A PHYLOGENOMIC PERSPECTIVE ON ANNELID EVOLUTION WITH EMPHASIS ON THE EVOLUTION OF BLOODFEEDING IN LEECHES (CLITELLATA: HIRUDINIDA)

A Dissertation
submitted to the Faculty of
The Richard Gilder Graduate School
at the
American Museum of Natural History
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy

By

Sebastian Kvist, M.S.

Richard Gilder Graduate School
at the
American Museum of Natural History
New York, NY
July 30, 2012

A PHYLOGENOMIC PERSPECTIVE ON ANNELID EVOLUTION WITH EMPHASIS ON THE EVOLUTION OF BLOODFEEDING IN LEECHES (CLITELLATA: HIRUDINIDA)

Sebastian Kvist, M.S.

Chair: Mark E. Siddall, Ph.D.

ABSTRACT

Annelida Lamarck, 1809 embodies over 17,000 species of segmented worms such as leeches, earthworms, lugworms, sandworms and clamworms. The phylum has traditionally been divided into two main orders: Sedentaria Lamarck, 1818 and Errantia Andouin & Milne Edwards, 1834, yet Sedentaria is seldom recovered as monophyletic and Errantia has only been recovered as monophyletic in one study. Recently, a large amino acid data set of expressed sequence tags (EST's) was created for Annelida and its close allies. The phylogenetic analyses of this data set, based on Bayesian inference and Maximum Likelihood estimations, recovered both Errantia and Sedentaria as monophyletic groups. Herein, I assess whether or not this hypothesis is also recovered by a nucleotide representation of the amino acids, and if the result is general across optimality criteria. Whereas parsimony analyses of the largest molecular character data set compiled for Annelida fails to recover Sedentaria or Errantia as monophyletic entities, re-analysis of the original amino acid data set does recover Errantia as monophyletic but with low support. In conjunction with previous studies, the analyses presented here suggest that the phylogenetic hypotheses of relationships both within

Annelida, and between the phylum and its constituent taxa are still unstable and that finding suitable data for resolving this is an important yet problematic issue.

Although medicinal leeches have long been used as treatment for various ailments because of their potent anticoagulation factors, the debates regarding the evolution of bloodfeeding and the ancestral feeding preference of leeches are still contentious. Moreover, neither the full diversity of salivary components that inhibit coagulation, much less the evolutionary selection acting on them, has been thoroughly investigated in a comparative manner. To address these questions, the full genome of the non-bloodfeeding leech *Helobdella robusta* was screened for anticoagulation factors. In addition, EST libraries from salivary glands of two species of medicinal leeches, *Hirudo verbana* and *Aliolimnatis fenestrata*, were constructed. For *H. robusta*, a total of eight loci matching leech antiplatelet proteins and positioned as a tandem array, were recovered with significantly low e-values, suggesting that this leech possesses ancestrally-inherited anticoagulants. In the medicinal leeches, expression of salivary peptides greatly exceed expectations and also suggest the feasibility of identifying the important active sites of the proteins through selective pressure analyses.

Although symbiotic associations between bacteria and leeches are well documented, several questions concerning the function of the bacterial symbionts and their phylogenetic positioning still remain. To address this and other issues, herein I characterize and annotate a large subset of the genome of *Reichenowia parasitica*, an alphaproteobacterial endosymbiont of the freshwater leech *Placobdella parasitica*. Results suggest that *R. parasitica* possesses genes coding for proteins related to nitrogen fixation, iron/vitamin B translocation and plasmid survival, and that the bacterium interacts with its host in part by transmembrane signaling. The phylogenetic analyses support the nesting of *R. parasitica* within the plant-inhabiting Rhizobiaceae, as sister to a group containing *Agrobacterium* and *Rhizobium* species.

ACKNOWLEDGEMENTS

First and, without a shadow of a doubt, foremost, I thank my adviser, colleague and friend Mark Siddall. Throughout the dissertation process you have always provided your stupendous knowledge and skillset, proclaimed your trust in me, presented your unquestionable loyalty, and permitted me to make mistakes and learn from them. If someone would have told me four years ago that I would be discussing scientific and other philosophical topics at your dinner table in New York City, I would have laughed. Ironically, thinking back on the times we've spent discussing, arguing, singing, eating, drinking and airguitaring, during this all-too-brief period, also makes me laugh. There are so many things that you've taught me about (e.g.,) science and for that I will be forever grateful. While I will remember and continue to apply numerous of these, some are more conspicuously engrained in my brain, almost to a tantric extent. RTFM (Read The F*&#ing Manual) immediately springs to mind. I can never thank you enough for making me perform tasks that I did not know I could do, showing me the art of scientific writing, reminding me to be critical and imaginative in my own science and for making me feel at home in the most overwhelming of cities. All I can do is tell you this: when time comes, I will pay it forward!

To my dearest academic committee, Susan Perkins, Rob DeSalle, George Amato and Neil Sarkar – thank you for instilling calm in me and for guiding me and teaching me throughout this process. Seldom is a Ph.D. student this happy with her/his choice of committee! I also thank my BS and MS adviser and my dear friend Christer Erséus for all the great guidance, advice and friendship.

It is true, to some extent, that your academic family becomes as much a part of you as your real family. Like a real family, you never get to choose who your lab members are. Luckily, I shared the same bit of floor with two of my best friends and two of the best scientists that I have ever had the joy to know, my academic siblings Alejandro Oceguera-Figueroa and Anna Phillips.

Alejandro, your passion for all things invertebrate, your philosophical nature, your inquisitive mind and your incredible skillset makes for an astonishing concoction. Thank you for keeping me motivated and adventurous, and thank you for showing me the outside of the box on so many occasions. ¡Eso fue lo que trajo el barco!

Anna, thank you for your warmth, kindness and courage. No matter what field you dive in to, if you set your mind to something, it happens. Your patience and skills were pivotal for my understanding of leech biology, as well as my well-being and advancement during my dissertation. Although my heels may be stained with tar from time to time, I will always route for the team that you are on!

Mother, you have always been my biggest fan, routing for and encouraging me in whatever task I have undertaken. Well know this: I am also your biggest fan. Your scientific integrity is gargantuan, like your knowledge and your love of science. In fact, the only thing surpassing these is your love for your family. I am truly grateful for your confidence in me, your support, your love and your advice. I reflect you!

To my father, the kindest person in the world, I want to extend my deepest gratitude from the entire family. So much of what we have accomplished is because of you. You untiring support and hard work allows us to be free-thinkers, to develop as people and to feel safe and cared for. Thank you for always being there and for being so interested in everything I have ever done.

I also greatly thank Alex, my brother. Your involvement in my dissertation is larger than I suspect that you think. Thank you for paving the way, for the laughs, for support, care and love.

It would be wrong of me to take all credit for the accomplishment of finishing my dissertation. Notwithstanding all of the above-mentioned people, there are so many more that I need to thank. Everyone who has helped me get through this period with my sanity intact – Thank you! In particular, I thank Shae (with family), Bryan, Antonia and Zach for being equal parts pioneers and guineapigs, and for all the good times! I extend a special thanks to Liz Wright who beyond making the office look like the Top Chef kitchen has helped me with various analyses and who makes it a bit easier to come into the office every day. Para los latinos locos: Eliecer, Jairo, Mariano, Maite, Miguel, Edmundo y Ofelia - ¡Muchas gracias por todo! I wish that I had more space to thank everyone personally here but I don't. I thank Ed (the toad is the ugliest of all the amphibians), Snorri (the night is young), Alejandro (remember when I almost swallowed that leech), Isabelle, John, Pedro, Phil, Ansel, Dawn, Carly, Eugenia, Andre, Stephanie, Nichole and Yi for being such awesome schoolmates.

Finally, I thank my beautiful wife Charlotte, singlehandedly the most important person both throughout this process and during the last 9 years of my life. Thank you for making me strive towards being the best I can each day, for showing such courage in your choices and for your unquestionable love. I could not have done this without you! I admire and love you beyond words!

Nature does not proceed by leaps and bounds

- Carl von Linné

TABLE OF CONTENTS

FIGURES AND TABLES

Figures

Tables

CHAPTER I

BACKGROUND

The phylum Annelida (Greek, *annulatus*, “annulated” or “ringed”) comprises over 17,210 currently described and valid species of bristle worms, leeches, earthworms and their allies (Zhang, 2011). Colloquially, members of the phylum are known as “segmented worms” due to the compartmentalization of their body, in which several internal and external organs are repeated within each segment. This organization of the body is known as serial homology (Brusca and Brusca, 2002) and is characteristic of the entire phylum. Moreover, all annelids are bilaterally symmetrical and triploblastic (arising from teloblastic development), they possess a complete gut (excepting species of *Inanidrilus* and *Olavius*), a closed circulatory system, an advanced nervous system and excretory mechanisms in the form of protonephridia or metanephridia. Many marine annelids produce trochophora larvae, a characteristic shared by several other lophotrochozoan taxa (Brusca and Brusca, 2002). Whereas group-specific conservation of some morphological characteristics (autapomorphies) exists, especially in leeches, the body sizes of annelid worms vary tremendously; the full spectrum ranging from lengths of less than a millimeter to over 240 centimeters (e.g., Maser and Rice, 1962; Healy, 1975). Unique among macroscopic animal phyla, annelids occur on every continent, in freshwater, marine and terrestrial ecosystems (Pettibone, 1982; Purschke, 1999). Even more astonishing, representatives of the phylum can be found in extreme and extremely varied environments such as ice cores (Shain et al., 2001; Shain, 2009) and hydrothermal vents (Felbeck, 1981; Shain, 2009). Some taxa are also sustained by hypoxic environments, while others thrive in more temperate ecosystems. In short, annelids exist virtually everywhere, regardless of geography and environment.

Historically, Annelida has been divided into two orders: Sedentaria Lamarck, 1818 and Errantia Andouin & Milne Edwards, 1834 (Bartolomaeus et al., 2005). As the names suggest, the former included taxa with a more sedentary or sessile lifestyle (but awkwardly also includes leeches and other highly motile groups), whereas the latter was used for more vagile and errant taxa (Perrier, 1897). Errantia strictly includes members of the morphologically diverse class Polychaeta (bristle worms), whereas Sedentaria includes both polychaetes and members of the class Clitellata. Clitellata, in turn, is comprised of the subclasses Oligochaeta (earthworms and their close relatives) and Hirudinida *sensu lato* (leeches, Branchiobdellida [crayfish worms] and Acanthobdellida; Siddall et al., 2001; Erséus, 2005).

Over two thirds (Erséus, 2005) of the described annelid diversity occurs within Polychaeta (Greek *poly*, “several”, *chaites*, “hairs” or “bristles”). The taxonomy of this large group of organisms is still problematic but most investigators agree that the class includes approximately 20 orders and over 80 families (Fauchald, 1977; Fauchald and Rouse, 1997; Rouse and Fauchald, 1997). The main divisions of polychaetes, based on morphology, have been difficult to corroborate with molecular data. Scolecida (including Arenicolidae, Capitellidae, Cossuridae, Maldanidae, Opheliidae, Orbiniidae, Paraonidae, Scalibregmayidae and *Questa* [?]; Rouse and Pleijel, 2001) is defined by the presence of parapodia with similar rami and the presence of two or more pygidial cirri, whereas the Palpata (further subdivided into Canalipalpata and Aciculata, which jointly contain the remaining families) includes species possessing palps. Members of

Polychaeta demonstrate numerous impressive behavioral, morphological and evolutionary characteristics. For example, *Riftia pachyptila* Jones, 1981 grows to be 240 cm and lives exclusively on hydrothermal vents in association with endosymbiotic mutualistic bacteria (López-Garcia et al., 2002), and members of the family Syllidae display every mode of reproduction known to occur across Polychaeta (e.g., Franke, 1999; Aguado et al., 2007, 2011).

The class Clitellata comprises over 5,800 species of Annelida. The name refers to a glandular, epidermal structure known as the clitellum (girdle or saddle) that is involved in the formation of cocoons. This structure is common to all clitellates but only becomes conspicuous when the worm reaches sexual maturity. For example, in the common earthworm *Lumbricus terrestris* Linnaeus, 1758, this structure is paler (sometimes yellow) and much thicker than the rest of the body, making it remarkably distinct. Clitellates are primarily hermaphroditic and parthenogenesis is known to occur in some taxa (Erséus, 2005). As mentioned above, Clitellata is commonly divided into Hirudinida (Greek, *hirudo*, "to suck") and Oligochaeta (Greek, *oligos,* "poor/few", *chaites*, "hairs/bristles"), although contemporary studies suggest that Clitellata is a synonym of Oligochaeta (Siddall et al., 2001; Erséus, 2005; see below). Hirudinida embodies about 800 species and Oligochaeta includes about 5000 species, 3400 terrestrial and 1600 aquatic (about 600 marine forms) (Erséus, 2005). Hirudinida is commonly divided into two main groups: Arhynchobdellida consists of the jaw-bearing species (including the medicinal leeches) and Rhynchobdellida consists of the proboscis-bearing species. The taxonomy of Oligochaeta, on the other hand, is not as straightforward. Frequently, the class is divided into Megadrili (or Metagynophora;

sensu Jamieson, 1988), which includes earthworm-like taxa and Microdrili, containing the remaining, predominantly aquatic and more undersized taxa.

The annelid phylogenetic predicament

Notwithstanding recent efforts to resolve the evolutionary history and phylogenetic relationships of Annelida (e.g., Erséus, 2005; Struck et al., 2007, 2008, 2011; Rousset et al., 2007; Zrzavý et al. 2009), hypotheses concerning certain groups are unstable. The instability is suggested by the fact that the class Polychaeta is regularly rendered paraphyletic with respect to Clitellata, as well as the non-segmented “phyla” Echiura and Sipuncula (see also Struck et al., 2007; Dunn et al., 2008; Dordel et al., 2010), and several of the long-standing clades among polychaetous annelids are themselves paraphyletic (see also Fauchald and Rouse, 1997). Beyond this, while Clitellata is frequently recovered as monophyletic, Hirudinida commonly nests within oligochaetous clitellates similarly rendering the latter paraphyletic (see also Brinkhurst and Nemec, 1987; Erséus, 1987; Siddall et al., 2001). Equally disconcerting is the fact that the traditionally held order Sedentaria is seldom recovered as monophyletic (Day, 1967; Westheide et al., 1999) and Errantia has only been recovered as monophyletic in one study, in which only morphological characters were used (Rouse & Fauchald, 1997). Whereas Errantia is upheld by a number of synapomorphic morphological characters (Bartolomaeus et al., 2005), Sedentaria includes such morphologically disparate taxa that establishing homologies is often difficult, if not impossible. Until Fauchald’s (1977) treatment of the two orders, there had also been doubt as to whether or not these different modes of life accurately reflect common ancestry (e.g., Day, 1967). The taxonomy behind these groups was likely a matter of convenience as opposed to their being reflective of true evolutionary relationships and, as a result, Fauchald (1977) eliminated the Errantia / Sedentaria dichotomy, erecting 17 new, taxonomically equivalent groups; this was later expanded to include more orders (Rouse and Fauchald, 1997; Rouse and Pleijel, 2001; Bartolomaeus et al., 2005).

Until recently, the phylogenetic analyses of molecular characters that underlie some of the taxonomic changes mentioned above have been based on only a small number of different loci (e.g., McHugh, 1997, 2000; Bleidorn et al., 2003; Bely and Wray, 2004; Erséus and Källersjö, 2004; Rousset et al., 2007), which may have contributed to inconsistency in the monophyletic groups recovered. To address this and other issues, a large expressed sequence tag data set was recently compiled for Annelida and its constituent taxa, including ~48,000 aligned amino acid sites for 39 taxa (Struck et al., 2011). The phylogenetic analyses employing this data set, based on Bayesian inference and Maximum Likelihood, recovered both Errantia and Sedentaria as monophyletic, thus being at odds with the contemporary view on annelid systematics. In light of these findings, in the second chapter of this thesis, I recovered a more data-rich nucleotide representation of the amino acid data set and analyze this (and re-analyze the original amino acid data set) under a cladistic, as opposed to probabilistic, framework. Beyond considerations of whether or not a phylogenetic analysis of the nucleotides coding for the amino acids result in the same hypothesis, this data set also lends itself well to a timely discussion on orthology statements and a desire for consistency of results across methods. The nucleotide representation used here represents the most comprehensive molecular character set compiled for Annelida and its constituent taxa and sheds further light on the difficulties of finding reliable and suitable loci for phylogenetic inferences of a group as ancient as Annelida (Struck et al., 2007).

The evolution of bloodfeeding in leeches

To our knowledge, bloodfeeding has been enabled in leeches by virtue of two “key innovations”, the details of which set them apart from other annelids. First, in order to both keep blood flowing in and around the incision wound of the prey and to maintain the blood meal in a suitable state during the long periods of digestion, leeches have evolved pharmacological cocktails of anticoagulation factors (Salzet, 2001). These salivary-gland-secreted bioactive peptides have a storied history when it comes to their use in human medicine. The documented use of leeches for medicinal purposes dates back over two millennia and the utility of leeches in modern medicine is becoming even more authoritative (Whitaker et al., 2004; Phillips and Siddall, 2009; Min et al., 2010). The most conspicuous application of leeches and their anticoagulants is that for relief of venous congestion following flap and digit replantation surgery (Derganc and Zdravic, 1960; Dabb et al., 1992; Soucacos et al., 1994) and the US Food and Drug Administration recently approved them as medical devices (Rados, 2004). Indeed, the first successful attempt at human clinical dialysis treatment was only facilitated by the introduction of a fine lining of the leech anticoagulant hirudin to the dialysis flow-tubes (Haas, 1924). Leech-derived hirudin, the most potent natural direct thrombin-inhibitor known (Greinacher and Warkentin, 2008), has remained of considerable interest to the field of medicine, especially in cases of heparin-induced thrombocytopenia (Greinacher et al., 1999). Contemporary studies seem to agree on the notion that bloodfeeding is a plesiomorphic strategy in leeches (Siddall and Burreson, 1995, 1996; Trontelj et al., 1999; Min et al., 2010). To test this, in the third chapter of this thesis, I investigated the presence of leech antiplatelet protein in the genome of the glossiphoniid leech *Helobdella robusta* Shankland et al., 1992. This non-bloodfeeding leech is frequently found at the base of the leech phylogeny (Siddall et al., 2005; Light & Siddall, 1999), suggesting that any possession of anticoagulants by the leech is by virtue of their presence in the most recent common ancestor of leeches.

Despite the range of applications of leech anticoagulants, neither the full diversity of salivary components that inhibit coagulation, much less the evolutionary selection acting on them, has been thoroughly investigated. Because of this conspicuous gap in our knowledge of leech-derived anticoagulants, I constructed expressed sequence tag libraries from salivary glands of two species of medicinal leeches, *Hirudo verbana* and *Aliolimnatis fenestrata*, and identify anticoagulant-orthologues present in the data. Moreover, to predict the level and type of selection acting on the proteins, and to identify putative active regions in the anticoagulant molecules, I used four different statistical methods for predicting signatures of positive and negative evolutionary pressures. The results of this study are presented in chapter IV of this thesis.

The second “key innovation” allowing for the leeches’ restricted hematophagous diet is the acquisition of bacterial symbionts. In bloodfeeding taxa of the family Glossiphoniidae, the bacterial symbionts are housed in specialized structures known as mycetomes (stemming from the early misconception that the symbionts were fungi) or bacteriomes, and these structures are not encountered outside of the family. Indeed, even among glossiphoniid taxa that have given up bloodfeeding entirely (e.g., species of *Glossiphonia* and *Helobdella*), these specific structures are lacking altogether. The mycetomes are attached to the esophagus and have no other obvious function than housing bacterial symbionts. It has been hypothesized that the bacterial associates provide the leeches with essential nutrients, such as vitamins and enzymes, otherwise lacking in their restricted diet (Nogge, 1981; Perkins et al., 2005). The importance of the leech bacterial symbionts is further evidenced by their vertical transovarial transmission (Siddall et al., 2004). Although symbiotic associations between leeches and bacteria are well-documented (Kikutchi and Fukatsu, 2002; Siddall et al., 2004, 2011; Graf et al., 2006), several questions concerning the details of the symbioses remain unanswered. In particular, neither the function of the symbionts nor their genomic makeup had previously been scrutinized. In the fifth chapter of this dissertation, I characterize and annotate a large subset of the genome of a leech-endosymbiotic alphaproteobacterium, *Reichenowia parasitica* Siddall et al., 2004, in an attempt to investigate how the symbiont may affect the host and to assess the symbiont's phylogenetic position among a wide range of bacteria, with much greater genomic coverage than that of previous phylogenetic hypotheses (358 orthologous loci).

In conclusion, this thesis focuses both on trying to disentangle general annelid evolutionary relationships and on understanding the evolution of bloodfeeding in leeches as it pertains to salivary peptides and bacterial symbionts in a more conclusive manner than previous studies. Each chapter of this dissertation involves a genomic perspective on the questions asked and by employing bioinformatics tools in conjunction with powerful phylogenetic software, this thesis aimed to more thoroughly address questions concerning the genomic evolution of this charismatic group of organisms.

References

Aguado, M.T., Nygren, A., Siddall, M.E. 2007. Phylogeny of Syllidae (Polychaeta) based on combined molecular analysis of nuclear and mitochondrial genes. Cladistics 23: 552-564.

Aguado, M.T., San Martin, G., Siddall, M.E. 2011. Systematics and evolution of syllids (Annelida, Syllidae). Cladistics 27: 1-17.

Bartolomaeus, T., Purschke, G., Hausen, H. 2005. Polychaete phylogeny based on morphological data – a comparison of current attempts. Hydrobiologia 535/536, 341-356.

Bely, A.E., Wray, G.A. 2004. Molecular phylogeny of naidid worms (Annelida: Clitellata) based on cytochrome oxidase I. Mol. Phylogenet. Evol. 30, 50-63.

Bleidorn, C., Posiadlowski, L., Bartolomaeus, T. 2006. The complete mitochondrial genome of the orbiniid polychaete *Orbinia latreillii* (Annelida, Orbiniidae) – a novel gene order for Annelida and implications for annelid phylogeny. Gene 370: 96-103.

Bleidorn, C., Vogt, L., Bartolomaeus, T. 2003. New insights into polychaete phylogeny (Annelida) inferred from 18S rDNA sequences. Mol. Phylogenet. Evol. 29, 279-288.

Brinkhurst, R.O., Nemec, A.F.L. 1987. A comparison of phenetic and phylogenetic methods applied to the systematics of Oligochaeta. Hydrobiologia 155, 65-74.

Brusca, R.C., Brusca, G.J. 2002. Invertebrates 2nd ed. Sinauer Associates, Sunderland, MA, USA.

Dabb, R.W., Malone, J.M., Leveret, L.C. 1992. The use of medicinal leeches in the salvage of flaps with venous congestion. Ann. Plast. Surg. 29: 250-256.

Day, J.H. 1967. A monograph on the Polychaeta of Southern Africa. Vol. British Museum (Natural History) Publication 656 (Part I. Errantia, Part II. Sedentaria). British Museum (Natural History), London, UK.

