journals: 1T8Q (Midwest Center for Structural Genomics; Zhang *et al.*, unpublished data), 1YDY (New York Structural Genomics Research Consortium; Malashkevich *et al.*, unpublished data), and 1V8E (Riken Structural Genomics/Proteomics Initiative; Ishijima *et al.*, unpublished data). The Swiss-Prot ID for the bacterial toxin sequence from *Corynebacterium pseudotuberculosis* is PLD_CORPS. Active-site superpositions were performed iteratively using Deep View, beginning from superpositions of the overall TIM barrel structure. Sequence pattern searches were performed on the Swiss-Prot protein sequence database using MyHits (Pagni *et al.*, 2004) (http://myhits.isb-sib.ch/cgi-bin/pattern_search). Consensus structure predictions for the bacterial toxin were performed using 3D-Jury (Ginalski *et al.*, 2003,

*To whom correspondence should be addressed.

<http://bioinfo.pl/meta>). Main-chain structural motif searches were performed using Dennis Madsen's server (<http://portray.bmc.uu.se/cgi-bin/spasm/scripts/spasm.pl>) for the SPASM program (Kleywegt, 1999). PSI-BLAST searches (Altschul *et al.*, 1997) were performed using the NCBI server (<http://www.ncbi.nlm.nih.gov/BLAST>).

RESULTS AND DISCUSSION

Just prior to the publication of the spider SMaseD structure, we reported evidence for distant homology between spider SmaseD and the ubiquitous protein domain family GDPD (Binford *et al.*, 2005). At the time, the primary evidence for this relationship was a significant *E*-value in a Pfam HMM search. A 3D-PSSM search had also identified the single published representative GDPD structure (from *Thermotoga maritima*; Santelli *et al.*, 2004) as the best fold recognition hit for the spider SMaseD. PSI-BLAST searches (Altschul *et al.*, 1997) initiated from individual spider toxins did not yield any GDPDs as hits. More recently, however, the deposition of new spider toxin homolog sequences into databases has yielded better sequence conservation profiles, such that PSI-BLAST now gives annotated GDPDs as hits (*E*-value < 0.005) as early as the second round. Thus, although no direct pairwise similarity exists between SMaseD and any annotated GDPD, the broad similarity in the chemistry catalyzed by SMaseD and GDPDs, coupled with the Pfam, 3D-PSSM and PSI-BLAST hits, indicates that spider SMaseD diverged from this family.

Structural comparisons between GDPDs and spider SMaseD further cement this conclusion. In addition to the *Thermotoga maritima* GDPD structure, there are now structures of two *Escherichia coli* GDPDs and one *Thermus thermophilus* GDPD in the PDB. The four GDPDs and *Loxosceles laeta* SMase I share an eight-stranded $\beta\alpha$ barrel or 'TIM barrel' fold commonly found in enzymes. Structural superpositions (Fig. 1) show that the GDPDs contain a set of well-conserved residues at the C-terminal end of the barrel that

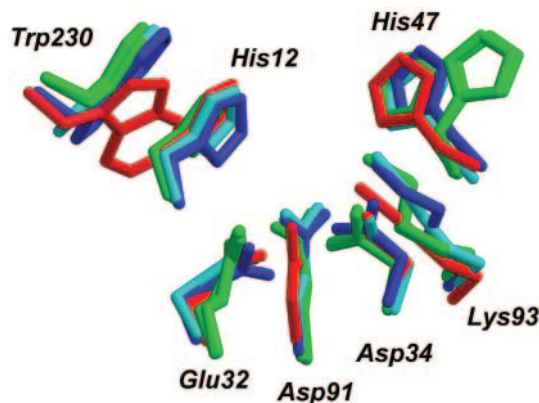


Fig. 1. Partial active-site superposition of *Escherichia coli* cytoplasmic and periplasmic GDPDs (three structures 1T8Q_A, 1YDY_A and 1YDY_B; all in green), *Thermotoga maritima* GDPD (1O1Z; cyan), *Thermus thermophilus* GDPD (1V8E; blue) and *Loxosceles laeta* SMase I (1XX1_A; red). For 1T8Q and 1XX1, all chains in the asymmetric unit showed similar rotamers for all residues. For 1YDY, one of the two chains showed a conformation for His 47 (SMase I numbering) similar to that of the *E. coli* cytoplasmic GDPD (1T8Q), while the other showed a conformation similar to that of the other GDPDs and the *L. laeta* SMase I. Superpositions were performed using DeepView.