Derganc, M., Zdravic, F. 1960. Venous congestion of flaps treated by application of leeches. Brit. J. Plast. Surg. 13: 187-192.

Dordel, J., Fisse, F., Purschke, G., Struck, T.H. 2010. Phylogenetic position of Sipuncula derived from multi-gene and phylogenomic data and its implication for the evolution of segmentation. J. Zool. Syst. Evol. Res. 48, 197-207.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745-749.

Erséus, C. 1987. Phylogenetic analysis of the aquatic Oligochaeta under the principle of parsimony. Hydrobiologia 155, 75-89.

Erséus, C. 2005. Phylogeny of oligochaetous Clitellata. Hydrobiologia 535/536, 357-372.

Erséus, C., Källersjö, M. 2004. 18S rDNA phylogeny of Clitellata (Annelida). Zool. Scr. 33, 187-196.

Fauchald, K. 1977. The polychaete worms: definitions and keys to the orders, families and genera. Natural History Museum of Los Angeles County, Science Series 28, Los Angeles, CA.

Fauchald, K., Rouse, G.W. 1997. Polychaete systematics: past and present. Zool. Scr. 26, 71-138.

Felbeck, H. 1981. Chemoautotrophic potential of the hydrothermal vent tube worm *Riftia pachyptila* Jones (Vestimintifera). Science 213: 336-338.

Franke, H-D. 1999. Reproduction of the Syllidae (Annelida: Polychaeta). Hydrobiologia 402: 39-55.

Graf, J., Kikuchi, Y., Rio, R.V.M. 2006. Leeches and their microbiota: naturally simple symbiosis models. Trends Microbiol. 14: 365-371.

Gregory, T.R., Hebert, P.D.N. 2002. Genome size estimates for some oligochaete annelids. Can. J. Zool. 80: 1485-1489.

Greinacher, A. Warkentin, T.E. 2008. The direct thrombin inhibitor hirudin. Thromb. Haemostasis 99: 819-829.

Greinacher, A., Völpel, H., Janssens, U., Hach-Wunderle, V., Kemkes-Matthes, B., Eichler, P., Mueller-Velten, H.G., Pötzsch, B. 1999. Recombinant hirudin (Lepirudin) provides safe and effective anticoagulation in patients with heparin-induced thrombocytopenia. Circulation 99: 73-80.

Haas, G. 1924. Versuche der Blutauswashung am Lebenden mit Hilfe der Dialyse. Wien. Klin. Wochenschr. 4: 13-14.

Healy, B. 2008. A description of five new species of Enchytraeidae (Oligochaeta) from Scotland. Zool. J. Linn. Soc. 56: 315-326.

Jamieson, B.G.M. 1988. On the phylogeny and higher classification of the Oligochaeta. Cladistics 4: 367-410.

Jennings, R.M., Halanych, K.M. 2005. Mitochondrial genomes of *Clymenella torquata* (Maldanidae) and *Riftia pachyptila* (Siboglinidae): evidence for conserved gene order in Annelida. Mol. Biol. Evol. 22: 210-222.

Kikutchi, Y., Fukatsu, T. 2002. Endosymbiotic bacteria in the esophageal organ of glosiphoniid leeches. Appl. Environ. Microbiol. 68: 4637-4641.

Light, J.E., Siddall, M.E. 1999. Phylogeny of the leech family Glossiphoniidae based on mitochondrial gene sequences and morphological data. J. Parasitol. 85:815-823.

López-Garcia, P., Gaill, F., Moreira, D. 2002. Wide bacterial diversity associated with tubes of the vent worm *Riftia pachyptila*. Env. Microbiol. 4: 204-215.

Maser, M.D., Rice, R.V. 1962. Biophysical and biochemical properties of earthworm-cuticle collagen. Biochim. Biophys. Acta 63: 255-265.

McHugh, D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. Proc. Natl. Acad. Sci. USA 94, 8006-8009.

McHugh, D. 2000. Molecular phylogeny of the Annelida. Can. J. Zool. 78, 1873-1884.

Min, G-S., Sarkar, I.N., Siddall, M.E. 2010. Salivary transcriptome of the North American medicinal leech, *Macrobdella decora*. J. Parasitol. 96: 1211-1221.

Mwinyi, A., Meyer, A., Bleidorn, C., Lieb, B., Bartolomaeus, T., Podsiadlowski, L. 2009. Mitochondrial genome sequence and gene order of *Sipunculus nudus* give additional support for an inclusion of Sipuncula into Annelida. BMC Genomics 10: 27 doi: 10.1186/1471-2164-10-27.

Nogge, G. 1981. Significance of symbionts for the maintenance of an optimal nutritional state for successful reproduction in hematophagous arthropods. Parasitol. 82: 101-104.

Perkins, S.L., Budinoff, R.B., Siddall, M.E. 2005. New Gammaproteobacteria associated with blood-feeding leeches and a broad phylogenetic analysis of leech endosymbionts. Appl. Environ. Microbiol. 71: 5219-5224.

Perrier, E. 1897. Traité de Zoologie, Fascicule IV. Vers. Molusques, Tuniciers, Masson et Cie, Paris, France.

Pettibone, M.H. 1982. Annelida. In: Parker, S.P. (Ed.), Synopsis and Classification of Living Organisms 2. McGraw Hill, New York, USA.

Phillips, A.J., Siddall, M.E. 2009. Poly-paraphyly of hirudinidae: many lineages of medicinal leeches. BMC Evol. Biol. 9: 246.

Purschke, G. 1999. Terrestrial polychaetes – models for the evolution of the Clitellata (Annelida)? Hydrobiologia 406: 87-99.

Rados, C. 2004. Beyond bloodletting: FDA gives leeches a medical makeover. FDA Consumer 38: 9.

Reis-Filho, J.S. 2009. Next-generation sequencing. Breast Cancer Res. 11: Suppl. 3:S12.

Rouse, G.W., Fauchald, K. 1997. Cladistics and polychaetes. Zool. Scr. 26, 139-204.

Rouse, G.W., Pleijel, F. 2001. Polychaetes. Oxford University Press, New York, NY.

Rousset, V., Pleijel, F., Rouse, G.W., Erséus, C., Siddall, M.E. 2007. A molecular phylogeny of annelids. Cladistics 23, 41-63.

Salzet, M. 2001. Anticoagulants and inhibitors of platelet aggregation derived from leeches. FEBS Lett. 429: 187-192.

Struck, T.H., Nesnidal, M.P., Purschke, G., Halanych, K.M. 2008. Detecting possibly saturated positions in 18S and 28S sequences and their influence on phylogenetic reconstruction of Annelida (Lophotrochozoa). Mol. Phylogenet. Evol. 48, 628-645.

Struck, T.H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G., Bleidorn, C. 2011. Phylogenomic analyses unravel annelid evolution. Nature 471, 95-98.

Struck, T.H., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., Halanych, K.M. 2007. Annelid phylogeny and the status of Sipuncula and Echiura. BMC Evol. Biol. 7, 57.

Shain, D.H. 2009. Annelids in Modern Biology. John Wileys and Sons, New York, USA.

Shain, D.H., Mason, T.A., Farrell, A.H., Michalewicz, L.A. 2001. Distribution and behavior of ice worms (*Mesenchytraeus solifugus*) on south-central Alaska. Can. J. Zool. 79: 1813-1821.

Siddall, M.E., Apakupakul, K., Burreson, E. M., Coates, K. A., Erséus, C., Gelder, S. R., Källersjö, M., Trapido-Rosenthal, H. 2001. Validating Livanow: molecular data agree that leeches, branchiobdellidans, and *Acanthobdella peledina* form a monophyletic group of oligochaetes. Mol. Phylogenet. Evol. 21, 346-351.

Siddall, M.E., Budinoff, R.B., Borda, E. 2005. Phylogenetic evaluation of systematics and biogeography of the leech family Glossiphoniidae. Invertebr. Syst. 19: 105-112.

Siddall, M.E., Perkins, S.L., Desser, S.S. 2004. Leech mycetome endosymbionts are a new lineage of alphaproteobacteria related to the Rhizobiaceae. Mol. Phylogenet. Evol. 30: 178-186.

Siddall, M.E., Min, G-S., Fontanella, F.M., Phillips, A.J., Watson, S.C. 2011. Bacterial symbiont and salivary peptide evolution in the context of leech phylogeny. Parasitol. 138: 1815-1827.

Soucacos, P.N., Beris, A.E., Malizos, K.N., Kabani, C.T., Pakos, S. 1994. The use of medicinal leeches, *Hirudo medicinalis*, to restore venous circulation in trauma and reconstructive microsurgery. 13: 251-258.

Westheide, W., McHugh, D., Purschke, G., Rouse, G. 1999. Systematization of the Annelida: different approaches. Hydrobiologia 402, 291-307.

Whitaker, I.S., Izadi, D., Oliver, D.W., Monteath, G.M., Butler, P.E. 2004. *Hirudo Medicinalis* and the plastic surgeon. Brit. J. Plast. Surg. 57: 348-353.

Zhang, Z-Q. 2011. Animal biodiversity: an introduction to higher-level classification and taxonomic richness. Zootaxa 3148, 7-12.

Zrzavý, J., Říha**,** P., Piálek, L., Janouškovec, J. 2009. Phylogeny of Annelida (Lophotrochozoa): total-evidence analysis of morphology and six genes. BMC Evol. Biol. 9, 189.

CHAPTER II

PHYLOGENOMICS OF ANNELIDA REVISITED: A CLADISTIC APPROACH USING GENOME-WIDE EST DATA MINING

(Adapted from Kvist, S., Siddall, M.E. Accepted with revisions. Phylogenomics of Annelida revisited: a cladistics approach using genome-wide EST data mining. Cladistics.)

Abstract

Here we present phylogenomic re-analyses of the most comprehensive molecular character set compiled for Annelida and its constituent taxa, including over 347 000 aligned nucleotide sites for 39 taxa across the phylum. The nucleotide data set was recovered using a preëxisting amino acid data set of almost 48 000 aligned sites as a backbone for tBLASTn searches against NCBI. In addition, orthology determinations of the loci in the original amino acid data set were scrutinized using an All vs. All Reciprocal Best Hit approach, employing BLASTp, and examining for statistical interdependency among the loci. This approach revealed considerable sequence redundancy among the loci in the original data set and a new data set was compiled, with the redundancy removed. Each of the newly compiled nucleotide data set, the original amino acid data set, and the new reduced amino acid data set were subjected to parsimony analyses and two forms of bootstrap resampling with two main objectives: (i) to examine the general topology, including support, resulting from the analyses of the new data sets and (ii) to assess the consistency of the branching patterns across optimality criteria by comparison to previous probabilistic approaches. The phylogenetic hypotheses resulting from analyses of the three data sets are not strongly supported reflecting the continued difficulty of finding numerous, reliable and suitable loci for a group as ancient as Annelida. Resulting parsimonious hypotheses disagree, in some respects, with the previous probabilistic approaches; Sedentaria and, in most cases, Errantia are not supported as monophyletic groups but Pleistoannelida is recovered as a (unsupported) monophyletic group in one of the three analyses.

Introduction

Notwithstanding recent efforts to resolve the evolutionary history and phylogenetic relationships of Annelida (e.g., Erséus, 2005; Struck et al., 2007, 2008, 2011; Rousset et al., 2007; Zrzavý et al. 2009), the debates relating to this are still contentious and the hypotheses concerning certain groups are unstable. While the class Clitellata (Hirudinida, oligochaetous clitellates, Branchiobdellida and Acanthobdellida) is frequently recovered as monophyletic, Hirudinida commonly nests within oligochaetous clitellates rendering the latter paraphyletic (see also Brinkhurst and Nemec, 1987; Erséus, 1987; Siddall et al., 2001). Moreover, the class Polychaeta, including the vast majority of the 17 210 annelid species currently recognized (Zhang, 2011), is regularly rendered paraphyletic with respect to Clitellata, as well as the non-segmented Echiura and Sipuncula (see also Struck et al., 2007; Dunn et al., 2008; Dordel et al., 2010), and several of the long-held clades among polychaetous annelids are themselves paraphyletic (see also Fauchald and Rouse, 1997). Equally disconcerting is the fact that the traditionally held order Sedentaria is seldom recovered as monophyletic (Day, 1967; Westheide et al., 1999) and its counterpart, Errantia, has only been recovered as monophyletic in one study, in which only morphological characters were used (Rouse & Fauchald, 1997). Sedentaria has historically referred to annelid worms with a more sessile or semi-sessile lifestyle (but including Clitellata) and with weakly developed, more or less absent parapodia, while Errantia includes worms with a more vagile lifestyle and more well-developed parapodia and chaetae (Perrier, 1897; Westheide et al., 1999; Bartolomaeus et al., 2005). Whereas Errantia is upheld by a number of synapomorphic morphological characters (Bartolomaeus et al., 2005), Sedentaria includes such morphologically disparate taxa that establishing homologies is often difficult, if not impossible. Until Fauchald's (1977) treatment of the two orders, there had also been doubt as to whether these different modes of life accurately reflected common ancestry (e.g., Day, 1967). The taxonomy behind these groups was likely a matter of convenience as opposed to their being reflective of true evolutionary relationships and, as a result, Fauchald (1977) eliminated the Errantia and Sedentaria dichotomy, whilst erecting 17 new, taxonomically equivalent groups, which were later expanded to include more orders (Rouse and Fauchald, 1997; Rouse and Pleijel, 2001; Bartolomaeus et al., 2005).

Until recently, the phylogenetic analyses of molecular characters that underlie some of the taxonomic changes mentioned above have been based on only a small number of different loci (e.g., McHugh, 1997, 2000; Bleidorn et al., 2003; Bely and Wray, 2004; Erséus and Källersjö, 2004; Rousset et al., 2007), which could be the reason for inconsistency in the monophyletic groups recovered. To address this and other issues, Struck et al. (2011) compiled a large data set of expressed sequence tags (EST's), including 231 loci for 39 taxa across Annelida. Their phylogenetic analyses, based on Bayesian inference and Maximum Likelihood (ML) estimations of almost 48

000 aligned amino acid sites, recovered both Errantia and Sedentaria as monophyletic groups (albeit somewhat taxonomically modified by Struck et al. [2011]). While their rediscovered monophyly may be reason enough to re-erect these old groupings, we believe that some re-considerations of the analyses are pertinent before restoring an older taxonomy of such a large group of organisms. Chiefly, we investigate whether or not a phylogenetic analysis of the nucleotides coding for the amino acids used by Struck et al. (2011) result in the same hypothesis. In addition, we examine whether or not locus interdependencies could lead to artificial clades or at least artificially inflated support values for clades. The large data set compiled by Struck et al. (2011) also lends itself well to a timely discussion on orthology statements and a desire for consistency of results across methods.

To this end, we here recovered orthologous nucleotide sequences for each locus and taxon using the amino acids as a source for targeted searches, and analyze this larger data set (as well as the amino acid data set independently) under a parsimony, as opposed to probabilistic, framework.

Material and methods

Data set reconstruction

The amino acid alignment compiled by Struck et al. (2011) (47 953 aligned sites for 39 taxa) was parsed and transferred to 231 separate files, each representing a single locus as defined in Supplementary Table 6 in Struck et al. (2011). The nucleotide data set then was compiled using the individual loci as queries for independent BLAST searches as described below.

Using BLAST client 3 (NCBI), the separate loci were compared remotely against both the EST database and the non-redundant (nr) sequence database at NCBI using a tBLASTn protocol (searching translated nucleotide databases using a protein query) and employing a cutoff e-value of $1E^{-5}$, retaining the best hit for each query; insofar as the best hit in a database should represent the sequence used by Struck et al. (2011). Data were obtained in this way for each of the following taxa: Polychaeta: *Alvinella pompejana* (Alvinellidae), *Arenicola marina* (Arenicolidae), *Cirratulus* sp. (Cirratulidae), *Eulalia clavigera* (Phyllodocidae), *Eurythoe complanata* (Amphinomidae), *Flabelligera affinis* (Flabelligeridae), *Glycera tridactyla* (Glyceridae), *Lanice conchilega* (Terebellidae), *Lumbrineris zonata* (Lumbrineridae), *Malacoceros fuliginosus* (Spionidae), *Onuphis iridescens* (Onuphidae), *Ophelia limacina* (Opheliidae), *Pectinaria koreni* (Pectinariidae), *Platynereis dumerilii* (Nereididae), *Pomatoceros lamarckii* (Serpulidae), *Ridgeia piscesae* (Siboglinidae), *Scoloplos armiger* (Orbiniidae), *Sthenelais boa* (Sigalionidae), *Typosyllis pigmentata* (Syllidae); Clitellata: *Eisenia andrei* (Lumbricidae), *Eisenia fetida* (Lumbricidae), *Haementeria depressa* (Glossiphoniidae), *Hirudo medicinalis* (Hirudinidae), *Lumbricus rubellus* (Lumbricidae), *Perionyx excavatus* (Megascolecidae), *Tubifex tubifex,*(Naididae); Bivalvia: *Crassostrea gigas* (Ostreidae); Myzostomida: *Myzostoma cirriferum* (Myzostomidae); and Sipuncula *Sipunculus nudus* (Sipunculidae). In cases where the best hit among the targets did not match the taxonomic identities of the queries, a new tBLASTn search was performed on the NCBI website using the Entrez–option to confine the search to the respective taxon.

Complementary to the foregoing, the amino acid sequences for each of the following taxa were compared against the annotated genomes of the respective taxon using a local tBLASTn search: Polychaeta: *Capitella teleta* (Capitellidae); Clitellata: *Helobdella robusta* (Glossiphoniidae); and Gastropoda: *Lottia gigantea* (Lottiidae). Again, the search employed a cutoff e-value of $1E^{-5}$ and the best hit was retrieved. The nucleotide sequences from the genomic hits were extracted and added to the data set acquired from the EST and nr databases.

In addition, tBLASTn searches ($1E^{-5}$ cutoff) were performed against assembled trace archive data for each of: Polychaeta: *Chaetopterus variopedatus* (Chaetopteridae); Myzostomida: *Myzostoma seymourcollegiorum* (Myzostomidae); Ectoprocta: *Bugula neritina* (Bugulidae); Nemertea: *Cerebratulus lacteus* (Cerebratulidae); Brachiopoda: *Terebratalia transversa* (Laqueidae); Sipuncula: *Themiste lageniformes* (Themistidae); and Echiura: *Urechis caupo* (Urechidae). To ensure maximum correspondence, the exact same trace archive assemblies as used by Struck et al. (2011) were used as targets. Nucleotide sequences from the best hit for each locus and taxon were extracted and added to the total data set now consisting of data from all four sources. All of these data, including the entire nucleotide data set as well as hit descriptions for each of the tBLASTn searches, are available from the first author upon demand.

Repeat masking

All of the data retrieved from the tBLASTn searches were imported to and indexed in a database created in FileMaker Pro ver. 5 (FileMaker). Polyadenosine tails were identified and, in cases where they were longer than four bases, each was truncated at the upstream polyadenylation signal (AATAA; Zaret and Sherman, 1982). In addition, homopolymer leading and trailing sequences were excised where they occurred in multiples of 2 or more (largely in *Platynereis dumerilii*). Other dinucleotide, trinucleotide and tetranucleotide repeat patterns also were identified with FileMaker Pro and excised using the RepBase repeats library for the nematode *Caenorhabditis elegans* (Rhabditidae) as implemented in the software EGassembler (Masoudi-Nejad et al., 2006) employing the “slow” option with a default cutoff score of 225.

Alignment and phylogenetic analyses

Sequences for each locus in the nucleotide data set were aligned independently using MAFFT (Katoh et al., 2005) on the European Bioinformatics Institute website employing a gap-opening cost of 3 and default settings for all other parameters. Each of the joined nucleotide data set and the amino acid data set from Struck et al. (2011) then were subjected to parsimony analyses using TNT (Goloboff et al., 2008). New Technology searches were conducted employing sectorial searching, with the tree fusing and ratcheting algorithms turned on. Trees were retrieved by a driven search using 100 initial addition sequences and requiring that the minimum length tree be found at least 5 times. All characters were equally weighted and non-additive, and gaps were treated as missing data. The results of the New Technology searches were subsequently resubmitted to TNT for TBR branch swapping. Support values for nodes also were estimated in TNT through both standard bootstrap resampling and partition (i.e., locus) bootstrapping (Siddall, 2009). Both bootstrap analyses employed 100 iterations, each subjected to five iterations of ratcheting and three rounds of tree fusing after an initial five rounds of Wagner tree building. The trees were rooted at *Bugula neritina* following Struck et al. (2011).

In addition, to assess the level at which the taxa grouped together based on mere presence or absence of data (and the support values related to this), rather than true phylogenetic signal, all positions in the amino acid data set were changed to an "A" such that the only information present in the data set consisted of patterns of presence or absence of data. The analysis then was re-run using the same parameters as mentioned above but now treating gaps as a fifth state and supported clades were crosschecked against clades present in the Bayesian tree found by Struck et al. (2011).

Examination of interdependent loci

An "all vs. all" reciprocal BLAST or bi-directional best hit search (Ge et al., 2005; Fang et al., 2010) using a BLASTp protocol (searching protein databases using a protein query) with a cutoff e-value of $1E^{-20}$ was performed on the amino acid loci to determine the level of similarity and coverage between them (hits that were less similar than $1E^{-20}$ were also retained separately to be used in the next step). Using the loci that showed hits at $1E^{-20}$ as a guide, all data (i.e., even hits less similar than $1E^{-20}$ but excluding self-hits) both for intra-locus (within a particular locus) and inter-locus (between loci) similarity values were compiled. In turn, averages, standard deviations and range intervals, as well as 95% confidence intervals were calculated both for intra-locus and inter-locus values. The putative overlap between the 95% confidence interval of intra-locus and inter-locus range values was assessed assuming that an overlap between 95% confidence intervals of e-value ranges indicated orthologous gene families or redundant use of sequences in more than one locus.

An additional phylogenetic analysis was performed on the amino acid data set following the removal of supposedly redundant loci, retaining the representative locus with the highest taxonomic coverage. The analysis used the same parameters as mentioned above with gaps treated as missing data.

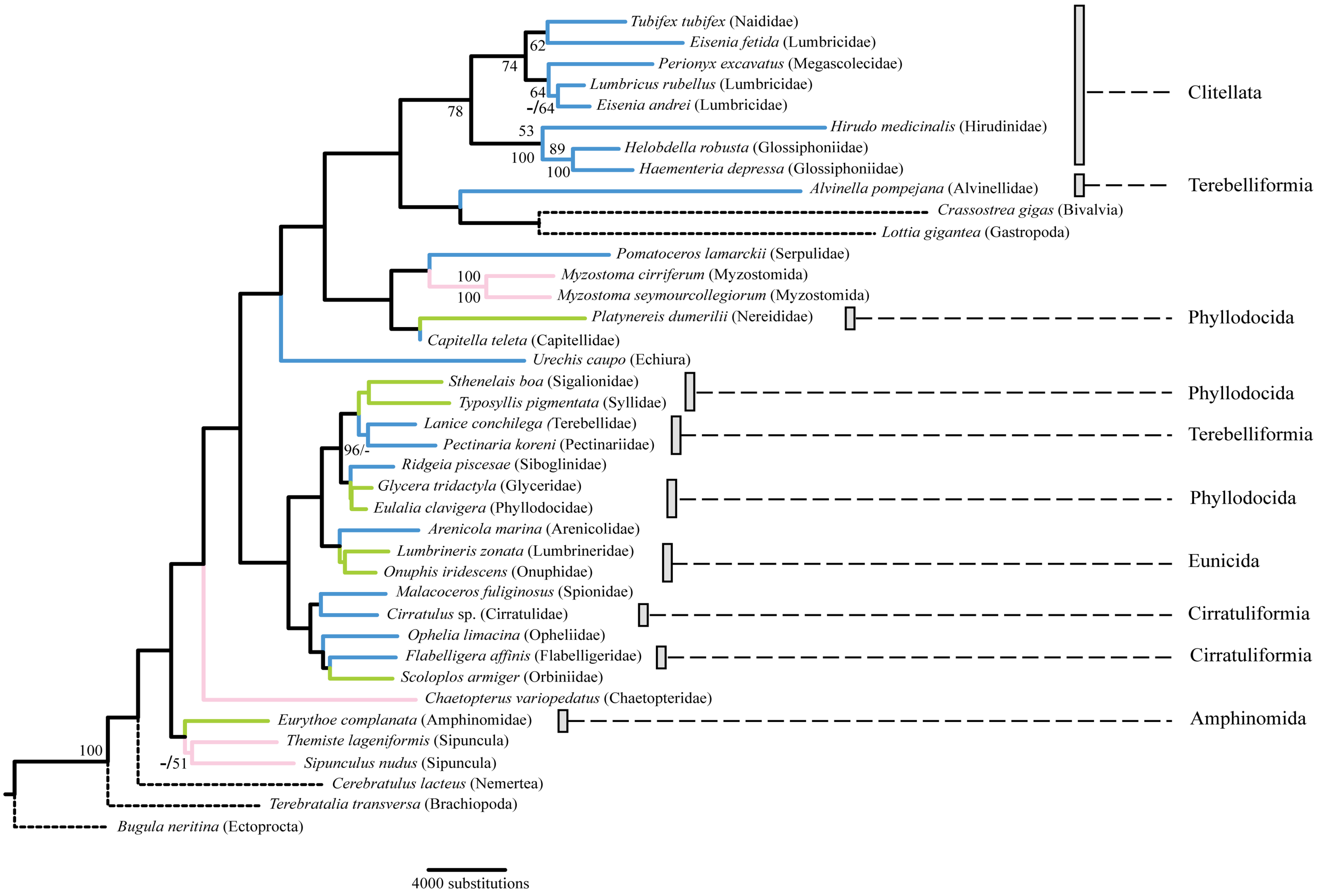
Results

Phylogenetic analyses

While a lenient e-value of 1E^{-5} was used as a cutoff for the tBLASTn searches, in almost all cases, resulting e-values were well below 1E^{-20}. With a rather high incidence, however, the best BLAST-hits for the amino acid sequences represented a different taxon than the query, requiring an additional tBLASTn search using the Entrez-option to confine the search to the same taxon. All of the individual BLAST searches for *Onuphis iridescens* matched *Lumbrineris zonata*, and vice versa, at a better e-value due to a confusion of these taxa (Torsten Struck pers. comm.) in the final data set used by Struck et al. (2011). That is, the NCBI submissions for these taxa were correct, whereas there is an error in the final data set used by Struck et al. (2011). Users should note that an NCBI search using EST data from the study by Struck et al. (2011) for *O. iridescens* will more than likely result in data for *L. zonata* and vice versa; the TreeBase submission for this data set now includes a disclaimer noting this confusion. In the present study, this was remedied by extracting the best hit for the opposite taxon in every case. As an aside, Supplementary Table 6 in Struck et al. (2011) states that data for *Myzostoma* *seymourcollegiorum*, *Platynereis dumerilii* and *Glycera tridactyla* are available in the NCBI EST database but, rather, data for *M. seymorcollegiorum* is found in NCBI trace archives and for the latter two in the NCBI nr database. Hits equal to or better than $1E^{-5}$ were not found for *Glycera tridactyla* at locus 163, 164, 171, 175, 187 and 189 or for *Sthenelais boa* at locus 228. Either no sequence (or an erroneous one) was deposited in NCBI for those taxa at those loci or the translation from nucleotides to amino acids by Struck et al. (2011) was not equivalent. For information on data coverage for each taxon, see Supplementary material in Struck et al. (2011).

The final nucleotide data set consisted of 347 298 aligned sites, 118 163 of which were parsimony informative. The phylogenetic analysis of this data set returned a single most parsimonious tree with 932 748 steps (Fig. 2.1). Support values were low across the entire topology of the tree, with the exception of support values relating to Clitellata, which were relatively high. Both of the mollusks (*Lottia gigantea* and *Crassostrea gigas*) nest within Annelida (standard bootstrap support; BS <50%, partition bootstrap support; PBS <50%), rendering the phylum non-monophyletic. Neither Errantia nor Sedentaria is monophyletic by virtue of taxa from each group nesting within the other. Clitellata is recovered as a monophyletic group (BS <50%, PBS 78%) as sister to an unsupported clade containing *Alvinella pompejana, Crassostrea gigas* and *Lottia gigantea.* Neither Canalipalpata (including Terebelliformia, Cirratuliformia, Siboglinidae, Serpulidae and Spionidae) nor Scolecida (including Capitellidae, Ophellidae and Arenicolidae) were recovered as monophyletic. The echiuran *Urechis caupo* nests within Annelida as sister to the clade containing *Platynereis*/*Capitella, Pomatoceros*/Myzostomida, *Alvinella/*Molluska and Clitellata.

Figure 2.1. Single most parsimonious tree recovered from the analysis of the nucleotide data set (L=932 748; CI=0.487; RI=0.236). Standard bootstrap values ≥50% are indicated above each node and partition bootstrap values ≥50% below each node. Note the low support for most nodes, which is discussed further in the text. Terminal branches are colored following taxonomic affiliations in Struck et al. (2011): blue=Sedentaria, green=Errantia, red=Annelida but not Errantia or Sedentaria. Dashed lines denote non-annelid taxa and grey bars denote additional taxonomic information. Branch lengths are drawn proportional to change.


Tubifex tubifex (Naididae)
Eisenia fetida (Lumbricidae)
Perionyx excavatus (Megascolecidae)
Lumbricus rubellus (Lumbricidae)
Eisenia andrei (Lumbricidae)
Hirudo medicinalis (Hirudinidae)
Helobdella robusta (Glossiphoniidae)
Haementeria depressa (Glossiphoniidae)
Alvinella pompejana (Alvinellidae)
Crassostrea gigas (Bivalvia)
Lottia gigantea (Gastropoda)
Pomatoceros lamarckii (Serpulidae)
Myzostoma cirriferum (Myzostomida)
Myzostoma seymourcollegiorum (Myzostomida)
Platynereis dumerilii (Nereididae)
Capitella teleta (Capitellidae)
Urechis caupo (Echiura)
Sthenelais boa (Sigalionidae)
Typosyllis pigmentata (Syllidae)
Lanice conchilega (Terebellidae)
Pectinaria koreni (Pectinariidae)
Ridgeia piscesae (Siboglinidae)
Glycera tridactyla (Glyceridae)
Eulalia clavigera (Phyllodocidae)
Arenicola marina (Arenicolidae)
Lumbrineris zonata (Lumbrineridae)
Onuphis iridescens (Onuphidae)
Malacoceros fuliginosus (Spionidae)
Cirratulus sp. (Cirratulidae)
Ophelia limacina (Opheliidae)
Flabelligera affinis (Flabelligeridae)
Scoloplos armiger (Orbiniidae)
Chaetopterus variopedatus (Chaetopteridae)
Eurythoe complanata (Amphinomidae)
Themiste lageniformis (Sipuncula)
Sipunculus nudus (Sipuncula)
Cerebratulus lacteus (Nemertea)
Terebratalia transversa (Brachiopoda)
Bugula neritina (Ectoprocta)
62
74
64
-/64
78
53
89
100
100
100
100
96/-
100
-/51
Clitellata
Terebelliformia
Phyllodocida
Phyllodocida
Terebelliformia
Phyllodocida
Eunicida
Cirratuliformia
Cirratuliformia
Amphinomida
4000 substitutions

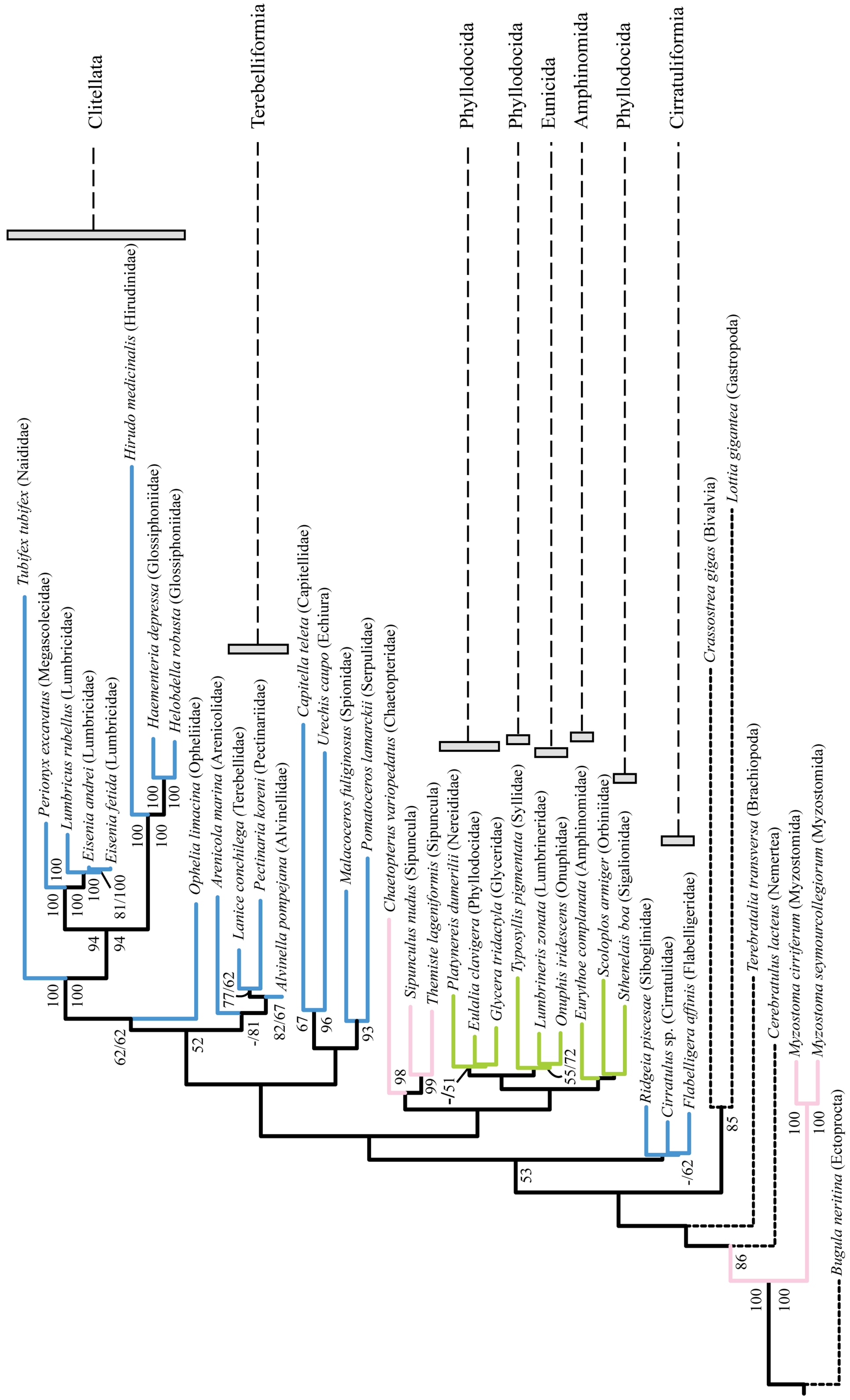
The original amino acid data set used by Struck et al. (2011) consisted of 47 953 aligned sites, 18 628 of which were parsimony informative. Analysis of that data set returned a single most parsimonious tree with 110 516 steps (Fig. 2.2). Again, support values were low across most of the topology, but show relatively high values for clitellate clades. In the tree, both species of *Myzostoma* (these were included in Annelida by Struck et al. [2011], but see their discussion) nest among the outgroup taxa (BS <50%, PBS 86%). Sedentaria is rendered paraphyletic (BS <50%, PBS <50%) with respect to both the monophyletic Errantia (BS <50%, PBS <50%) and another clade containing both the sipunculids (placing together with BS 98%, PBS 99%) and *Chaetopterus variopedatus* (BS <50%, PBS <50%). Clitellata is monophyletic (BS 100%, PBS 100%) but places as sister to Opheliidae. Within Sedentaria, neither Canalipalpata nor Scolecida are monophyletic. *Urechis caupo* nests well within Annelida; its position as sister to *Capitella teleta* is supported by BS of 67% and PBS of 96%.

The topology of the phylogenetic tree based on missing data only (not shown) is completely incongruent with the trees shown here, as well as the Bayesian tree recovered by Struck et al. (2011), affirming that the taxa are not grouping based on mere presence/absence of data.

Data independency

In the results of the BLASTp search, loci were deemed orthologous if the 95% confidence interval of the e-value range for hits between sequences from at least two

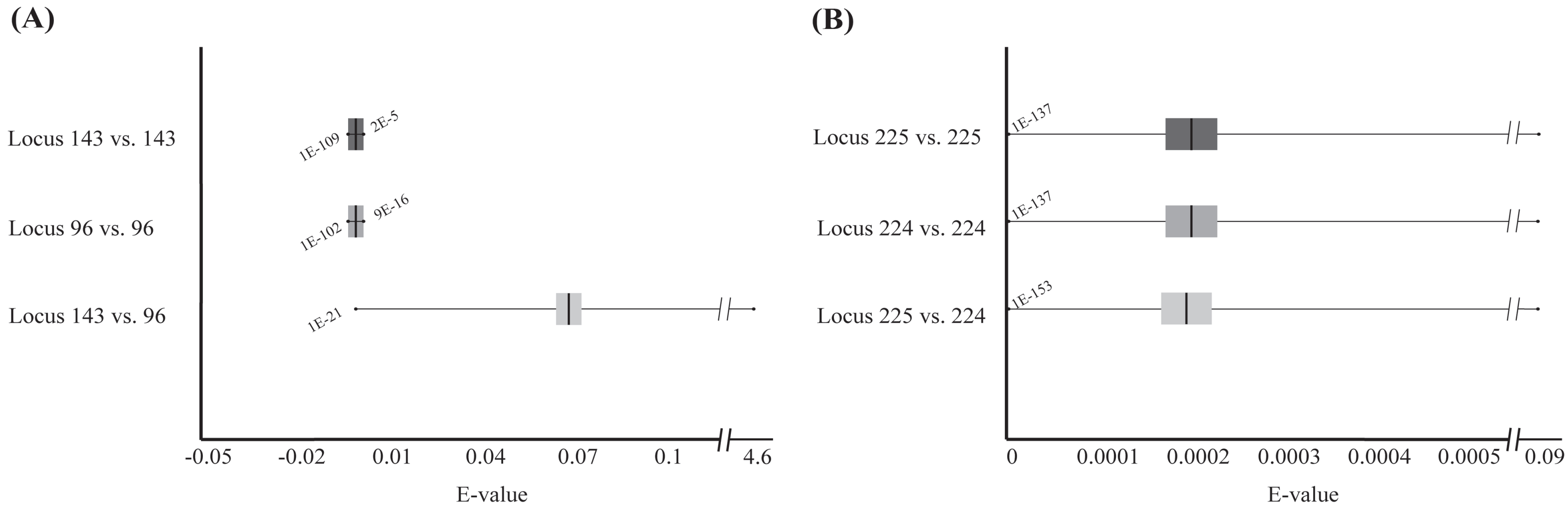
Figure 2.2. Single most parsimonious tree recovered from the analysis of the amino acid data set (L=110 516; CI=0.619; RI=0.347). Standard bootstrap values ≥50% are indicated above each node and partition bootstrap values ≥50% below each node. Note the low support for most nodes, which is discussed further in the text. Terminal branches are colored following taxonomic affiliations in Struck et al. (2011): blue=Sedentaria, green=Errantia, red=Annelida but not Errantia or Sedentaria. Dashed lines denote non-annelid taxa and grey bars denote additional taxonomic information. Branch lengths are drawn proportional to change.


Clitellata
Terebelliformia
Phyllodocida
Phyllodocida
Eunicida
Amphinomida
Phyllodocida
Cirratuliformia
Tubifex tubifex (Naididae)
Perionyx excavatus (Megascolecidae)
Lumbricus rubellus (Lumbricidae)
Eisenia andrei (Lumbricidae)
Eisenia fetida (Lumbricidae)
Hirudo medicinalis (Hirudinidae)
Haementeria depressa (Glossiphoniidae)
Helobdella robusta (Glossiphoniidae)
Ophelia limacina (Opheliidae)
Arenicola marina (Arenicolidae)
Lanice conchilega (Terebellidae)
Pectinaria koreni (Pectinariidae)
Alvinella pompejana (Alvinellidae)
Capitella telera (Capitellidae)
Urechis caupo (Echiura)
Malacoceros fuliginosus (Spionidae)
Pomatoceros lamarckii (Serpulidae)
Chaetopterus variopedatus (Chaetopteridae)
Sipunculus nudus (Sipuncula)
Themiste lageniformis (Sipuncula)
Platynereis dumerilii (Nereididae)
Eulalia clavigera (Phyllodocidae)
Glycera tridactyla (Glyceridae)
Typosyllis pigmentata (Syllidae)
Lumbrineris zonata (Lumbrineridae)
Onuphis iridescens (Onuphidae)
Eurythoe complanata (Amphinomidae)
Scoloplos armiger (Orbiniidae)
Sthenelais boa (Sigalionidae)
Ridgeia piscesae (Siboglinidae)
Cirratulus sp. (Cirratulidae)
Flabelligera affinis (Flabelligeridae)
Crassostrea gigas (Bivalvia)
Lottia gigantea (Gastropoda)
Terebratalia transversa (Brachiopoda)
Cerebratulus lacteus (Nemertea)
Myzostoma cirriferum (Myzostomida)
Myzostoma seymourcollegiorum (Myzostomida)
Bugula neritina (Ectoprocta)
100
100
100
100
100
81/100
94
94
100
100
100
100
100
62/62
52
77/62
-/81
82/67
67
96
93
98
99
-/51
55/72
-/62
53
85
86
100
100
100
100
700 substitutions

different loci overlapped with that of the range found within a locus. A total of 23 loci (9.96%) satisfied this criterion. Some of the redundancy was conjoined in triple-locus hits (i.e., if locus A=locus B and locus B=locus C then, transitively, locus A=locus C). A total of 13 loci (5.63%) were found to be redundant in that they are already represented by 10 other orthologues. Figure 2.3 illustrates this phenomenon using real similarity values from the data set generated by Struck et al. (2011). Specifically, non-redundant loci 96 and 143 each demonstrate intra-locus similarity values that do not significantly overlap with values obtained from comparisons between locus 96 and 143 (Fig. 2.3A). In contrast, the inter-locus similarities for loci 224 and 225 are indistinguishable from intra-locus values (Fig. 2.3B), illustrating their mutual redundancy. The following loci (with the single redundant locus retained for analysis appearing in parentheses) were removed from subsequent phylogenetic analyses: locus 14 (locus 7); locus 16 and locus 136 (locus 5); locus 32 and locus 109 (locus 35); locus 64 (locus 120); locus 125 and locus 137 (locus 134); locus 126 (locus 50); locus 129 (locus 74); locus 141 (locus 59); locus 212 (locus 53); locus 224 (locus 225).

After removing the 13 redundant loci (retaining 10 representatives), the data set comprised 44 262 aligned amino acid sites (92.30% of the total data set), 17 555 of which were parsimony informative. The analysis of these resulted in two equally parsimonious trees; the strict consensus of which is shown in Figure 2.4. The tree shows largely the same topology as the analysis of the original amino acid data set that included the redundant loci. Nevertheless, differences exist relative to the tree from the original data set as well as the tree found by Struck et al. (2011). Significantly, Errantia is not recovered as monophyletic: *Ridgeia piscesae* nests within the group. As well,

Figure 2.3. Two examples of the calculations of overlap between ranges of e-values within a 95% confidence interval; (A) shows an instance in which the 95% confidence interval of similarity values between two loci does not overlap with that of similarity values within each of those loci and (B) shows a case in which these values do overlap. Black horizontal lines bind the minimum and maximum e-values (i.e., ranges); shaded areas indicate the 95% confidence interval; and vertical lines within the shading indicate the average e-values for comparisons within and between the loci. Broken lines denote a compression of the actual ranges in order to fit the values within the ranges of the x-axes. Lower bound values of the e-value ranges, and in some cases also upper bound values, are printed as they are very close to 0. All e-values were recovered using a BLASTp protocol and the remaining comparisons for other loci are presented in the text.