clearly correspond to the putative active site identified for SMase I. Of the seven residues shown, five (His12, Glu32, Asp34, His47 and Lys93) are perfectly conserved in the bacterial GDPDs and occupy similar relative positions in three-dimensional (3D) space (Fig. 1 and 2). The other two, Asp91 and Trp230, have conservative mutations. Asp91 of SMase I is replaced by Glu in all the GDPDs, while Trp230 of SMase I is replaced by Tyr in the *E. coli* proteins. The active-site similarity leaves no doubt that the spider enzymes and the GDPDs are distantly homologous.

Although there is no experimental structure of a corynebacterial toxin to add to this comparison, and the pairwise sequence similarity to the spider toxins and the GDPDs is insignificant, several observations support inclusion of the bacterial SMases in this homologous group. First, the key active-site residues shown in Figure 1 appear to be conserved. Figure 2 shows a sequence alignment of the GDPDs and the spider toxin generated based on their structure superposition using Deep View. Using ClustalX we added the SMase sequence from *C. pseudotuberculosis* to this alignment. The seven active-site residues of Figure 1 are well conserved in the bacterial toxin, though His 47 is misaligned. This apparent conservation was also noted by Murakami *et al.* (2005) (see Introduction). Second, recent PSI-BLAST searches initiated from individual spider toxin sequences (see above) give near-hits to two bacterial toxins from *C. pseudotuberculosis* in the third round (*E*-values of 0.005–0.01). Third, structures of bacterial SMases computationally predicted by different methods exhibit a consensus consistent with homology to GDPDs and the spider toxins. In a 3D-Jury (Ginalski *et al.*, 2003) structure prediction performed on the *C. pseudotuberculosis* sequence, the top five scoring models were all based on spider SMaseD and the next three highest-scoring models were based on three different GDPD template structures. All eight top models had jury scores above 55, indicating a high degree of consensus.

If the two toxins and the GDPDs therefore belong to a common evolutionary lineage, three possible mechanisms may be envisioned for the origin of SMaseD activity in bacteria and spiders (Fig. 3): (A) independent gene duplication events of GDPD family members in bacteria and spiders followed by independent functional divergence to SMaseD in both; (B) gene duplication of a GDPD in spiders followed by functional divergence and LGT to bacteria; (C) gene duplication of a GDPD in bacteria followed by functional divergence and LGT to spiders. Note that in both the LGT mechanisms (B and C) it is also possible (scenario not shown) that the lateral transfer could have occurred prior to the origin of SMaseD activity, with the toxic activity independently originated twice at a later time in both lineages.

Phylogenetic trees based on multiple sequence alignments could in principle be used to distinguish LGT (B and C) from non-LGT (A) mechanisms as well as to infer the direction (B versus C) and timing of a putative LGT transfer. However, although sequences for over 15 spider toxins, 5 bacterial toxins and hundreds of GDPDs are available, a useful analysis of this kind is not feasible here due to limitations imposed by high sequence divergence (illustrated by the alignment in Fig. 2) and a lack of appropriate GDPD taxon sampling. At the low levels of identity present (~20% or less between members of the three groups), alignment errors and uncertainties in phylogenetic models limit confidence in tree topology. Moreover, the use of a phylogeny to infer the timing and direction of a putative lateral transfer event depends upon appropriate taxon