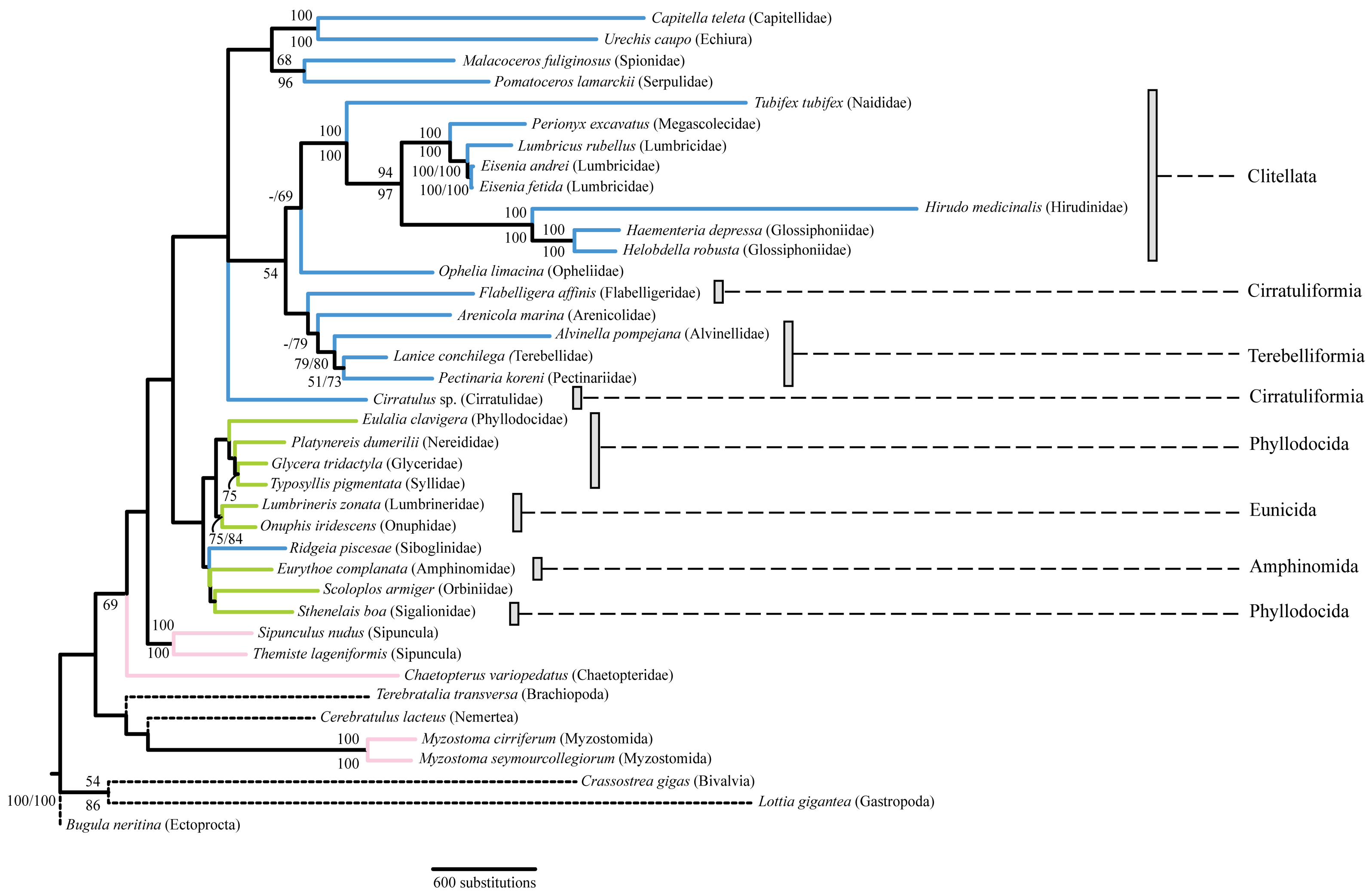

(A)
Locus 143 vs. 143
Locus 96 vs. 96
Locus 143 vs. 96
1E-109
2E-5
1E-102
9E-16
1E-21
-0.05
-0.02
0.01
0.04
0.07
0.1
4.6
E-value
(B)
Locus 225 vs. 225
Locus 224 vs. 224
Locus 225 vs. 224
1E-137
1E-137
1E-153
0
0.0001
0.0002
0.0003
0.0004
0.0005
0.09
E-value

Flabelligera affinis and *Cirratulus* sp. appear in different parts of the tree when redundant loci are excised (Fig. 2.4). However, Sedentaria+Errantia (named clade 1 by Struck et al. [2011], see Discussion below) is recovered as monophyletic. Neither the standard bootstrap values nor the partition bootstrap values associated with the tree vary markedly when excluding redundant loci but it can be noted that the standard bootstrap support for *Ophelia limacina* as sister to Clitellata is decreased to below 50%.

Discussion

Using the amino acid data set from Struck et al. (2011) as well as a newly compiled nucleotide representation of it, we show that parsimony analyses produce trees with topologies that are at odds with the Bayesian and ML trees recovered by that study. However, the resulting parsimony bootstrap values were found wanting for almost all clades, indicating that the topologies shown in Figs. 2.1 and 2.2 are no more reliable than topologies of previous studies (e.g., Rousset et al., 2007). That is, our study shows that even when using this large EST data set, cladistics analysis does not support the probabilistic hypotheses of phylogenetic relationships of Annelida, such as the monophyly of Sedentaria and, in most cases, Errantia. Regardless of one's preferred optimality criterion, these findings reflect the continued difficulty of finding numerous reliable and suitable loci for a group as ancient as Annelida (Struck et al., 2007). The fact that the Bayesian tree shown by Struck et al. (2011) receives dramatically higher support values than the parsimony trees shown here may not be surprising. Previous studies have shown that Bayesian posterior probabilities are often considerably higher when compared to parsimony or ML bootstrap values (Suzuki et al., 2002; Alfaro et al.,

Figure 2.4. Strict consensus of two most parsimonious trees recovered from the amino acid data set with 13 redundant loci removed (L=104 470 CI=0.615 RI=0.347). Standard bootstrap values ≥50% are indicated above each node and partition bootstrap values ≥50% below each node. Note the low support for most nodes, which is discussed further in the text. Terminal branches are colored following taxonomic affiliations in Struck et al. (2011): blue=Sedentaria, green=Errantia, red=Annelida but not Errantia or Sedentaria. Dashed lines denote non-annelid taxa and grey bars denote additional taxonomic information. Branch lengths are drawn proportional to change.


Capitella teleta (Capitellidae)
Urechis caupo (Echiura)
Malacoceros fuliginosus (Spionidae)
Pomatoceros lamarckii (Serpulidae)
Tubifex tubifex (Naididae)
Perionyx excavatus (Megascolecidae)
Lumbricus rubellus (Lumbricidae)
Eisenia andrei (Lumbricidae)
Eisenia fetida (Lumbricidae)
Hirudo medicinalis (Hirudinidae)
Haementeria depressa (Glossiphoniidae)
Helobdella robusta (Glossiphoniidae)
Ophelia limacina (Opheliidae)
Flabelligera affinis (Flabelligeridae)
Arenicola marina (Arenicolidae)
Alvinella pompejana (Alvinellidae)
Lanice conchilega (Terebellidae)
Pectinaria koreni (Pectinariidae)
Cirratulus sp. (Cirratulidae)
Eulalia clavigera (Phyllodocidae)
Platynereis dumerilii (Nereididae)
Glycera tridactyla (Glyceridae)
Typosyllis pigmentata (Syllidae)
Lumbrineris zonata (Lumbrineridae)
Onuphis iridescens (Onuphidae)
Ridgeia piscesae (Siboglinidae)
Eurythoe complanata (Amphinomidae)
Scoloplos armiger (Orbiniidae)
Sthenelais boa (Sigalionidae)
Sipunculus nudus (Sipuncula)
Themiste lageniformis (Sipuncula)
Chaetopterus variopedatus (Chaetopteridae)
Terebratalia transversa (Brachiopoda)
Cerebratulus lacteus (Nemertea)
Myzostoma cirriferum (Myzostomida)
Myzostoma seymourcollegiorum (Myzostomida)
Crassostrea gigas (Bivalvia)
Lottia gigantea (Gastropoda)
Bugula neritina (Ectoprocta)
Clitellata
Cirratuliformia
Terebelliformia
Cirratuliformia
Phyllodocida
Eunicida
Amphinomida
Phyllodocida
100
100
68
96
100
100
100
100
94
97
100/100
100/100
-/69
100
100
100
100
54
-/79
79/80
51/73
75
75/84
69
100
100
100
100
54
86
100/100
600 substitutions

2003; Cummings et al., 2003; Douady et al., 2003; Lewis et al., 2005), and might be treated with considerable caution. Note here, however, that Struck et al. (2011) base their conclusions only on nodes that also received high ML bootstrap support.

Stochastic models of evolution may be prone to difficulties in parameter estimation when the number of parameters is greatly increased (e.g., amino acid substitution matrices) or when a large number of characters are added in conjunction with the use of certain models of evolution (Felsenstein, 1982, 1983, 2004; Lartillot and Phillipe, 2004). Because (site-specific) mutations occur at the nucleotide level, and because nucleotides entail fewer parameters, we were curious to see how well a tree recovered from a nucleotide representation would compare to that of the amino acid data set used by Struck et al. (2011). Nucleotide data sets potentially hold an advantage over those of amino acids in that they consider potentially informative synonymous substitutions, a measure that is lost after translation into amino acids. If the rate of third-position nucleotide substitutions greatly exceeds "normal" evolutionary rates, this may in some cases lead to less resolved phylogeny reconstructions (Cunningham, 1997). However, in other cases, inclusion of such sites increases the resolution and support of the trees (e.g., Källersjö et al., 1999; Rydin et al., 2002; Kim et al., 2004). Bearing all of this in mind, our nucleotide data set produced a tree with a substantially different topology (Fig. 2.1) as compared to both the parsimony tree recovered from the original (Struck et al., 2011) amino acid data set (Fig. 2.2), as well as the parsimony tree recovered from the reduced amino acid data set used here (Fig. 2.4), much as it differs from the Bayesian tree shown by Struck et al. (2011). None of Annelida, Errantia and Sedentaria were monophyletic in the resulting nucleotide hypothesis, based on over 347

000 aligned sites. Nor was there much support for any but the already obviously well-supported groups (Rousset et al., 2007). It is also notable that the support values related to our analysis of nucleotide data agree, in most respects, to those estimated by the parsimony analysis performed by Zrzavý et al. (2009) on the basis of only six nucleotide loci, suggesting that a 38-fold increase in sequence information is no more or less helpful in resolving annelid relationships. However, this seems to be true only under the parsimony criterion as likelihood bootstrap values seem to greatly increase with an increased number of characters (see Struck et al., 2008; Struck, 2011). In comparing the parsimony analyses performed here with previous studies of annelid phylogenetics, it seems that using amino acids to reconstruct the trees is more propitious than using nucleotides; this was also noted by Dordel et al. (2010) for probabilistic analysis.

Redundancy of loci and orthology determination

All phylogenetic methods assume independence of characters (Farris, 1983); a prerequisite that is transitive to loci or other sets of characters. Whether or not automated orthology determinations are sensitive to this requirement is poorly evaluated in phylogenomics. Using a straightforward BLAST to establish orthology will not itself ensure locus independence, especially if coverage in a database is patchy (Koski and Golding, 2001). Our simple double-check was to perform an “all vs. all” BLAST search and contrast the ranges (rather than a single match) of the e-values *among* the suspected overlapping loci against the ranges of e-values *within* each putative locus. When applied to the amino acid data set used here, this simple method showed that 23 loci belonged to 10 distinct, redundant sets of loci. Exemplary of this, sequences for *Helobdella robusta,* *Lottia gigantea, Capitella teleta* and *Crassostrea gigas* at locus 59 and locus 141 all display perfect e-values (0), even for inter-locus comparisons. Non-independence of (e.g.,) loci inflates the number of ad-hoc hypotheses needed to describe the data on any tree (Farris, 1983). For example, if two taxa are supported as a group, any homoplasy supporting that group would be counted as many times as there were redundant non-independent sites (Farris, 1983; Farris and Kluge, 1985). Clearly, the use of non-independent data generates an artificial inflation of support values for clades that are themselves supported by the redundant loci involved. As the size of the data set increases, as it does when many orthologous loci are used more than once in a data set, the support values will also artificially increase (see de Queiroz et al., 1995).

A corollary problem, and one not fully explored here, concerns orthology statements that belie multiple loci in one. Several of the loci show abnormally dissimilar e-values for intra-locus comparisons, suggesting that the sequences have been forced to represent a single locus, regardless of sequence similarity or coverage. For example, within locus 48, *Cirratulus* sp. and *Platynereis dumerilii* return an e-value of 5E^{0} when compared against each other, much like *Sipunculus nudus* vs. *Arenicola marina* within locus 95. Of the 107 376 individual intra-locus BLAST hits, 2320 (2.16%) show e-values worse than 1E^{-5} and 1679 (1.56%) of them show e-values of 0.001 or less similar. This could of course be exacerbated by the often-fragmentary nature of EST data leading to very little overlap between taxa on any given locus. Regardless, if orthology statements for the loci involved in these hits were based on e-values alone, many would not accept that these loci should be considered orthologues. Indeed, BLAST-based orthology prediction is becoming more stringent (Kharchenko et al., 2006; Chen et al.,

2007) and there is some precedence for using an e-value cutoff of $1E^{-20}$ for unequivocal orthology determination (e.g., Putta et al., 2004; Pel et al., 2007). If these stringent e-values were to be applied to the current amino acid data set, fully 12 481 sequences (11.62%) would have to be removed from the partition to which they currently belong.

Errantia, Sedentaria and Pleistoannelida

Until the 1970's, Polychaeta was widely accepted to consist of Archiannelida, Errantia and Sedentaria (Bartolomaeus et al., 2005), although some authors had expressed doubt about the reliability of these taxa (e.g., Dales, 1962; Day, 1967). In the last quarter of the twentieth century, each of these large taxa was eliminated on the basis of both phylogenetic analysis and morphological examinations. While Hermans (1969) argued that archiannelid taxa are more closely related to each other than any other group based on morphological phylogenetics, Westheide (1985, 1987) based his arguments on ontogeny and a more comprehensive phylogeny, which suggested that Archiannelida is a paraphyletic assemblage. Errantia and Sedentaria also were eliminated during this period of rapid advancement of annelid systematics; Fauchald (1977) replaced them with 17 orders on the basis of morphological examination. Struck et al's. (2011) analyses imply a resurrection of Errantia and Sedentaria, and Struck (2011) also erects a new taxon, Pleistoannelida (defined as the group consisting of Errantia and Sedentaria, and exclusive of Sipuncula, Myzostomida and *Chaetopterus*).

In the most comprehensive analyses of Annelida prior to Struck et al. (2011) none of Errantia, Sedentaria or Pleistoannelida were reliably recovered as monophyletic groups (Zrzavý et al., 2009). With the most taxonomically broad analysis to date, Rousset et al. (2007), with 217 taxa for four loci, also failed to recover the monophyly of Errantia, Sedentaria or Pleistoannelida. Our reanalysis of nucleotides and of amino acids with redundant sequences removed continue to suggest a lack of compelling evidence for the monophyly of Errantia or Sedentaria, much like both the nucleotide data set and the original amino acid data set used by Struck et al. (2011) do not reliably find the monophyly of Pleistoannelida under a parsimony framework. Interestingly though, when removing the redundant loci from the amino acid data set, Pleistoannelida is recovered as monophyletic but without support. If the root of the tree resulting from our analysis of the original amino acid data set was applied at the node leading to *Chaetopterus*, however, our tree would result in monophyly of both Errantia and Sedentaria. Indeed, increased taxon sampling may be of even more importance than increased genetic coverage, and could potentially accurately resolve the relationships at the base of the tree of Annelida, which would have a large effect on the overall topology and groupings found (see Phillipe et al., 2011). Regardless, in conjunction with previous studies, our analyses suggest that the phylogenetic hypotheses of relationships both within Annelida, and between the phylum and its constituent taxa are still unstable; finding suitable data for resolving this is an important yet problematic issue.

References

Alfaro, M.E., Zoller, S., Lutzoni, F. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20, 255-266.

Bartolomaeus, T., Purschke, G., Hausen, H. 2005. Polychaete phylogeny based on morphological data – a comparison of current attempts. Hydrobiologia 535/536, 341-356.

Bely, A.E., Wray, G.A. 2004. Molecular phylogeny of naidid worms (Annelida: Clitellata) based on cytochrome oxidase I. Mol. Phylogenet. Evol. 30, 50-63.

Bleidorn, C., Vogt, L., Bartolomaeus, T. 2003. New insights into polychaete phylogeny (Annelida) inferred from 18S rDNA sequences. Mol. Phylogenet. Evol. 29, 279-288.

Brinkhurst, R.O., Nemec, A.F.L. 1987. A comparison of phenetic and phylogenetic methods applied to the systematics of Oligochaeta. Hydrobiologia 155, 65-74.

Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2, e383.

Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., Winka, K. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52, 477-487.

Cunningham, C.W. 1997. Can three incongruence tests predict when data should be combined?. Mol. Biol. Evol. 14, 733-740.

Dales, R.P. 1962. The polychaete stomatodeum and the interrelationships of the families of Polychaeta. Proc. Zool. Soc. London 139, 389-428.

Day, J.H. 1967. A monograph on the Polychaeta of Southern Africa. Vol. British Museum (Natural History) Publication 656 (Part I. Errantia, Part II. Sedentaria). British Museum (Natural History), London, UK.

de Queiroz, A., Donoghue, M.J., Kim, J. 1995. Separate versus combined analysis of phylogenetic evidence. Ann. Rev. Ecol. Syst. 26, 657-681.

Dordel, J., Fisse, F., Purschke, G., Struck, T.H. 2010. Phylogenetic position of Sipuncula derived from multi-gene and phylogenomic data and its implication for the evolution of segmentation. J. Zool. Syst. Evol. Res. 48, 197-207.

Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., Douzery, E.J.P. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20, 248-254.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745-749.

Erséus, C. 1987. Phylogenetic analysis of the aquatic Oligochaeta under the principle of parsimony. Hydrobiologia 155, 75-89.

Erséus, C. 2005. Phylogeny of oligochaetous Clitellata. Hydrobiologia 535/536, 357-372.

Erséus, C., Källersjö, M. 2004. 18S rDNA phylogeny of Clitellata (Annelida). Zool. Scr. 33, 187-196.

Fang, G., Bhardwaj, N., Robilotto, R., Gerstein, M.B. 2010. Getting started in gene orthology and functional analysis. PLoS Comput. Biol. 6:e1000703.

Farris, J.S. 1983. The logical basis of phylogenetic analysis. In: Platnick, N. I., Funk, V. A. (Eds.), Advances in cladistics 2. Columbia University Press, New York, NY.

Farris, J.S., Kluge, A.G. 1985. Parsimony, synapomorphy, and explanatory power: a reply to Duncan. Taxon 34, 130-135.

Fauchald, K. 1977. The polychaete worms: definitions and keys to the orders, families and genera. Natural History Museum of Los Angeles County, Science Series 28, Los Angeles, CA.

Fauchald, K., Rouse, G.W. 1997. Polychaete systematics: past and present. Zool. Scr. 26, 71-138.

Felsenstein, J. 1982. How can we infer geography and history from gene frequencies? J. Theor. Biol. 1, 9-20.

Felsenstein, J. 1983. Parsimony in systematics: biological and statistical issues. Ann. Rev. Ecol. Syst. 14, 313-333.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, MA.

Ge, F., Wang, L-S., Kim, J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3, 1709-1718.

Goloboff, P.A., Farris, J.S., Nixon, K.C. 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774-786.

Hermans, C.O. 1969. The systematic position of Archiannelida. Syst. Biol. 18, 85-102.

Källersjö, M., Albert, V.A., Farris, J.S. 1999. Homoplasy *increases* phylogenetic structure. Cladistics 15, 91-93.

Katoh, K., Kuma, K-I, Toh, H., Miyata, T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511-518.

Kharchenko, P., Chen, L., Freund, Y., Vitkup, D., Church, G. M. 2006. Identifying metabolic enzymes with multiple types of association evidence. BMC Evol. Biol. 7, 177.

Kim, S., Soltis, D.E., Soltis, P.S., Suh, Y. 2004. DNA sequences from Miocene fossils: an *ndhF* sequence of *Magnolia lathahensis* (Magnoliaceae) and an *rbcL* sequence of *Persea pseudocarolinensis* (Lauraceae). Am. J. Bot. 91, 615-620.

Koski, L.B., Golding, G.B. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52, 540-542.

Lartillot, N., Phillippe, H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095-1109.

Lewis, P.O., Holder, M.T., Holsinger, K.E. 2005. Polytomies and Bayesian inference. Syst. Biol. 54, 241-253.

Masoudi-Nejad, A., Tonomura, K., Kawashima, S., Moriya, Y., Suzuki, M., Itoh, M., Kanehisa, M., Endo, T., Goto, S. EGassembler: online bioinformatics service for large-scale processing, clustering and assembling EST's and genomic DNA fragments. Nucleic Acids Res. 34, W459-W462.

McHugh, D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. Proc. Natl. Acad. Sci. USA 94, 8006-8009.

McHugh, D. 2000. Molecular phylogeny of the Annelida. Can. J. Zool. 78, 1873-1884.

Pel, H.J., de Winde, J.H., Archer, D.B., Dyer, P.S., Hofmann, G. et al. 2007. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nat. Biotechnol. 25, 221-231.

Perrier, E. 1897. Traité de Zoologie, Fascicule IV. Vers. Molusques, Tuniciers, Masson et Cie, Paris.

Phillipe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9, e1000602.

Putta, S., Smith, J.J., Walker, J.A., Rondet, M., Weisrock, D.W., et al. Monaghan, J., Samuels, A.K., Kump, K., King, D.C., Maness, N.J., Habermann, B., Tanaka, E., Bryant, S.V., Gardiner, D.M., Parichy, D.M. Voss, S.R. 2004. From biomedicine to natural history research: EST resources for ambystomatid salamanders. BMC Genomics 5, 54.

Rouse, G.W., Fauchald, K. 1997. Cladistics and polychaetes. Zool. Scr. 26, 139-204.

Rouse, G.W., Pleijel, F. 2001. Polychaetes. Oxford University Press, New York, NY.

Rousset, V., Pleijel, F., Rouse, G.W., Erséus, C., Siddall, M.E. 2007. A molecular phylogeny of annelids. Cladistics 23, 41-63.

Rydin, C., Källersjö, M., Friis, E.M. 2002. Seed plant relationships and the systematic position of Gnetales based on nuclear and chloroplast DNA: conflicting data, rooting problems, and the monophyly of conifers. Int. J. Plant. Sci 163, 197-214.

Siddall, M.E. 2009. Unringing a bell: metazoan phylogenomics and the partition bootstrap. Cladistics 26, 444-452.

Siddall, M.E., Apakupakul, K., Burreson, E. M., Coates, K. A., Erséus, C., Gelder, S. R., Källersjö, M., Trapido-Rosenthal, H. 2001. Validating Livanow: molecular data agree that leeches, branchiobdellidans, and *Acanthobdella peledina* form a monophyletic group of oligochaetes. Mol. Phylogenet. Evol. 21, 346-351.

Struck, T.H. 2011. Direction of evolution within Annelida and the definition of Pleistoannelida. J. Zool. Syst. Evol. Res. 49, 340-345.

Struck, T.H., Nesnidal, M.P., Purschke, G., Halanych, K.M. 2008. Detecting possibly saturated positions in 18S and 28S sequences and their influence on phylogenetic reconstruction of Annelida (Lophotrochozoa). Mol. Phylogenet. Evol. 48, 628-645.

Struck, T.H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G., Bleidorn, C. 2011. Phylogenomic analyses unravel annelid evolution. Nature 471, 95-98.

Struck, T.H., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., Halanych, K.M. 2007. Annelid phylogeny and the status of Sipuncula and Echiura. BMC Evol. Biol. 7, 57.

Suzuki, Y., Glazko, G. V., Nei, M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA. 99, 16138-16143.

Westheide, W. 1985. The systematic position of Dinophilidae and the archiannelid problem. In: Conway Morris, S., George, J. D., Gibson, R., Platt, H. M. (Eds.), The origins and relationships of lower invertebrates. Oxford University Press, Oxford, UK.

Westheide, W. 1987. Progenesis as a principle in meiofauna evolution. J. Nat. Hist. 21, 843-854.

Westheide, W., McHugh, D., Purschke, G., Rouse, G. 1999. Systematization of the Annelida: different approaches. Hydrobiologia 402, 291-307.

Zaret, K.S., Sherman, F. DNA sequence required for efficient transcription termination in yeast. Cell 28, 563-573.

Zhang, Z-Q. 2011. Animal biodiversity: an introduction to higher-level classification and taxonomic richness. Zootaxa 3148, 7-12.

Zrzavý, J., Říha**,** P., Piálek, L., Janouškovec, J. 2009. Phylogeny of Annelida (Lophotrochozoa): total-evidence analysis of morphology and six genes. BMC Evol. Biol. 9, 189.

CHAPTER III

GENOME-WIDE SEARCH FOR LEECH ANTIPLATELET PROTEINS IN THE NON-BLOODFEEDING LEECH *HELOBDELLA ROBUSTA* (RHYNCHOBDELLIDA: GLOSSIPHONIIDAE) REVEALS EVIDENCE OF SECRETED ANTICOAGULANTS.

(Adapted from: Kvist, S., Sarkar, I.N., Siddall, M.E. 2011. Genome-wide search for leech antiplatelet proteins in the non-bloodfeeding leech *Helobdella robusta* (Rhyncobdellida: Glossiphoniidae) reveals evidence of secreted anticoagulants. Invertebrate Biology 130: 344-350.)

Abstract

The available genome of the non-bloodfeeding glossiphoniid leech *Helobdella robusta* was screened for leech antiplatelet protein (LAPP), an anticoagulant that specifically inhibits collagen-stimulated platelet aggregation. Previously identified LAPP sequences from *Haementeria officinalis* were used as queries against the predicted genes in the genome, employing a variety of BLAST protocols. Matches were reciprocally BLASTed against GenBank databases as a cross-validation of the predicted annotations of the genes. A total of eight loci, positioned as a tandem array, were recovered with significantly low e-values and showed high sequence similarity (32.49% average sequence similarity of shared amino acid positions) compared to the known anticoagulants. Moreover, six of these possess a predicted signal peptide towards the N-terminus, indicating their secretion by the leech. All eight loci, together with known LAPP sequences from *Haementeria officinalis*, as well as several sequences from publicly available expressed sequence tag libraries of *Haementeria depressa* and

Helobdella robusta were collectively aligned and subjected to phylogenetic analysis. The resulting tree showed a monophyletic clade consisting of the *Helobdella robusta* loci, placing as sister to the *Haementeria*-derived sequences. To corroborate the evolution of the anticoagulants with the evolution of leeches more generally, the topology of the LAPP-tree was compared to that of a previously published leech phylogeny, and these show compatible topologies concerning the included genera. These results corroborate contemporary phylogenetic work, which suggests that this non-bloodfeeding leech has a hematophagous ancestry.

Introduction

Leeches (Hirudinida) have evolved suites of salivary proteins (anticoagulants) that are injected into the feeding site to facilitate bloodfeeding. They are used by leeches both to prevent blood from coagulating around the incision wound of the prey, as well as to maintain the blood in a suitable state during the long periods of digestion by the leech (Salzet 2001).

Leech antiplatelet protein (LAPP) was first isolated from the glossiphoniid leech *Haementeria officinalis*, and characterized as an inhibitor of collagen-stimulated platelet aggregation (Connolly et al. 1992). A normal thrombus formation, following injury to vascular walls, is mediated by von Willebrand-factor (vWf) via its conformational change of proaggregatory collagen (Ruggeri 1997) and an irreversible binding to the surface glycoprotein complex GP Ib/IX/V (Obert et al. 1999; Watson et al. 2000). This leads to activation of the platelet and subsequent secretion of its granular contents, which form aggregates stimulating both thrombosis and the further activation of platelets (Connolly et al. 1992; Carter et al. 1998). Like saratin, which is a LAPP homologue first isolated from the European medicinal leech *Hirudo medicinalis* (Cruz et al. 2001), LAPP residues bind to subendothelial collagen thus inhibiting the vWf-mediated activation of the platelets. Already, LAPP and saratin have been isolated from a variety of leech species from different taxonomic families (Min et al. 2010; Barnes et al. 2001; Connolly et al. 1992) including *Haementeria depressa* and *H. officinalis* (Glossiphoniidae)*, Hirudo medicinalis* (Hirudinidae) and *Macrobdella decora* (Macrobdellidae).

Helobdella robusta is a non-bloodfeeding freshwater leech in the family Glossiphoniidae. As opposed to blood, the species feeds primarily on the haemolymphal fluids of freshwater snails, a strategy known as liquidosomatophagy. However, if the leech has had a recent hematophagous past, remnants of that ancestry may still be encoded in its genome. Owing largely to the transparent nature of the body of the organism, but also to the fact that its egg-cases (cocoons) are rather large and easily maintained in a laboratory setting, *H. robusta* has rapidly become central to evolutionary developmental studies of annelids (e.g. Shain 2009). In 2007, the DOE Joint Genome Institute released results from a full-genome sequencing effort of *H. robusta*. This genome enables investigations of the presence of anticoagulants also in a non-bloodfeeding leech. Here, we focus on investigating the presence of LAPP in *Helobdella robusta*, the putative secretion of the protein and on corroborating the evolution of the anticoagulant in the context of the evolution of the leech species that possess it.

Methods

Characterization of putative salivary peptides

The full genome of *Helobdella robusta* is available on the Joint Genome Institute (JGI) portal website (http://genome.jgi-psf.org/Helro1/Helro1.home.html). It consists of 2,354,463 reads in 1993 scaffolds for a total of 235.4 Mbp.

Amino acid sequences of previously characterized LAPP and saratin from bloodfeeding leeches were employed as queries to matches against the *H. robusta* genome. Queries were conducted using the tBLASTn and BLASTp algorithms on the JGI portal website with an e-value cut-off of $1E^{-2}$, without filtering low complexity regions, and using a gapped alignment with a BLOSUM62 scoring matrix. Complementary to this, word-searches were performed for both anticoagulants among the gene annotations already accomplished for the *H. robusta* genome. All matching regions were retained as nucleotide sequence records and localized on the JGI genome browser so as to correlate with gene predictions at each locus. Where more than one gene prediction model identified multiple identical loci, only one was retained to avoid redundancy. For all anticoagulant loci recovered in the *H. robusta* genome, both the nucleotide sequences (with introns removed) and their translated amino acid sequences were individually compared against the non-redundant (nr) GenBank nucleotide and protein sequence databases (using tBLASTx and BLASTp, respectively) as a definitive cross-validation of the predicted annotation (reciprocal BLAST or "bi-directional best hit" approach; Fang et al. 2010). To localize expressed putative orthologues, candidate anticoagulants from the *H. robusta* genome each were compared using the tBLASTx and BLASTn algorithms also to available expressed sequence tag (EST) libraries both in GenBank and against a stand-alone *Hirudo medicinalis* EST library at http://genomes.ucsd.edu/leechmaster/database. Moreover, to recover potential phylogenetic outgroups, LAPP from *Haementeria officinalis* and saratin from *Hirudo medicinalis* were queried against the entire GenBank nr nucleotide and protein databases, and the GenBank EST database, as well as the entire genome of the polychaete *Capitella teleta* on the JGI website, using BLASTn and BLASTx algorithms.

Signal peptides at the N-terminus were predicted using the SignalP 3.0 (Bendtsen et al. 2004) server at the Center for Biological Sequence Analysis website (http://www.cbs.dtu.dk/services/SignalP/) employing both neural networks and hidden Markov models for prediction. To identify and characterize conservation levels within the full genome-derived sequences, these were aligned with known anticoagulants using RevTrans 1.4 (Wernersson & Pedersen 2003) in accordance with their inferred amino acid states. The alignment used Clustal W 1.83 (Thompson et al. 1994) as implemented in RevTrans and used the Standard Genetic Code translation table. The amino acid alignment was visualized using Jalview ver. 2 (Waterhouse et al. 2009) where percent similarity was calculated by hand.

Phylogenetic analysis

Alignment of nucleotide sequences used for the phylogenetic analysis was also accomplished with RevTrans 1.4 in accordance with their inferred amino acid states. Aligned sequences were subjected to phylogenetic analysis under the parsimony criterion using TNT (Goloboff et al. 2008). A heuristic search was performed using the traditional search option with 100 random addition sequences and employing TBR branch swapping. Support values for the nodes were retrieved through standard bootstrap re-sampling with 1000 iterations each consisting of 10 random addition sequence replicates and with the same settings as above. For branch length comparisons, the matrix and resulting TNT trees were imported into PAUP* ver. 2.0b10 (Swofford 2002) and branch lengths were calculated. The tree was rooted using saratin from *H. medicinalis*.

Results

LAPP loci

A total of eight loci matching LAPP from *Haementeria officinalis* were found in the *Helobdella robusta* genome (Table 3.1). When reciprocally BLASTed against GenBank protein database , all of these loci matched leech antiplatelet proteins better than anything else; six of these were matched at e-values <1E^{-5} whereas two of them only hit with marginal e-values (BLASTp scores of 0.004 and 0.005). All eight of these putative LAPP loci co-localized in a tandem array in the *H. robusta* genome. Significant matches for the eight full-genome derived loci also were found in EST libraries that are available for leeches (Table 3.1). No matches were found in the *H. robusta* genome for saratin and no putative LAPP orthologues were found in the *Hirudo medicinalis* EST-library.

Table 3.1.Top matches in the GenBank nr. nucleotide and protein databases, as well as EST libraries using *H. robusta* loci as queries. The *H. robusta* loci (left column) were identified through a tblastn search using a known anticoagulant sequence as query (GenBank Accn. M81489).

***H. robusta* locus**	**tblastx GenBank nr.**	**blastp GenBank nr.**	**tblastx *Helobdella robusta* EST library (GenBank)**	**tblastx *Haementeria depressa* EST library (GenBank)**
e_gw1.2.270.1	M81489 LAPP (1E^{-12})	Q01747 antiplatelet protein (1E^{-13})	EY344275 (2E^{-64})	CN807637 LAPP (2E^{-16})
fgenesh4_pg.C_scaffold_2000933	-	Q01747 antiplatelet protein (5E^{-3})	EY344275 (2E^{-64})	CN807637 LAPP (2E^{-7})
e_gw1.2.291.1	-	1I8N_A LAPP (2E^{-7})	EY349090 (2E^{-56})	CN807637 LAPP (1E^{-5})
fgenesh4_pg.C_scaffold_2000928	-	Q01747 antiplatelet protein (2E^{-5})	EY361718 (4E^{-61})	CN807637 LAPP (5E^{-6})
estExt_Genewise1.C_21054	-	Q01747 antiplatelet protein (2E^{-7})	EY349090 (1E^{-75})	CN807637 LAPP (8E^{-5})
fgenesh4_pg.C_scaffold_2000934	-	Q01747 antiplatelet protein (3E^{-5})	EY341220 (3E^{-59})	CN807637 LAPP (3E^{-5})
fgenesh4_pg.C_scaffold_2000930	M81489 LAPP (4E^{-3})	Q01747 antiplatelet protein (5E^{-7})	EY392612 (5E^{-56})	CN807637 LAPP (2E^{-6})
fgenesh4_pg.C_scaffold_2000932	-	Q01747 antiplatelet protein (4E^{-3})	EY378603 (6E^{-59})	CN807637 LAPP (4E^{-5})

Protein conservation and secretion

The amino acid alignment included nine taxa; all eight of the full genome-derived loci and LAPP from *H. officinalis.* The average similarity of shared amino acid positions was 32.49%. The six disulphide-bond-forming cysteines, indicative of antiplatelet proteins (Min et al. 2010), showed full conservation across the eight full genome-derived sequences. In addition, Signal P predicted the presence of signal peptide regions in all but two sequences from the *H. robusta* genome (positions 10-27; Fig. 3.1).

Phylogeny

The BLASTx and BLASTn analyses using LAPP from *Haementeria officinalis* and saratin from *Hirudo medicinalis* against the entire GenBank nr database returned no hits, suggesting that no orthologous proteins are present outside of the leech taxa used in the present study. Thus, no outgroups could be used for the phylogenetic analysis performed here. The LAPP data set included 23 nucleotide sequences; eight loci derived from the full genome, 10 *H. robusta* EST sequences, three LAPP-loci from *Haementeria officinalis* and *Haementeria depressa* EST libraries*,* and two saratin-loci from *Hirudo medicinalis* and *Macrobdella decora*. A total of 549 aligned sites were analyzed, 279 of which were parsimony informative. The heuristic search returned a single most parsimonious tree with 1068 steps (Fig. 3.2a). The tree revealed a cluster grouping LAPP from *Haementeria officinalis* (the annotated sequence M81489) with two EST loci from *Haementeria depressa* (CN807637, CN807641). Sister to that group is a monophyletic cluster of all *H. robusta* loci with full-genome derived and EST derived

Fig. 3.1. Alignment of inferred amino acid sequences for *H. robusta* LAPP loci together with the known anticoagulant from *Haementeria officinalis*. Red boxes indicate predicted signal peptides, black boxes indicate fully conserved cysteines and green underlining denotes the taxon with a known sequence of the anticoagulant. Shading intensity corresponds to BLOSUM62 conservation.

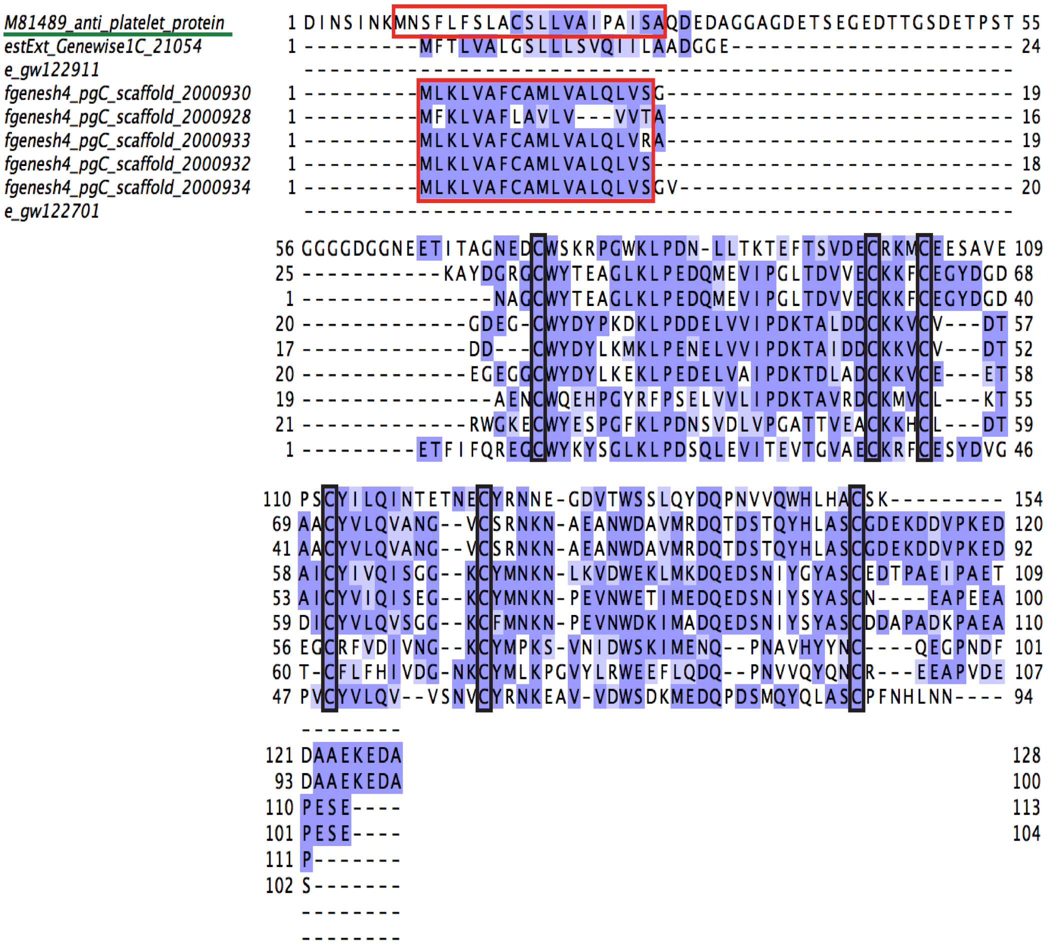
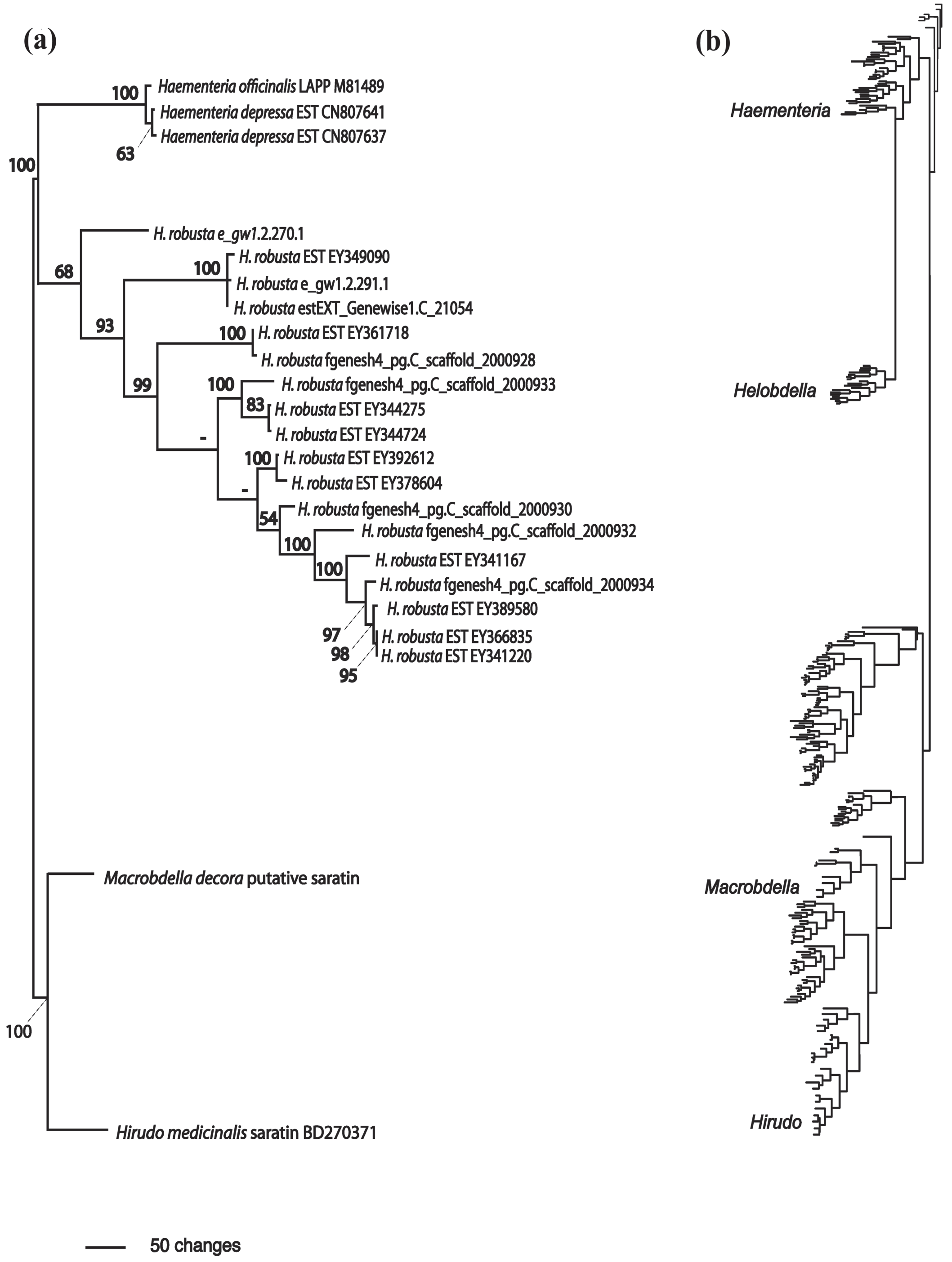

M81489_anti_platelet_protein
estExt_Genewise1C_21054
e_gw122911
fgenesh4_pgC_scaffold_2000930
fgenesh4_pgC_scaffold_2000928
fgenesh4_pgC_scaffold_2000933
fgenesh4_pgC_scaffold_2000932
fgenesh4_pgC_scaffold_2000934
e_gw122701
1 DINSINKMNSFLFSLACSLLVAIPAISAQDEDAGGAGDETSEGEDTTGSDETPST 55
1 ---------MFTLVALGSLLLSVQIILAADGGE---------------------- 24
----
1 ---------MLKLVAFCAMLVALQLVSG--------------------------- 19
1 ---------MFKLVAFLAVLV---VVTA--------------------------- 16
1 ---------MLKLVAFCAMLVALQLVRA--------------------------- 19
1 ---------MLKLVAFCAMLVALQLVS---------------------------- 18
1 ---------MLKLVAFCAMLVALQLVSGV-------------------------- 20
----
56 GGGGDGGNEETITAGNEDCWSKRPGWKLPDN-LLTKTEFTSVDECRKMCEESAVE 109
25 -----------KAYDGRGCWYTEAGLKLPEDQMEVIPGLTDVVECKKFCEGYDGD 68
1 ---------------NAGCWYTEAGLKLPEDQMEVIPGLTDVVECKKFCEGYDGD 40
20 -------------GDEG-CWYDYPKDKLPDDELVVIPDKTALDDCKKVCV---DT 57
17 -------------DD---CWYDYLKMKLPENELVVIPDKTAIDDCKKVCV---DT 52
20 -------------EGEGGCWYDYLKEKLPEDELVAIPDKTDLADCKKVCE---ET 58
19 ---------------AENCWQEHPGYRFPSELVVLIPDKTAVRDCKMVCL---KT 55
21 -------------RWGKECWYESPGFKLPDNSVDLVPGATTVEACKKHCL---DT 59
1 ---------ETFIFQREGCWYKYSGLKLPDSQLEVITEVTGVAECKRFCESYDVG 46
110 PSCYILQINTETNECYRNNE-GDVTWSSLQYDQPNVVQWHLHACSK--------- 154
69 AACYVLQVANG--VCSRNKN-AEANWDAVMRDQTDSTQYHLASCGDEKDDVPKED 120
41 AACYVLQVANG--VCSRNKN-AEANWDAVMRDQTDSTQYHLASCGDEKDDVPKED 92
58 AICYIVQISGG--KCYMNKN-LKVDWEKLMKDQEDSNIYGYASCEDTPAEIPAET 109
53 AICYVIQISEG--KCYMNKN-PEVNWETIMEDQEDSNIYSYASCN----EAPEEA 100
59 DICYVLQVSGG--KCFMNKN-PEVNWDKIMADQEDSNIYSYASCDDAPADKPAEA 110
56 EGCRFVDIVNG--KCYMPKS-VNIDWSKIMENQ--PNAVHYYNC----QEGPNDF 101
60 T-CFLFHIVDG-NKCYMLKPGVYLRWEEFLQDQ--PNVVQYQNCR---EEAPVDE 107
47 PVCYVLQV--VSNVCYRNKEAV-VDWSDKMEDQPDSMQYQLASCPFNHLNN---- 94
----
121 DAAEKEDA 128
93 DAAEKEDA 100
110 PESE---- 113
101 PESE---- 104
111 P-------
102 S-------

Fig. 3.2. Phylogenetic hypotheses of LAPP. a) Single most parsimonious tree recovered from the heuristic search using the LAPP data set (L=1068 steps; CI= 0.658; RI=0.784). Bootstrap values (>50%) are shown above each node. b) Phylogenetic hypothesis of leeches modified from Min et al. (2010) with each of *Haementeria, Helobdella, Macrobdella and Hirudo* (all discussed in the text) highlighted. The two trees are in agreement concerning the aforementioned genera.