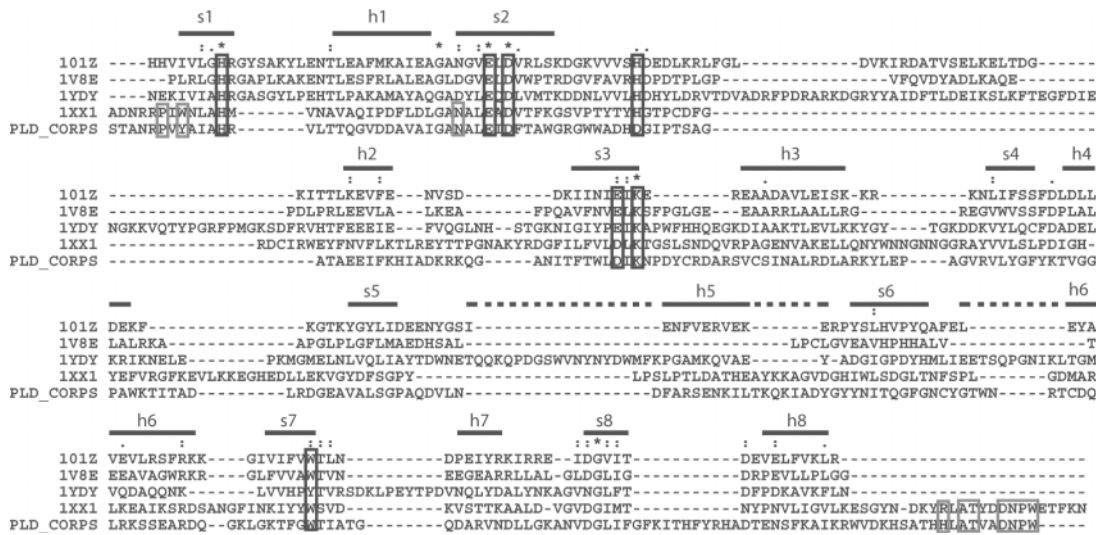


Fig. 2. Structure-based sequence alignment of three GDPDs (101Z, 1YDY and 1V8E) with *L.laeta* SMase I (1XX1), generated using DeepView. The sequence of *Corynebacterium pseudotuberculosis* SMase (PLD_CORPS) has been added to this alignment using ClustalX. Conserved active-site residues are highlighted in black, while conserved structurally important residues in the ‘plug motif’ common to the two toxins are highlighted in gray. Sequence positions of the eight strands and helices for the TIM barrel in 101Z are also shown.

sampling of GDPDs, including family members from lineages containing SMaseD. Here we are limited by the absence of a sequenced genome for a representative spider lineage.

As an alternative to a tree-based approach, we reasoned that the new structural data and structure-guided sequence comparisons might help us to identify motifs or unique features present exclusively within some subset of the GDPD/toxin proteins. This could aid in broadly distinguishing between the LGT mechanisms (B and C) and independent divergence (A). For example, suppose that the spider toxin has a sequence/structure motif not present in the GDPDs, but that sequence comparisons between the bacterial and spider SMaseDs supported conservation of this feature in both toxins. If such a motif were sufficiently unique that it was unlikely to have arisen twice independently, its presence in both toxins but not in non-SMaseD GDPDs would constitute evidence for the two toxins sharing a more recent common ancestor with each other than either does with the other GDPDs. Such a finding would be consistent with the LGT mechanisms (B) and (C) but not with the independent divergence mechanism (A) (Fig. 3).

Such a signature sequence/structure motif indeed exists in the spider toxin. Following the last helix of the TIM barrel fold, *Loxosceles* SMase I contains a stretch of sequence (residues 269–280) which essentially ‘plugs’ the end of the TIM barrel opposite the active site (Fig. 4). Residues 269–273 form a beta-strand that pairs with an extension of the barrel’s first strand (residues ~4–8). Asn 29’s side chain, at the beginning of strand 2, makes hydrogen bonds to both the end of the 269–273 strand and to strand 1, wedging the two apart. Residues 275–276 form a type I beta-turn, while the flanking residues Thr 274 and Asp 277 participate in polar interactions, including a salt bridge between Asp 277 and Arg 271. The chain then curves over the top of strand 1 of the barrel, forming a bridge over Pro 6 and Trp 8 with the side chains of Ala 273 and Pro 279 sandwiching Trp 8. Finally, Pro 279 and Trp 280 form a second type I beta-turn, with the side chain of Asn 278 contributing a

backbone hydrogen bond and the Pro and Trp side chains plugging the N-terminal end of the beta-barrel.