(a)
100
Haementeria officinalis LAPP M81489
Haementeria depressa EST CN807641
Haementeria depressa EST CN807637
63
100
H. robusta e_gw1.2.270.1
68
H. robusta EST EY349090
100
H. robusta e_gw1.2.291.1
H. robusta estEXT_Genewise1.C_21054
93
100
H. robusta EST EY361718
H. robusta fgenesh4_pg.C_scaffold_2000928
99
100
H. robusta fgenesh4_pg.C_scaffold_2000933
83
H. robusta EST EY344275
H. robusta EST EY344724
-
100
H. robusta EST EY392612
H. robusta EST EY378604
-
H. robusta fgenesh4_pg.C_scaffold_2000930
54
H. robusta fgenesh4_pg.C_scaffold_2000932
100
100
H. robusta EST EY341167
H. robusta fgenesh4_pg.C_scaffold_2000934
H. robusta EST EY389580
97
H. robusta EST EY366835
98
H. robusta EST EY341220
95
Macrobdella decora putative saratin
100
Hirudo medicinalis saratin BD270371
50 changes
(b)
Haementeria
Helobdella
Macrobdella
Hirudo

loci interspersed. Several of the loci derived from the full genome are almost identical to the corresponding *Helobdella robusta* EST sequence. In turn, the *Helobdella* / *Haementeria*-group was recovered as sister to the *Hirudo medicinalis* / *Macrobdella decora* group. To corroborate the phylogenetic hypothesis derived from the *Helobdella* and *Haementeria* LAPP data set, the topology of the LAPP tree was compared to that of a previously published phylogenetic hypothesis of leeches (Fig. 3.2b; Min et al. 2010). The topology of the LAPP tree, concerning the relationships between LAPP orthologues of *Haementeria* and *Helobdella* versus the saratin orthologues, mirrors the topology of the more data-rich phylogeny of leeches (see Discussion).

Discussion

Hematophagous ancestry?

The similarity of predicted genes in the *Helobdella robusta* genome with a known leech salivary gland-secreted anticoagulant (LAPP) is powerful corroboration of contemporary phylogenetic work, which suggest that this non-bloodfeeding leech has a hematophagous ancestry (Siddall & Burreson, 1995, 1996; Trontelj et al. 1999). Because Glossiphoniidae is consistently recovered towards the base of the phylogeny of leeches in independent studies (Siddall & Burreson, 1998; Apakupakul et al. 1999; Siddall et al. 2001; but see Trontelj et al. 1999), the finding of anticoagulants in this leech suggests that possession of anticoagulants is not restricted to taxa in the more derived parts of the phylogenetic tree. That is, our finding belies the long-held notion that hematophagy is a recently derived trait (Mann 1962; Sawyer 1986; see Siddall & Burreson, 1996).

The present study represents the first evidence of an anticoagulant in a non-bloodfeeding glossiphoniid leech. Interestingly, Kim et al. (1996) isolated and biochemically characterized the serine protease inhibitor Guamerin II from the macrophagous hirudinid (*sensu* Phillips & Siddall, 2009) *Whitmania edentula*. Furthermore, Hovingh & Linker (1999) identified a hyaluronoglucuronidase from each of *Erpobdella obscura* and *Erpobdella punctata* (macrophagous Erpobdellidae). Taken together, these findings corroborate independent losses of bloodfeeding throughout the evolutionary history of leeches as hypothesized by Siddall & Burreson (1995, 1996) and Trontelj et al. (1999).

Putative role of LAPP and its genomic positioning in Helobdella robusta

It is not yet clear for what purpose *H. robusta* uses the anticoagulant. However, as almost identical sequences were found also in *H. robusta* EST-libraries, anticoagulant-like proteins seem to be expressed in this leech. Furthermore, several of the putative vWf-blocking proteins contain a signal-peptide-encoding region towards the N-terminus, indicating their secretion by the leech (Fig. 3.1).

Anticoagulation factors and lytic proteases have previously been isolated from oligochaetous clitellates (Popovic et al. 1998; Jeon et al. 1995; Mihara et al. 1991), some specifically able to lyse collagen and laminin (Jeon et al. 1995; Alberts et al. 1994). The paraphyletic status of clitellata with several oligochaetous clitellates diverging earlier than Hirudinida (Rousset et al. 2007), coupled with the apparent patchy expression pattern of anticoagulants and (paralogous) collagen proteases across this group, raises some questions concerning the evolutionary history of the proteins. One hypothesis would be the retention of these proteins throughout clitellata, enabling the change in feeding strategy from macrophagy in several clitellates including oligochaetes to liquidosomatophagy in e.g. *Helobdella robusta*. Like a macrophagous lifestyle, liqiuidosomatophagy would require the breaking down of subendtothelial collagen, with anticoagulation properties as a secondary effect. In turn, this would suggest that the LAPP-like proteins already present in the genomes of related proboscis-bearing (rhyncobdellid) species may be used for dual purposes; as a means for keeping blood flowing in and around the incision wound and as a collagen protease. This breaking down of collagen is not as obvious a need for jaw-bearing (arhynchobdellid) hematophagous species, who restrict their incisions to the skin surface, and this may be why the expression of collagen-binding/lysing proteins is not obvious in these species (e.g. *Hirudo medicinalis* as mentioned above).

Interestingly, all of the eight predicted genes matching LAPP are positioned as tandem repeats in the *H. robusta* genome. This may be due to linked functionality between the loci. For example, where no independent promoter region exists between the loci, the RNA polymerase may simultaneously transcribe them all in a single pass. This would enable rapid, high-copy translation of a variety of LAPP's; given the diversity of collagen (Eyre 1980), it is tempting to speculate that this tandem array of LAPP's would be capable of simultaneously targeting an array of different collagen types. Notably, platelet glycoproteins such as Ib and IIIa exhibit variable number tandem repeats (Carter et al. 1998). Although the platelets targeted by LAPP and saratin are of different kinds than glycoprotein Ib and IIIa, at this stage we cannot rule out that the agonists also have a structural connection to the platelets. Koh & Kini (2009) demonstrated that the structure of both Kazal-type proteinase inhibitors and antistasin-like inhibitors include tandem repeat domains. Unfortunately, that study does not include antiplatelet proteins. X-ray crystallography and/or nuclear magnetic resonance studies would likely shed light on any interactions both between the domains in each LAPP and between these consecutive loci.

Phylogeny

Although our search of the entire GenBank protein database failed to return any putative LAPP orthologues, thus preventing any outgroups to be used in the phylogenetic analysis, there are still three possible major topological outcomes of the phylogenetic analysis conducted in the present study. Specifically, (i) LAPP from the *Haementeria* species could place as sister to saratin, (ii) LAPP from *Helobdella robusta* could place as sister to saratin, and (iii) the *Haementeria* and *Helobdella robusta* orthologues could place as sister to each other. The hypothesis presented here shows the latter topology, and this is congruent with the phylogenetic hypothesis of leeches presented by Min et al. (2010). In other words, the topology of the phylogenetic tree of leeches is in agreement with the tree derived from the LAPP loci in terms of the concerned genera.

Conclusions

Through a combination of similarity analyses (BLASTn, BLASTp, tBLASTn and tBLASTx) and phylogenetic analysis, we have shown that the non-bloodfeeding glossiphoniid leech *Helobdella robusta* possesses putative orthologous to a known anticoagulant, LAPP. In light of previous phylogenetic hypotheses recovering *Helobdella* at the base of the leech tree, this finding suggests that the presence of anticoagulants is plesiomorphic in Hirudinea. Eight LAPP-like loci were found in *H. robusta* and these are represented as a tandem array, a phenomenon that already characterizes several other anticoagulation factors, and that may bring significant benefits to the transcription efficiency of these specific DNA regions.

References

Alberts B, Bray D, Lewis J, Raff M, Roberts K, & Watson, JD, editors 1994. Molecular Biology of the Cell 3rd Edition, 1294 pp. Garland, New York, USA.

Apakupakul K, Siddall ME, & Burreson EM 1999. Higher-level relationships of leeches (Annelida: Clitellata: Euhirudinea) based on morphology and gene sequences. Mol. Phylogenet. Evol. 12: 350-359.

Barnes CS, Krafft B, Frech M, Hofmann UR, Papendieck A, Dahlems U, Gelliessen G, & Hoylaerts MF 2001. Production and characterization of saratin, an inhibitor of von Willebrand factor-dependent platelet adhesion to collagen. Semin. Thromb. Hemost. 27: 337-348.

Bendsten JD, Nielsen H, von Heijne G, & Brunak S 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340: 783-795.

Carter AM, Catto AJ, Bamford JM, & Grant PJ 1998. Platelet GP IIIa, PlA and GP Ib variable number tandem repeat polymorphisms and markers of platelet activation in acute stroke. Arterioscler. Thromb. Vasc. Biol. 18: 1124-1131.

Connolly TM, Jacobs JW, & Condra C 1992. An inhibitor of collagen-stimulated platelet activation from the salivary glands of the *Haementeria officinalis* leech. I. Identification, isolation, and characterization. J. Biol. Chem. 267: 6893-6898.

Cruz CP, Eidt J, Drouilhet J, Brown AT, Wang Y, Barnes CS, & Moursi MM 2001. Saratin, an inhibitor of von Willebrand factor-dependent platelet adhesion, decreases platelet aggregation and intimal hyperplasia in a rat carotid endarterectomy model. J. Vasc. Surg. 34: 724-729.

Eyre DR 1980. Collagen: molecular diversity in the body's protein scaffold. Science 207: 1315-1322.

Fang G, Bhardwaj N, Robilotto R, Gerstein MB 2010. Getting started in gene orthology and functional analysis. PLoS Comput. Biol. 6: e1000703.

Goloboff PA, Farris JS, & Nixon KC 2008. TNT, a free program for phylogenetic analysis. Cladistics 24: 774-786.

Hovingh P, & Linker A 1999. Hyaluronidase activity in leeches (Hirudinea). Comp. Biochem. Physiol. B. 124: 319-326.

Jeon O-H, Moon W-J, & Kim D-S 1995. An anticoagulant/fibrinolytic protease from *Lumbricus rubellus*. J. Biochem. Mol. Biol. 28: 138-142.

Kim DR, Hong SJ, Ha K-S, Joe CO, & Kang KW 1996. A cysteine-rich protease inhibitor (Guamerin II) from the non-blood sucking leech *Whitmania edentula*: biochemical characterization and amino acid sequence analysis. J. Enzyme Inhib. 10: 81-91.

Koh CY, & Kini RM 2009. Molecluar diversity of anticoagulants from hematophagous animals. Expert Review of Hematology 1: 135-139.

Mann KH 1962. Leeches (Hirudinea) Their structure, physiology, ecology and embryology. Pergamon Press, New York, USA.

Mihara H, Sumi H, Yoneta T, Mizumoto H, Ikeda R, & Maruyama M 1991. A novel fibrinolytic enzyme extracted from the earthworm, *Lumbricus rubellus*. Japanese J. Physiol. 41: 461-472.

Min G-S, Sarkar IN, & Siddall ME 2010. Salivary transcriptome of the North American medicinal leech, *Macrobdella decora*. J. Parasitol. 96: 1211-1221.

Obert B, Houllier A, Meyer D, & Girma JP 1999. Conformational changes in the A3 domain of von Wilebrand factor modulate the interaction of the A1 domain with platelet glycoprotein Ib. Blood 93: 1959-1968.

Phillips AJ, Siddall ME 2009. Poly-paraphyly of hirudinidae: many lineages of medicinal leeches. BMC Evol. Biol. 9: 246.

Popovic M, Hrzenjak T, Grdisa M, & Vukovic S 1998. Adhesins of immunoglobulin-like superfamily from earthworm *Eisenia foetida*. Gen. Pharmac. 30; 795-800.

Rousset V, Pleijel F, Rouse GW, Erséus C, & Siddall ME 2007. A molecular phylogeny of annelids. Cladistics 23: 41-63.

Ruggeri ZM 1997. Mechanisms initiating platelet thrombus formation. Thromb. Haemostasis 78: 611-616.

Salzet M 2001. Anticoagulants and inhibitors of platelet aggregation derived from leeches. FEBS Let. 429: 187-192.

Sawyer RT 1986. Leech biology and behaviour. Oxford University Press, Oxford, UK.

Shain DH 2009. Annelids in modern biology. John Wiley, Sons, New York, USA.

Siddall ME, Apakupakul K, Burreson EM, Coates KA, Erséus C, Gelder SR, Källersjö M, & Trapido-Rosenthal H 2001. Validating Livanow: molecular data agree that leeches, branchiobdellidans, and *Acanthobdella peledina* form a monophyletic group of oligochaetes. Mol. Phylogenet. Evol. 21: 346-351.

Siddall, M.E. & Burreson EM 1995. Phylogeny of the Euhirudinea: independent evolution of blood feeding by leeches? Can. J. Zool. 73: 1048-1064.

Siddall, M.E. & Burreson EM 1996. Leeches (Oligochaeta?: Euhirudinea), their phylogeny and the evolution of life history strategies. Hydrobiologia 334: 277-285.

Siddall, M.E. & Burreson EM 1998. Phylogeny of leeches (Hirudinea) based on mitochondrial cytochrome *c* oxidase subunit I. Mol. Phylogenet. Evol. 9: 156-162.

Swofford D 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.0b. Computer software and manual: Sinauer Associates, Sunderland, USA.

Thompson JD, Higgins DG, & Gibson TJ 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

Trontelj P, Sket B, & Steinbrück G 1999. Molecular phylogeny of leeches: congruence of nuclear and mitochondrial rDNA data sets and the origin of bloodsucking. J. Zool. Syst. Evol. Research 37: 141-147.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, & Barton GJ 2009. Jalview version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

Watson S, Berlanga O, Best D, & Frampton J 2000. Update on collagen receptor interactions in platelets: Is the two-state model still valid? Platelets 11: 252-258.

Wernersson R, & Pedersen AG. 2003. RevTrans - Constructing alignments of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31: 3537-3539.

CHAPTER IV

DIVERSITY AND SELECTIVE PRESSURES OF ANTICOAGULANTS IN THREE MEDICINAL LEECH SPECIES (HIRUDINIDA: HIRUDINIDAE, MACROBDELLIDAE).

Abstract

Although medicinal leeches have long been used as treatment for various ailments because of their potent anticoagulation factors, neither the full diversity of salivary components that inhibit coagulation, much less the evolutionary selection acting on them has been thoroughly investigated. Here, we constructed expressed sequence tag libraries from salivary glands of two species of medicinal leeches, *Hirudo verbana* and *Aliolimnatis fenestrata*, and identify anticoagulant-orthologues through BLASTx searches. The identified orthologues then were compared and contrasted to known anticoagulants from a variety of leeches with different feeding habits, including non-sanguivorous species. Moreover, four different statistical methods for predicting signatures of positive and negative evolutionary pressures were used to assess the level and type of selection acting on the molecules as a whole and on specific sites. In total, sequences showing BLASTx-orthology with eleven and seven known anticoagulants were recovered in the *A. fenestrata* and *H. verbana* EST libraries, respectively. Selection pressure analyses estimated high levels of purifying selection across the anticoagulant diversity, yet some isolated sites, some with important an positioning, showed signs of positive selection. These results represent a first attempt at mapping the anticoagulant repertoires in a comparative fashion across several leech families, and not only show that the diversities found in the expression of salivary peptides greatly exceed

expectations but also suggest the feasibility of identifying the important active sites of the proteins through selective pressure analyses.

Introduction

Whereas the documented use of leeches for medicinal purposes dates back over two millennia, emphasis on the utility of leeches in modern medicine is becoming more authoritative (Whitaker et al., 2004; Phillips and Siddall, 2009; Min et al., 2010). The most conspicuous application of leeches is that for relief of venous congestion following flap and digit replantation surgery (Dabb et al., 1992; Soucacos et al., 1994). Critical to this application are leech anticoagulants, proteins that interfere with a normal thrombus formation at various stages of the coagulation cascade, and that play an important role in the leeches ability to feed for extended periods. Most widely exploited of these is hirudin, first extracted from the European medicinal leech *Hirudo medicinalis* Linnaeus, 1758, which binds irreversibly to the fibrinogen exosite of thrombin as well as to the catalytic pocket (Rydel et al., 1990). With an inhibition constant in the picomolar range, it remains the most potent natural direct thrombin inhibitor known (Greinacher and Warkentin, 2008). However, leech salivary glands produce a more diverse pharmacological cocktail of a wide variety of anticoagulants (e.g., Min et al., 2010; Alaama et al., 2011) that not only assist in phlebotomy by keeping blood flowing in and around an incision wound but that also keeps the blood from coagulating inside the leech crop during the substantial periods of digestion, thus preventing inflexibility of the leech body (Salzet, 2001). As an example of the diversity of coagulation-factors targeted by leech anticoagulants, leech antiplatelet protein (LAPP) from the Mexican leech

Haementeria officinalis de Fillippi, 1849, in contrast to hirudin, inhibits von Willebrand factor-mediated, and collagen-stimulated, platelet aggregation by binding to subendothelial collagen (Connolly et al., 1992). Other leech bioactive salivary peptides target (e.g.,) factor Xa, XIIIa, plasmin and hyaluronic acid. Despite the renaissance of leech anticoagulants in medicine, anticoagulant profiles are known for only three of the more than 800 species.

Whereas the European *Hirudo verbana* Carena, 1820 remains the model for biomedical studies on leeches (not *H. medicinalis* as previously thought; Siddall et al. [2007]), much as it is the focal point for several other areas of invertebrate biology (Shain, 2009), other continents are inhabited by hirudiniform counterparts equivalent to *Hirudo verbana* in terms of feeding habits but with less documented use in medicine. These include, but are not limited to, *Macrobdella* spp. in North America, *Aliolimnatis* spp. in Africa, *Hirudinaria* spp in Asia, *Goddardobdella* spp. in Australia and (e.g.,) *Oxyptychus* spp. in South America. Despite the infrequent mention of these leeches in medical contributions, there is some evidence that these leeches historically have been used to treat medical conditions in light of their equivalent bloodfeeding behaviors (Phillips and Siddall, 2009). Sanguivory, however, also occurs in several other, only distantly related, leech families including Glossiphoniidae, Piscicolidae, Praobdellidae, Haemadipsidae and Xerobdellidae (Min et al., 2010). Whereas it has also been hypothesized that bloodfeeding is a derived strategy in leeches, contemporary studies seem to agree on the notion that bloodfeeding is a plesiomorphic strategy (Siddall and Burreson, 1995, 1996; Trontelj et al., 1999; Min et al., 2010). It has even been demonstrated that at least one non-bloodfeeding leech, *Helobdella robusta* Shankland et al., 1992 (Glossiphoniidae), while often recovered in the basal part of the phylogeny of leeches (Siddall et al., 2005; Light & Siddall, 1999), nonetheless possesses ancestrally inherited anticoagulants (Kvist et al., 2011). Min et al. (2010), described the partial transcriptome of the North American medicinal leech, *Macrobdella decora* (Say, 1824), and found several loci with very high sequence similarity to eight previously known anticoagulants in addition to predicted serine protease inhibitors, lectoxin-like c-type lectins, ficolin, disintegrins and histidine-rich proteins. In the same contribution, the authors conclude that “the goal of identifying evolutionarily significant residues associated with biomedically significant phenomena implies continued insights from a broader sampling of blood-feeding leech salivary transcriptomes”. To this end, sampling in a phylogenetic framework through focusing on sanguivorous taxa across the fullness of the leech phylogeny will greatly increase our understanding of the evolution of bloodfeeding in leeches. Moreover, identifying regions under negative and positive selection within the anticoagulant molecules holds the potential to highlight gene regions that are potentially critical to the functionality of the proteins providing a more convincing understanding of the structure-function relationships of anticoagulant proteins.

To address these topics, we here investigate and contrast salivary transcriptomes from three different arynchobdellid leech species, from two different families. We also investigate the general phylogenetic relationships of the orthologues, and assess the level and type of selection pressures acting on the molecules and on specific sites.

Material and methods

Taxon sampling and EST-library creation

Based on their geographic distribution and phylogenetic relationships, two species were chosen for salivary EST-library creation; the European medicinal hirudinoid leech *Hirudo verbana* and the African medicinal hirudinoid leech *Aliolimnatis fenestrata.* These EST libraries then were compared with, and contrasted to, a previously constructed EST-library for the North American medicinal macrobdelloid leech *Macrobdella decora* (Min et al., 2010). Specimens of *Hirudo verbana* were obtained from Leeches USA Ltd. (Westbury, New York) and specimens of *Aliolimnatis fenestrata* were collected in Kasanka National Park, Zambia, from exposed skin while wading in ponds.

Prior to RNA extraction, leeches were washed in 0.5% bleach for 1 min and rinsed in deionized water for 1 min in order to minimize contamination of surface bacteria. Using sterilized tools, salivary tissue masses (glandular tissue) were removed aseptically by dissection while immersed in RNA*later* (Qiagen, Valencia, California) and subsequently rinsed in 0.5% bleach for 1 min and rinsed in deionized water for 1 min. RNA then was isolated using RNeasy Tissue kit (Qiagen). Subsequent construction of cDNA libraries, as well as low-quality sequence and repeat masking, follow the protocol detailed by Min et al. (2010).

Similarity and identification using BLASTx

A relational database for all EST sequences from all three species was created in FileMaker Pro (FileMaker, Santa Clara, California) following the removal of low-quality sequences as determined with Sequence Analysis Software ver. 5.4 (Applied Biosystems). At this point, sequences shorter than 150 bp in length also were removed from the data set. As noted by Min et al. (2010), vector and adaptor sequence removal was not necessary due to the use of Smart-seq sequencing primer, which anneals to within 3 bp of the cloned insert. Nonetheless, the first 20–30 bp were automatically trimmed so as to minimize the inclusion of 5' sequencing errors. Sequences were clustered locally using a BLASTn protocol based on an inclusion criterion of $1E^{-5}$ similarity score, and each cluster was assigned a unique identifier number.