Sequence conservation leaves little doubt that this ‘plug motif’ is also present in the bacterial toxin. In the alignment between *Loxosceles* SMase I and *C.pseudotuberculosis* SMase shown in Figure 2, residues 5–8 and 271–280 are among the most well-conserved regions, and Asn 29 is also conserved (Fig. 4). All putatively structurally critical residues in the plug motif of SMase I are identical or similar in the bacterial toxin, while other residues such as Asp 276, which do not play any clear structural role in the motif, show less conservative substitutions. In the absence of structural information one might normally dismiss the limited sequence conservation pattern as insignificant. The coincidence of the conservation with important residues in an irregular structural motif in related proteins, however, suggests either that it reflects a feature present in a common ancestor or, less plausibly, that a remarkable convergence of sequence and structure has occurred.

The GDPDs of known structure lack any C-terminal sequence at all corresponding to spider SMaseD residues 271–280 and instead terminate with the last helix of the barrel (Fig. 2). While some annotated GDPDs of unknown structure may have longer C-termini, none has any sequence fragment reflecting the conservation seen between the bacterial and spider proteins. Specifically, sequence pattern searches against the Swiss-Prot database for the motif [HRK]xATxxDNPW using MyHits (Pagni *et al.*, 2004) (http://myhits.isb-sib.ch/cgi-bin/pattern_search) found only *Loxosceles* and *Corynebacterium* SMaseDs. The shorter motif DNPW is found in many proteins, but not in any annotated GDPD family members. Importantly, the database includes many GDPD representatives in arthropods and Corynebacteria, with 10 sequences from *Drosophila melanogaster* alone. This increases our confidence that the uniqueness of the motif is not a function of lack of representation of broadly conserved SMaseD orthologs in the database. Together, these findings point to a lack of any plug motif in GDPDs in diverse organisms that might have explained

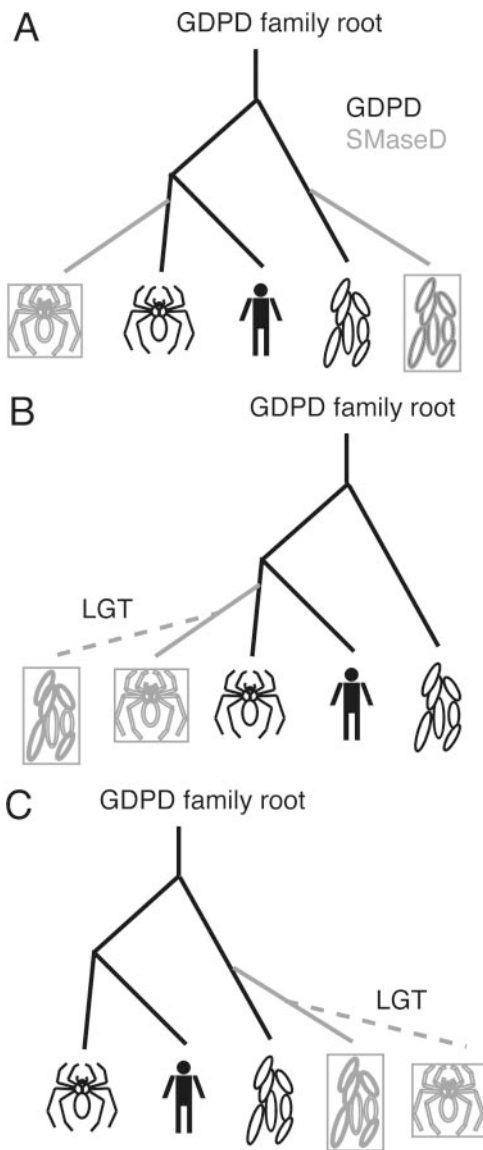


Fig. 3. Phylogenetic scenarios representing hypothetical mechanisms of origin of SMaseD activity (gray) from within the glycerophosphoryl diester phosphodiesterase (GDPD) family: (A) independent duplication and functional divergence in both bacteria and spiders; (B) duplication and functional divergence in spiders followed by LGT to bacteria; (C) duplication and functional divergence in bacteria followed by LGT to spiders. Variants of (B) and (C) are also possible in which acquisition of SMaseD activity occurs independently in bacteria and spiders following an LGT event.

its presence in both spider and bacterial toxins as a result of independent divergence from a common GDPD ancestor with this motif.