High-quality, non-repetitive sequences were employed as queries in a BLASTx search (searching a protein data base using a translated nucleotide query) against a locally compiled set of known anticoagulants including the following accessions: Q07558 hirudin from *Hirudo medicinalis*, P84590 hirudin from *Poecilobdella viridis* (Moore, 1927), P28504 hirudin II from *H. medicinalis*, P26631 hirulin from *Hirudinaria* manillensis (Lesson, 1842), P09865 bdellin from *H. medicinalis*, AAA96144 destabilase I from *H. medicinalis*, AAA96143 destabilase II from *H. medicinalis*, AAN28679 cystatin from *Theromyzon tessulatum* (Müller, 1774), 0905140A eglin c from *H. medicinalis*, Q01747 leech antiplatelet protein from *Haementeria officinalis*, P17350 decorsin from *Macrobdella decora*, Q9NBW4 therostasin from *T. tessulatum*, AAB21233 ghilanten from *Haementeria ghilianii* de Filippi, 1849, P16242 ghilanten from *H. ghilianii*, AAA29193 antistasin from *H. officinalis,* P15358 antistasin from *H. officinalis*, AAD09442 guamerin from *Hirudo nipponia* Whitman, 1886, P80302 hirustasin from *H. medicinalis*, 2K13-X saratin from *H. officinalis,* patent 2006_US_7.049.124_B1 manillase from *H. manillensis* and a transcript from a previous

M. decora EST library (Min et al. 2010) matching ficolin (cluster 686). The BLASTx search used a similarity cut-off value of 1E^{-5}. Moreover, the anticoagulant data set was queried both against a stand-alone EST database for *H. medicinalis* on the Hirudinea Genomics Consortium website (http://genomes.sdsc.edu/leechmaster/database/) and against the genome of *Helobdella robusta*, available at the Joint Genome Institute (JGI) portal website (http://genome.jgi-psf.org/Helro1/Helro1.home.html). Both searches employed a cutoff e-value of 1E^{-5}.

Sequences matching known anticoagulants were submitted to CodonCode Aligner (Codoncode Corp., Dedham, Massachusetts) where they were reconciled into unigene sequences using a 95% minimum percent identity cut-off and a 25% minimum overlap length between sequences. The longest open-reading frame (ORF) of the single representative (i.e., the "reference" sequence) of the multiple reconciled sequences was retrieved. When irreconcilable, the longest ORF for each individual sequence was retained. As a cross-control of the ORF's, reference sequences were translated into amino acids using six-frame translation on the ExPASy Bioinformatics Resource Portal website (http://web.expasy.org/translate/). The longest nucleotide ORF within the sequences was confirmed and all sequences were asserted to be in first frame (i.e., the first position of the sequence was the first codon position in all cases). In cases where newly generated EST sequences were substantially longer than the archetypal anticoagulant sequences, these were truncated at either the 5' end, 3' end or both. Prediction of signal peptides, which suggest secretion of the protein, was performed on the SignalP 4.0 (Petersen et al., 2011) server at http://www.cbs.dtu.dk/services/SignalP/. Jalview ver. 2 (Waterhouse et al. 2009) was used to visualize the alignments in order to assess the level of conservation between translated sequences and protein sequences of the archetypal anticoagulants.

Phylogenetic analyses

Nucleotide sequences of the unigenes were aligned with the respective known anticoagulant and in accordance with their inferred amino acid states by employing Dialign-T (Subramanian et al., 2005) as implemented in RevTrans 1.4 (Wernersson and Pedersen, 2003). The separate alignments for each anticoagulant then were submitted to TNT (Goloboff et al., 2008) for phylogenetic analyses under the parsimony criterion. A traditional search was performed for each data set employing 100 initial addition sequences and TBR branch swapping. All characters were un-weighted and non-additive and gaps were treated as a fifth state. All trees were left unrooted.

Analysis of evolutionary selection

Each anticoagulant alignment was analyzed for selection pressures acting on the full molecule by implementation of the PARRIS method (Scheffler et al., 2006) in HyPhy (Kosakovsky Pond et al., 2005), and for site-specific selection using the codon-based likelihood ratio tests, Fixed Effects Likelihood (FEL), internal Fixed Effects Likelihood (iFEL) and Random Effects Likelihood (REL). Statistical significance ($p<0.05$) for ω (the ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site; *dN*/*dS*) was assessed in HyPhy and all models of evolution relating to these analyses were predicted using the same software. In addition, HyPhy was used to plot the likelihood ratio test (LRT) scores for each codon position resulting from the FEL analyses.

Results

Anticoagulant diversity

After removal of low-quality and repetitive sequences, 1555 and 1800 sequences remained for the EST libraries of *Aliolimnatis fenestrata* and *Hirudo verbana*, respectively; already, 2019 sequences were available for *Macrobdella decora* (Min et al., 2010).

For *Aliolimnatis fenestrata,* the 1555 sequences assembled into 408 distinct clusters. The BLASTx search returned hits within the *A. fenestrata* EST library for each of 11 well-characterized anticoagulants, as well as elastase inhibitors, eglin C and plasmin inhibitors at e-values better than $1E^{-5}$ (Table 4.1). Factor Xa-inhibiting proteins (antistasins, including the leech-isolated ghilanten, hirustasin, therostasin, guamerin and piguamerin) were the most frequently found anticoagulants; 209 out of the total 1555 sequences matched antistasin-family proteins at e-values of $1E^{-5}$ or better. Sequence reconciliation implied one major and one minor unigene transcript. The highest scoring transcript showed an average amino acid identity of 49% when compared to P15358 antistasin from *Haementeria officinalis*, AAB21233 ghilanten from *Haementeria ghilianii*, AAD09442 guamerin from *Hirudo nipponia* and P80302 hirustasin from *Hirudo medicinalis*. The second most frequently recovered anticoagulant was saratin/LAPP; 102 sequences matched the archetypal sequences at $1E^{-5}$ or better and these reconciled into three major unigene transcripts. Two additional transcripts, each

Table 4.1. Top anticoagulant BLASTx hits in each of the three EST-libraries using a locally compiled set of known anticoagulants as targets.

	Aliolimnatis fenestrata	*Hirudo verbana*	*Macrobdella decora*
Hirudin	$9.1E^{-31}$	-	$1.5E^{-7}$
Haemadin	$2.8E^{-12}$	-	-
Destabilase I	-	-	$1.6E^{-60}$
Destabilase II	-	-	$2.5E^{-13}$
Saratin	$1.7E^{-36}$	-	$2.6E^{-45}$
Bdellin	$4.3E^{-11}$	$1.1E^{-6}$	$2.0E^{-6}$
Piguamerin	$7.3E^{-7}$	$5.6E^{-7}$	$5.6E^{-7}$
Antistasin	$6.0E^{-20}$	$4.8E^{-20}$	$9.5E^{-22}$
Ghilanten	$1.6E^{-20}$	$4.8E^{-20}$	$5.6E^{-22}$
Hirustasin	$3.4E^{-19}$	$8.9E^{-6}$	-
Therostasin	$1.7E^{-7}$	$3.9E^{-7}$	$5.5E^{-6}$
Ornatin	-	-	$4.3E^{-7}$
LAPP	$4.9E^{-8}$	-	$4.8E^{-8}$
Decorsin	-	-	$7.0E^{-18}$
Elastase inhibitor	$2.2E^{-8}$	$5.1E^{-42}$	$3.5E^{-8}$
Eglin c	$4.5E^{-8}$	$2.2E^{-16}$	$1.9E^{-7}$
Heparanases	$4.5E^{-71}$	$7.0E^{-159}$	$5.7E^{-13}$
Plasmin inhibitor	$1.1E^{-12}$	$1.4E^{-12}$	$4.6E^{-9}$

represented by a single sequence, showed e-values equal to or better than $1E^{-5}$ when BLASTed against saratin. The transcript with the best e-value displayed 62% amino acid identity when compared to 2K13-X saratin from *Haementeria officinalis*. Putative manillase orthologues comprised six clones in one cluster, all of which reconciled into a single transcript. The percentage of shared amino acid positions between the highest scoring transcript and Patent no. 2006 US 7.049.124 B1 manillase from *Hirudinaria manillensis* was 69%. Further, three irreconcilable putative bdellin orthologs matched the archetypal sequence for bdellin, each thus corresponding to their own unigene. The shared amino acid identity between the best scoring of these and P09865 bdellin from *Hirudo medicinalis* was 43%. One single sequence representing a putative hirudin orthologue was recovered in the *A. fenestrata* EST library at $9.1E^{-31}$ (Table 4.1). This transcript showed 53% amino acid identity, on average, when compared to P28504 hirudin II from *Hirudo medicinalis*, P84590 hirudin from *Poecilobdella viridis* and P26631 hirulin from *Hirudinaria manillensis*. One cluster, including three transcripts, reconciled into a single unigene that matched a previously determined ficolin sequence from *Macrobdella decora*. When compared to the archetypal sequence, this unigene showed 61% amino acid identity. Finally, a single transcript matched 0905140A eglin c at $4.5E^{-8}$ (Table 4.1) and shared 35% amino acid identity with the archetypal sequence.

For *Hirudo verbana,* the 1800 total sequences assembled into 419 clusters. Putatively orthologous sequences were found in the *H. verbana* EST library for each of seven known anticoagulants in addition to elastase inhibitors, eglin C and plasmin inhibitors at equal to or better than $1E^{-5}$ (Table 4.1). The most frequently found anticoagulants again belonged to the antistasin-family; these were represented by 14 sequences, which reconciled into three major transcripts. An additional singleton sequence matched antistasin-family proteins at 1E^{-5}. The best scoring transcript showed 41% amino acid identity with the archetypal anticoagulants. A total of 10 sequences, nine of which reconciled into a single unigene, matched bdellin at 1E^{-5} or better. When compared to P09865 bdellin from *Hirudo medicinalis*, the best scoring transcript displayed 51% amino acid identity. A single transcript matching each of manillase, hirudin, eglin c, ficolin and a putative thrombin-inhibiting hirudin-like orthologue were represented by a single sequence in the *H. verbana* EST library (Table 4.1). The putative endoglucuronidase (manillase) transcript displayed 69% similarity with patent no. 2006US7049124B1 manillase from *Hirudinaria manillensis*, whereas the putative hirudin transcript showed an average identity of 32% with P28504 hirudin II from *Hirudo medicinalis*, P84590 hirudin from *Poecilobdella viridis* and P26631 hirulin from *Hirudinaria manillensis*.

In addition, when screening the stand-alone *Hirudo medicinalis* EST library at a 1E^{-5} cutoff level, sequences showing putative orthology with nine well-characterized anticoagulants and five other leech bioactive salivary peptides were found. These included manillase, orgelase, destabilase, bdellin, saratin/LAPP, ghilanten, antistasin, therostasin, ficolin, leucocyte elastase inhibitors, eglin c, c-type lectin and cystatin, Sequences showing putative orthology with saratin/LAPP proteins were also found in the genome of the non-bloodfeeding *Helobdella robusta*.

Phylogeny reconstructions

The antistasin-family data set comprised 627 aligned sites, 275 being parsimony-informative. The analysis of these yielded a single most parsimonious tree, 2079 steps long (Fig. 4.1). The known anticoagulants M24423 antistasin from *Haementeria officinalis* and U20787 ghilanten from *Haemeteria ghilianii* form a monophyletic cluster sister to a larger set of newly generated sequences from *H. verbana* and *A. fenestrata,* as well as a single sequence from *H.* medicinalis. By contrast, U38282 guamerin from *Hirudo nipponia* places as sister to a clade consisting of predicted piguamerin and hirustasin, and orthologues from each of *A. fenestrata*, *H. medicinalis, H. depressa* and *T. tessulatum* (the archetypal sequence for theromyzon).

The saratin data set consisted of 669 aligned sites, 336 of which were parsimony informative. Analysis of the saratin data set resulted in a single most parsimonious tree with 2279 steps (Fig. 4.2). Three main monophyletic clusters can be identified in the tree: the first exclusively containing orthologues derived from the EST library of *Macrobdella decora,* the second including sequences from only glossiphoniid taxa (*Helobdella robusta* and archetypal sequences from *Haementeria officinalis* and *Haementeria depressa*), and the third including sequences from each of *A. fenestrata, M. decora* and *H. medicinalis*. In the latter cluster, an antiplatelet protein from *Hirudo medicinalis* (GenBank accession BD270371) groups with a saratin-like orthologue from *Macrobdella decora* as well as a sequence derived from the stand-alone *H. medicinalis* EST library (with very short internal branches).

Fig. 4.1 Unrooted single most parsimonious tree recovered from analysis of the antistasin-family data set (L=2079, CI=0.632, RI=0.642). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are drawn proportional to change.

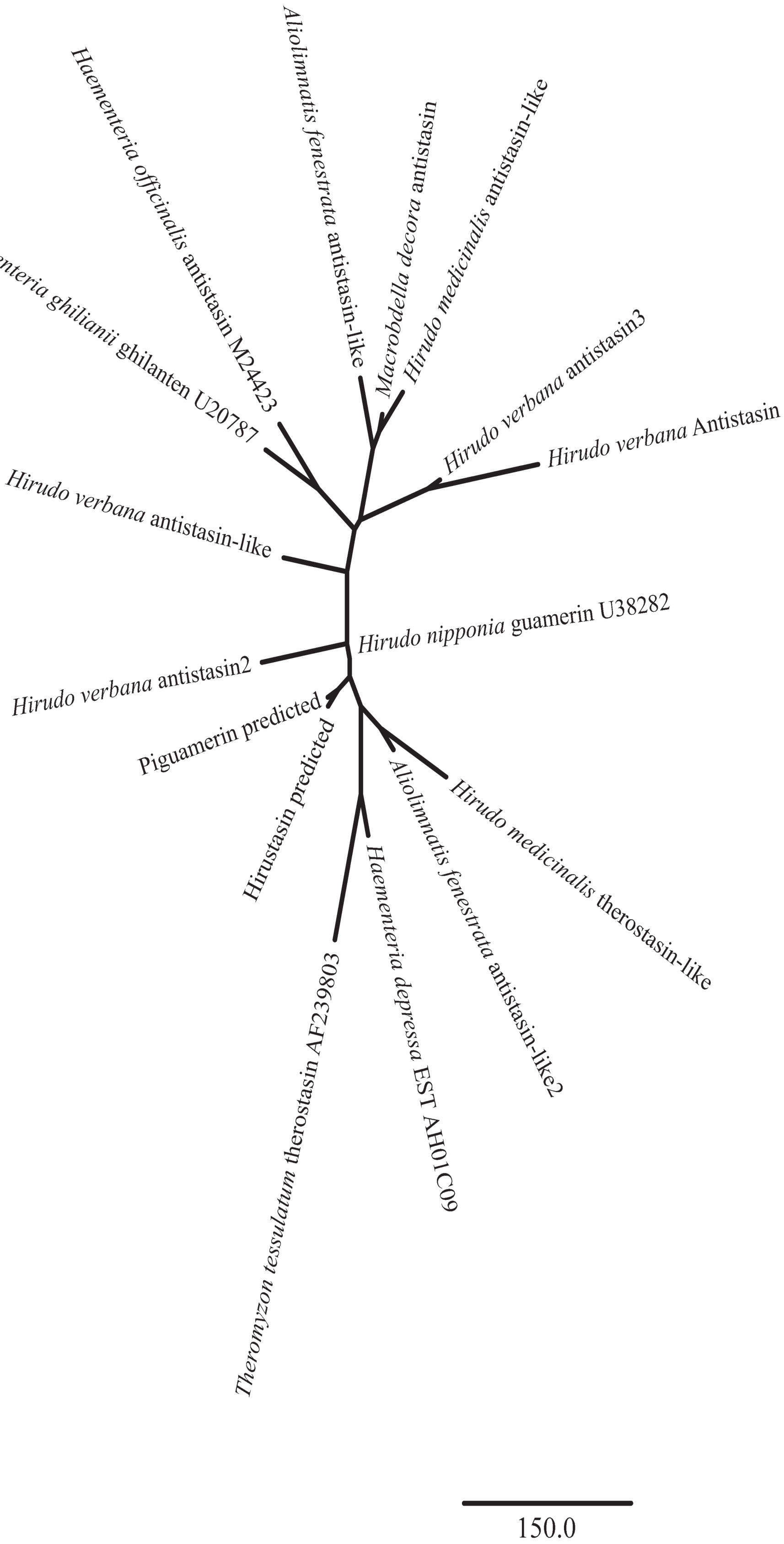
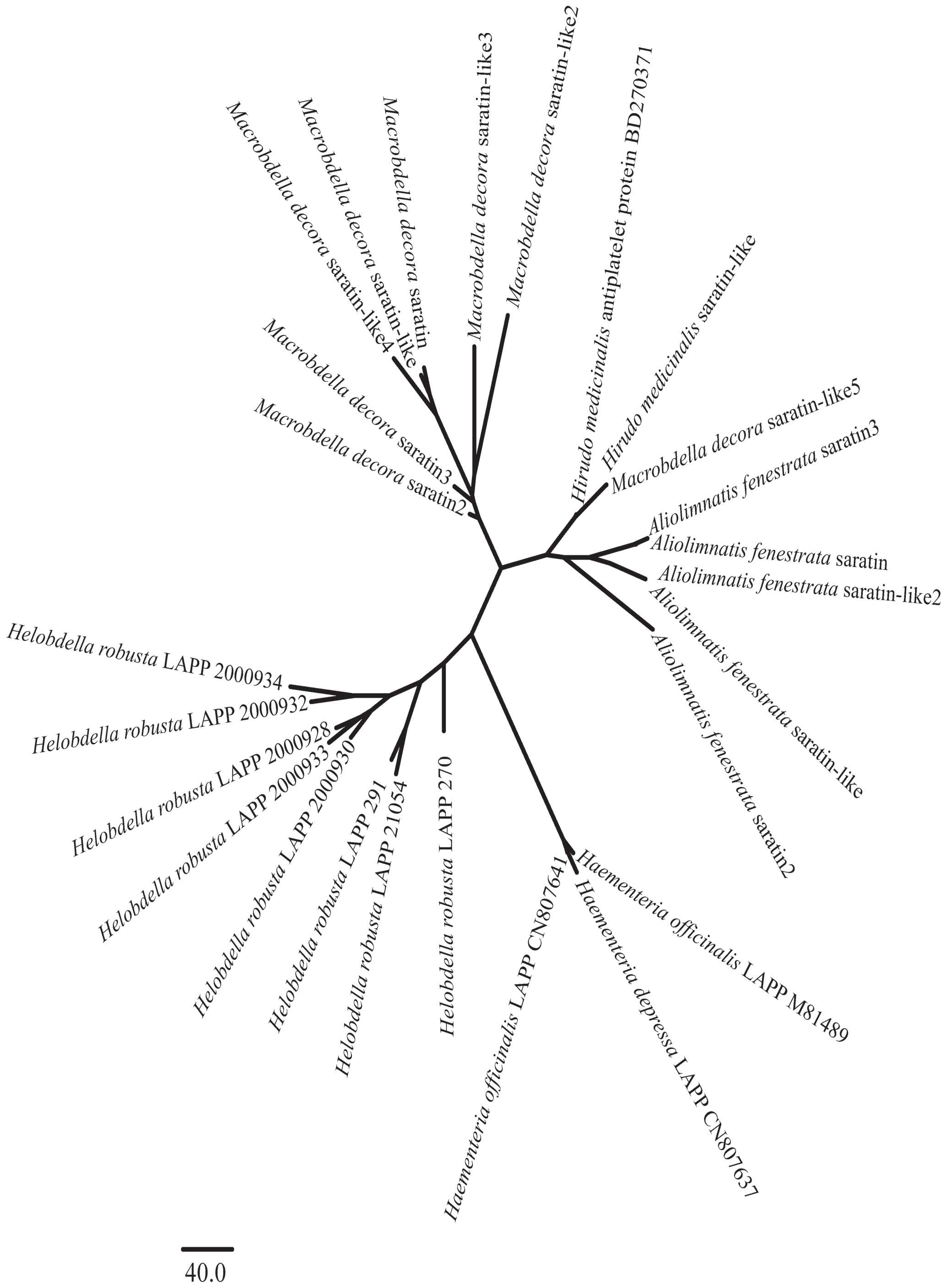

Haementeria ghilianii ghilanten U20787
Haementeria officinalis antistasin M24423
Aliolimnatis fenestrata antistasin-like
Macrobdella decora antistasin
Hirudo medicinalis antistasin-like
Hirudo verbana antistasin3
Hirudo verbana Antistasin
Hirudo verbana antistasin-like
Hirudo nipponia guamerin U38282
Hirudo verbana antistasin2
Piguamerin predicted
Hirustasin predicted
Aliolimnatis fenestrata antistasin-like2
Hirudo medicinalis therostasin-like
Theromyzon tessulatum therostasin AF239803
Haementeria depressa EST AH01C09
150.0

Fig. 4.2 Unrooted strict consensus of two equally parsimonious trees recovered from analysis of the saratin data set (L=2279, CI=0.587, RI=0.740). When appropriate, GenBank accession numbers follow taxon names. For *Helobdella robusta* orthologues, the numbers following the taxon name correspond to JGI scaffold for the full genome sequencing. Branch lengths are drawn proportional to change.


Macrobdella decora saratin-like4
Macrobdella decora saratin-like
Macrobdella decora saratin
Macrobdella decora saratin-like3
Macrobdella decora saratin-like2
Hirudo medicinalis antiplatelet protein BD270371
Hirudo medicinalis saratin-like
Macrobdella decora saratin3
Macrobdella decora saratin2
Macrobdella decora saratin-like5
Aliolimnatis fenestrata saratin3
Aliolimnatis fenestrata saratin
Aliolimnatis fenestrata saratin-like2
Aliolimnatis fenestrata saratin-like
Aliolimnatis fenestrata saratin2
Helobdella robusta LAPP 2000934
Helobdella robusta LAPP 2000932
Helobdella robusta LAPP 2000928
Helobdella robusta LAPP 2000933
Helobdella robusta LAPP 2000930
Helobdella robusta LAPP 291
Helobdella robusta LAPP 21054
Helobdella robusta LAPP 270
Haementeria officinalis LAPP CN807641
Haementeria depressa LAPP CN807637
Haementeria officinalis LAPP M81489
40.0

The hirudin data set consisted of 465 aligned sites, 60 of which were parsimony informative. The analysis resulted in a single most parsimonious tree with 649 steps (Fig. 4.3). The archetypal sequences from *H. medicinalis* (GenBank accessions M12693 and A14988) place as sister to the *Aliolimnatis fenestrata* EST sequence. By contrast, the thrombin inhibitor and haemadin sequences, both from *Haemadipsa sylvestris* (Z19864 and S58792, respectively), form a monophyletic cluster with the EST sequence derived from *Macrobdella decora*.

The bdellin data set included 240 aligned sites, 111 of which were parsimony informative. The parsimony analysis returned two most parsimonious trees 565 steps long; the unrooted strict consensus of these is shown in Fig. 4.4. The known sequence of bdellin from *Hirudo nipponia* (GenBank accession AF223972) forms a monophyletic group with two sequences derived from the *Hirudo medicinalis* stand-alone EST library. In turn, this group places as sister to a sequence derived from *Macrobdella decora*. Beyond this, the remaining sequences are interspersed across the tree without relevant taxonomic clustering.

The endoglucuronidase (manillase) data set included 1257 aligned sites, 106 being parsimony informative. The parsimony analysis returned a single most parsimonious tree, 2288 steps long (Fig. 4.5). As nucleotide sequences have yet to be generated for the leech-derived endoglucuronidases manillase and orgelase, an already existing EST sequence (EY484527) showing orthology at the amino acid level with manillase acted as the archetypal variant in the data set. In the resulting tree, this sequence is sister to a clade containing orthologues from *Hirudo medicinalis* and *Hirudo*

Fig. 4.3 Unrooted single most parsimonious tree recovered from analysis of the hirudin data set (L=649, CI=0.924, RI=0.809). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are drawn proportional to change.