One might still propose that the spider and bacterial toxins converged on a common sequence/structure motif following independent divergence from GDPDs, perhaps as a result of a shared evolutionary pressure to gird the TIM barrel structure against a hostile environment. However, there is no evidence that the unusual loop conformation observed in spider SMaseD is a common structural solution for plugging a TIM barrel, or that use of this loop conformation would necessarily impose the precise sequence

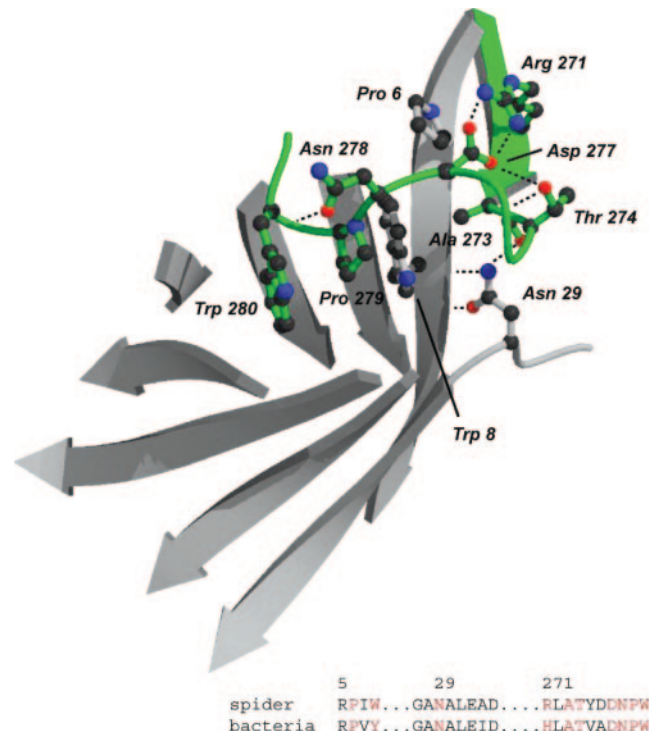


Fig. 4. Illustration of the unique C-terminal 'plug' motif of *L. laeta* SMase I, accompanied by a partial sequence alignment of *C. pseudotuberculosis* SMase and *L. laeta* SMase I, showing relevant conserved residues. The plug motif is shown in green, while the strands and selected chains of the TIM barrel are illustrated in gray. In the sequence alignment, residues which appear to play key structural roles in *L. laeta* SMase I are highlighted in red. The sequence alignment is derived from Figure 2 and the numbering reflects that of *L. laeta* SMase I.

conservation observed. Searches of the PDB using SPASM (Kleywegt, 1999) for main-chain structures within 1.2 Å C α RMSD to 271–280 of the spider SMaseD yielded 20 hits, including a single instance of a similar loop conformation perched at the N-terminal end of a TIM barrel. However, the motif adopted a very different relative orientation with respect to the barrel and had a completely different sequence. No significant sequence conservation was found among the hits to other loops of similar backbone structure. Thus, the observed conservation is almost certain to reflect homology rather than convergence.

The unique and homologous plug motif, which is present in both toxin proteins but absent in the ancestral GDPD superfamily, is a synapomorphy (shared derived character) indicating that the toxins share a more recent common ancestor with each other than either does with GDPDs. We specifically propose that the bacterial and spider toxins both descended from a single duplication of a GDPD-encoding gene (Fig. 3B and C mechanisms) followed by origination of the plug motif and retention in both toxin descendants despite the eventual loss of any other recognizable sequence similarity outside of the active site. Given the extraordinarily distant relationship between *Loxosceles* and *Corynebacteria*, and the paucity of similar proteins in other organisms, LGT is the most reasonable explanation for similarities between these toxins.

The mechanism and directionality of this transfer remain unknown. SMaseD has an ancient presence in the spider lineage (Binford and Wells, 2003). Heritable transfer from bacteria to spiders would require germ line insertion, an occurrence that has never been documented. In general, more cases of gene movement from eukaryotes to bacteria are on record (Brown, 2003). Opportunities for contact between *Corynebacteria* and *Loxosceles* exist, as the bacteria are found in soil and the spiders are ground dwelling. Furthermore, there is evidence for *Corynebacteria* being vectored among farm animals by dipteran flies (Spier *et al.*, 2004) and thus potentially being consumed by spiders.

ACKNOWLEDGEMENTS

The authors thank Wayne Maddison, Howard Ochman and Vahe Bandarian for thoughtful discussion and suggestions on the manuscript.

REFERENCES

- Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Bernheimer,A.W. *et al.* (1985) Comparative toxinology of *Loxosceles reclusa* and *Corynebacterium pseudotuberculosis*. *Science*, **228**, 590–591.
- Binford,G.J. and Wells,M.A. (2003) The phylogenetic distribution of sphingomyelinase D activity in venoms of Haplogyne spiders. *Comp. Biochem. Physiol. B*, **135**, 25–33.
- Binford,G.J. *et al.* (2005) Sphingomyelinase D from venoms of *Loxosceles* spiders: evolutionary insights from gene sequence and structure. *Toxicon*, **45**, 547–560.
- Brown,J.R. (2003) Ancient horizontal gene transfer. *Nat. Rev. Gen.*, **4**, 121–132.
- Fernandes-Pedrosa,M.F. *et al.* (2002) Molecular cloning and expression of a functional dermonecrotic and haemolytic factor from *Loxosceles laeta* venom. *Biochem. Biophys. Res. Commun.*, **298**, 638–645.
- Ginalski,K. *et al.* (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.
- Hacker,J. *et al.* (2004) Pathogenomics of mobile genetic elements of toxigenic bacteria. *Int. J. Med. Microbiol.*, **293**, 453–461.
- Kleywegt,G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, **285**, 1887–1897.
- Murakami,M.T. *et al.* (2005) Structural basis for metal ion coordination and the catalytic mechanism of sphingomyelinases D. *J. Biol. Chem.*, **280**, 13658–13664.
- Pagni,M. *et al.* (2004) MyHits: a new interactive resource for protein annotation and domain identification. *Nucleic Acids Res.*, **32** (Web Server issue), W332–W335.
- Ramos-Cerrillo,B. *et al.* (2004) Genetic and enzymatic characterization of sphingomyelinase D isoforms from the North American fiddleback spiders *Loxosceles boneti* and *Loxosceles reclusa*. *Toxicon*, **44**, 507–514.
- Santelli,E. *et al.* (2004) Crystal structure of a glycerophosphoryl diester phosphodiesterase (GDPD) from *Thermotoga maritima* at 1.60 Å resolution. *Proteins*, **56**, 167–170.
- Soucek,A. *et al.* (1967) Enzymatic hydrolysis of sphingomyelins by a toxin of *Corynebacterium ovis*. *Biochim. Biophys. Acta*, **144**, 180–182.
- Spier,S.J. *et al.* (2004) Use of a real-time polymerase chain reaction-based fluorogenic 5' nuclease assay to evaluate insect vectors of *Corynebacterium pseudotuberculosis* infections in horses. *Am. J. Vet. Res.*, **65**, 829–834.
- Tambourgi,D.V. *et al.* (2004) Molecular cloning, expression, function and immunoreactivities of members of a gene family of sphingomyelinases from *Loxosceles* venom glands. *Mol. Immunol.*, **41**, 831–840.
- Truett III,A.P., and King,J.L.E. (1993) Sphingomyelinase D: a pathogenic agent produced by bacteria and arthropods. *Adv. Lipid Res.*, **26**, 275–291.
- van Meeteren,L.A. *et al.* (2004) Spider and bacterial sphingomyelinase D target cellular lysophosphatidic acid receptors by hydrolyzing lysophosphatidylcholine. *J. Biol. Chem.*, **279**, 10833–10836.