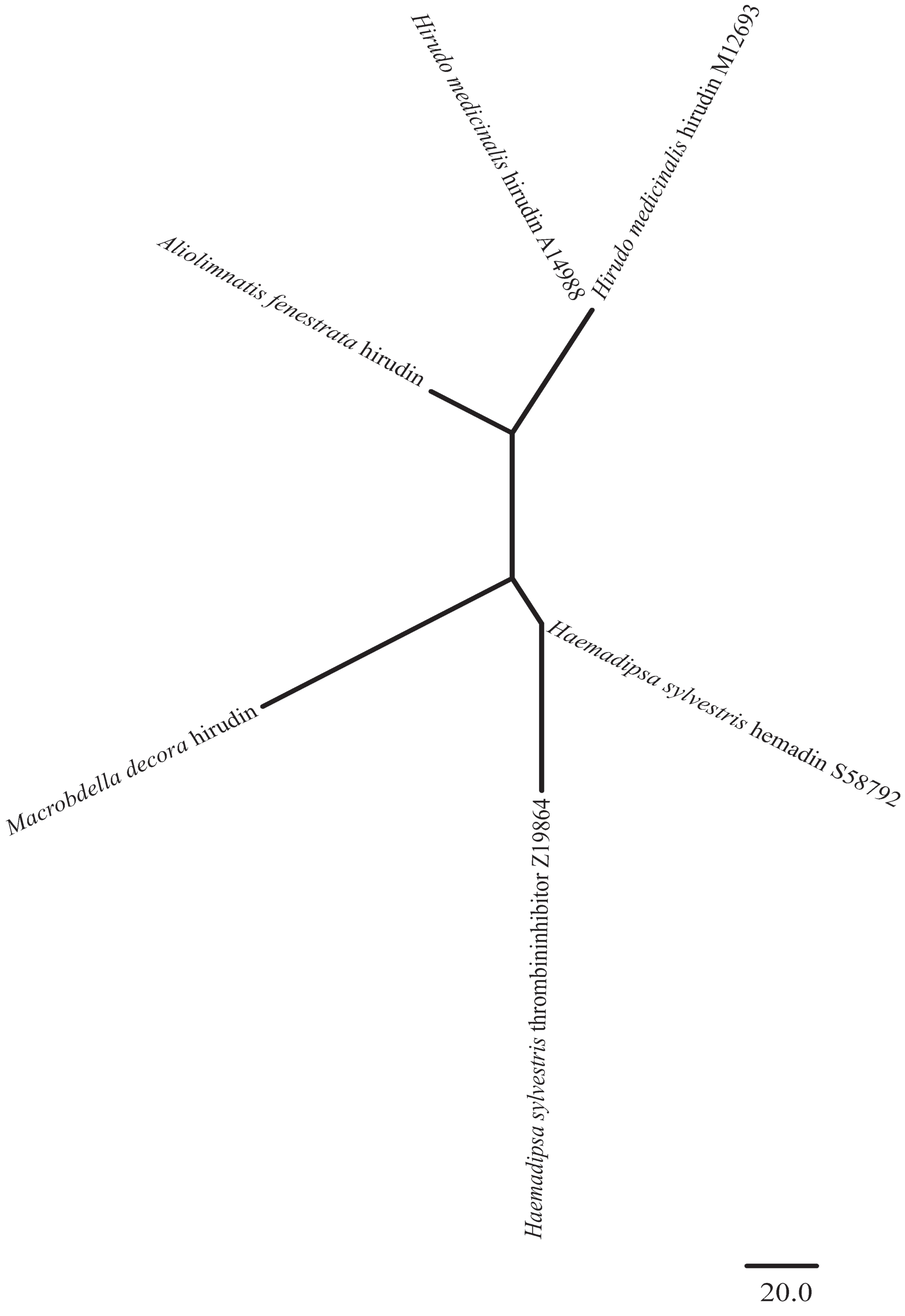

Hirudo medicinalis hirudin A14988
Hirudo medicinalis hirudin M12693
Aliolimnatis fenestrata hirudin
Macrobdella decora hirudin
Haemadipsa sylvestris hemadin S58792
Haemadipsa sylvestris thrombininhibitor Z19864
20.0

Fig. 4.4 Unrooted strict consensus of two equally parsimonious trees recovered from analysis of the bdellin data set (L=565, CI=0.742, RI=0.738). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are drawn proportional to change.

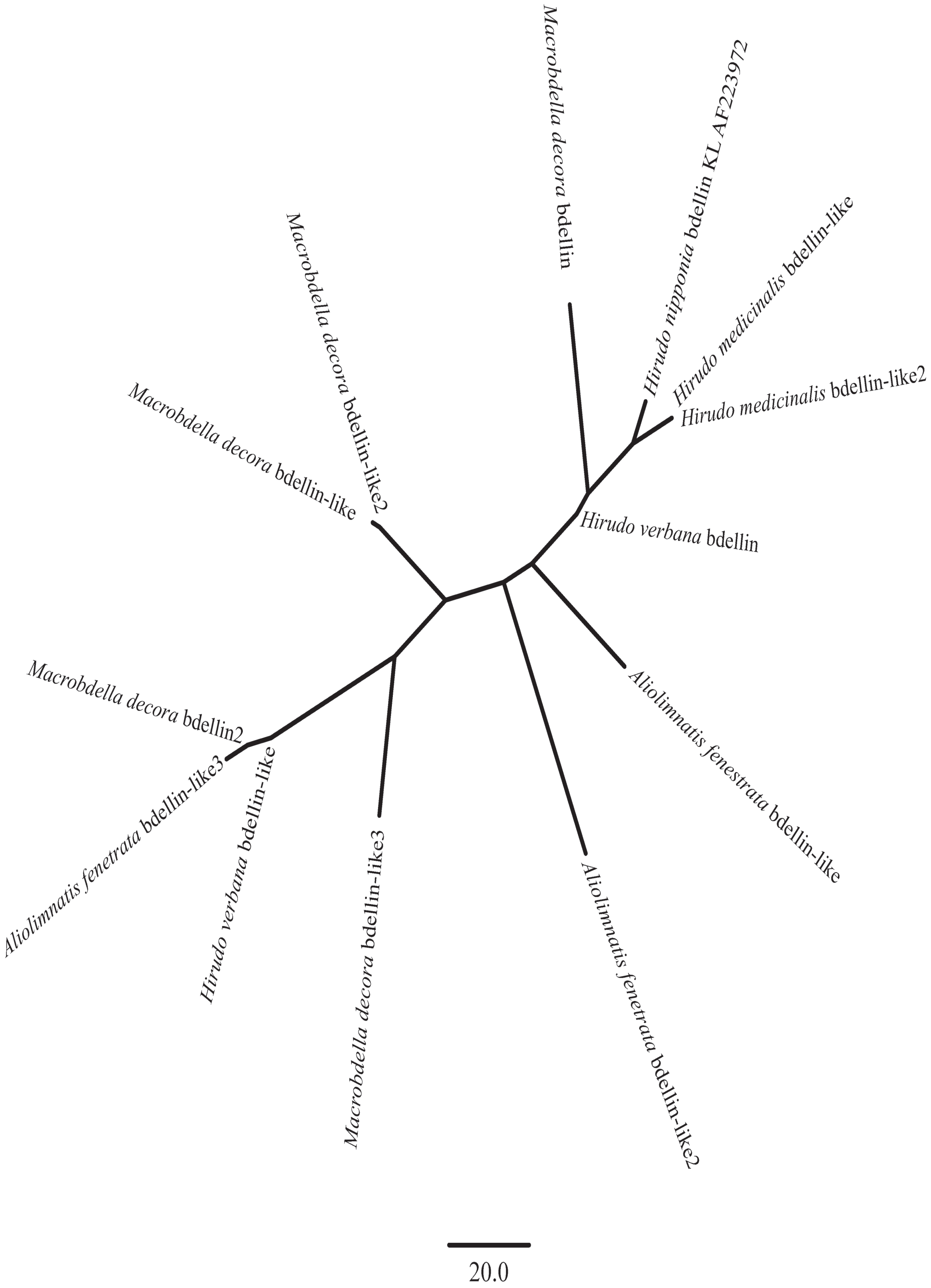
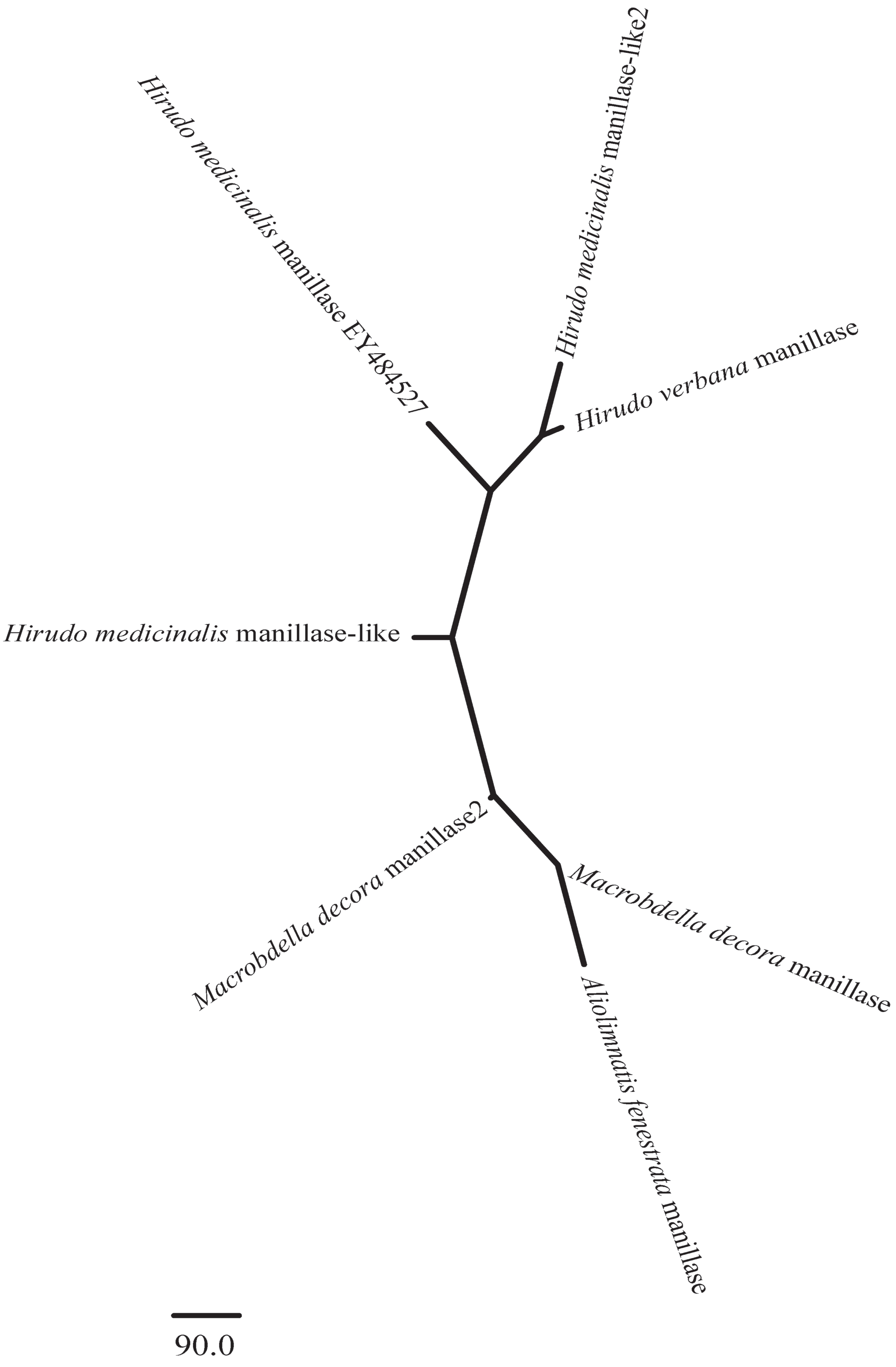

Macrobdella decora bdellin
Hirudo nipponia bdellin KL AF223972
Hirudo medicinalis bdellin-like
Hirudo medicinalis bdellin-like2
Hirudo verbana bdellin
Aliolimnatis fenestrata bdellin-like
Aliolimnatis fenetrata bdellin-like2
Macrobdella decora bdellin-like3
Hirudo verbana bdellin-like
Aliolimnatis fenetrata bdellin-like3
Macrobdella decora bdellin2
Macrobdella decora bdellin-like
Macrobdella decora bdellin-like2
20.0

Fig. 4.5 Unrooted single most parsimonious tree recovered from analysis of the manillase (heparanase-class endoglucuronidase) data set (L=2288, CI=0.891, RI=0.852). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are drawn proportional to change.


Hirudo medicinalis manillase EY484527
Hirudo medicinalis manillase-like2
Hirudo verbana manillase
Hirudo medicinalis manillase-like
Macrobdella decora manillase2
Macrobdella decora manillase
Aliolimnatis fenestrata manillase
90.0

verbana. In turn, this cluster places as sister to the remaining sequences from *A. fenestrata*, *M. decora* and *H. medicinalis*.

The decorsin data set comprised orthologues from *Macrobdella decora* and putative orthologues from the stand-alone *Hirudo medicinalis* EST library. The final data set included 210 aligned positions, 23 of which were parsimony informative. The analysis returned a single most parsimonious tree with 144 steps (Fig. 4.6). The two sequences from the patents for decorsin form a dichotomy and so do the *Macrobdella decora* orthologues.

Destabilase orthologues (matching at 1E^{-5} or better) were only found in *Macrobdella decora* and *Hirudo medicinalis*. The full data set comprised 882 aligned sites (211 parsimony informative characters) and analysis of these resulted in a single most parsimonious tree 453 steps long (Fig. 4.7). The archetypal sequence and the *Hirudo medicinalis* ESTs cluster together, as sister to two *Macrobdella decora* orthologues. Sister to this group is a large cluster of *M. decora* sequences.

The ficolin data set consisted of 648 aligned sites (132 parsimony informative) and the analysis of these resulted in a single tree, 967 steps long (Fig. 4.8). The "archetypal" ficolin sequence, derived from the *M. decora* EST library, places as sister to two of the three *Hirudo medicinalis* EST sequences; the remaining sequence groups with the single *Aliolimnatis fenestrata* orthologue.

For eglin c, the compiled data set consisted of 384 characters, 41 of which were parsimony informative. The parsimony analysis resulted in a single tree with 321 steps. In the tree (Fig. 4.9), the sequences derived from *H. medicinalis* cluster and place

Fig. 4.6 Unrooted single most parsimonious tree recovered from analysis of the decorsin data set (L=144, CI=1.000, RI=1.000). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are drawn proportional to change.

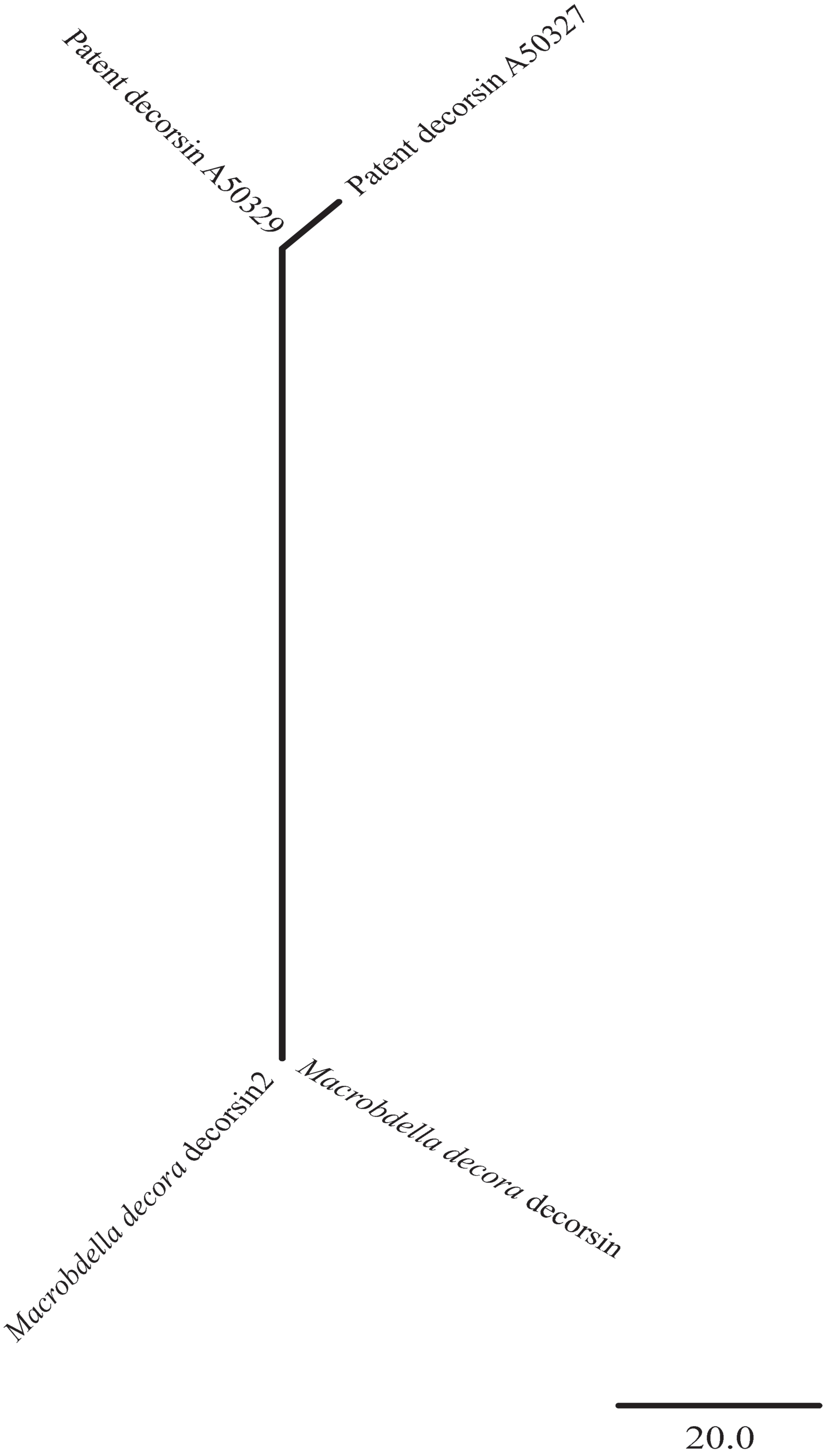

Patent decorsin A50329
Patent decorsin A50327
Macrobdella decora decorsin2
Macrobdella decora decorsin
20.0

Fig. 4.7 Unrooted single most parsimonious tree recovered from analysis of the destabilase data set (L=1023, CI=0.811, RI=0.664). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are drawn proportional to change.

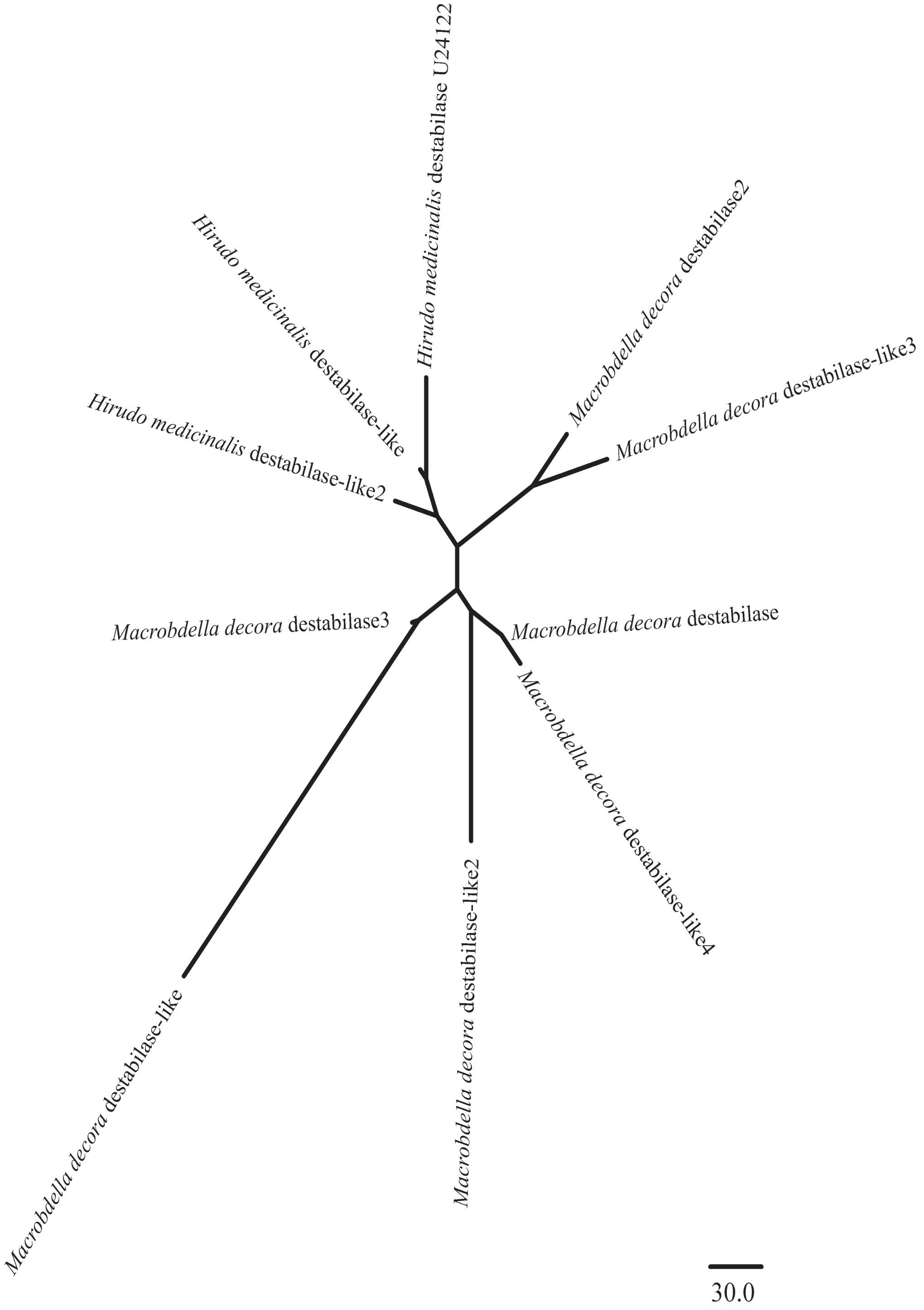

Hirudo medicinalis destabilase U24122
Hirudo medicinalis destabilase-like
Hirudo medicinalis destabilase-like2
Macrobdella decora destabilase2
Macrobdella decora destabilase-like3
Macrobdella decora destabilase3
Macrobdella decora destabilase
Macrobdella decora destabilase-like4
Macrobdella decora destabilase-like2
Macrobdella decora destabilase-like
30.0

Fig. 4.8 Unrooted single most parsimonious tree recovered from analysis of the ficolin data set (L=967, CI=0.896, RI=0.646). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are drawn proportional to change.

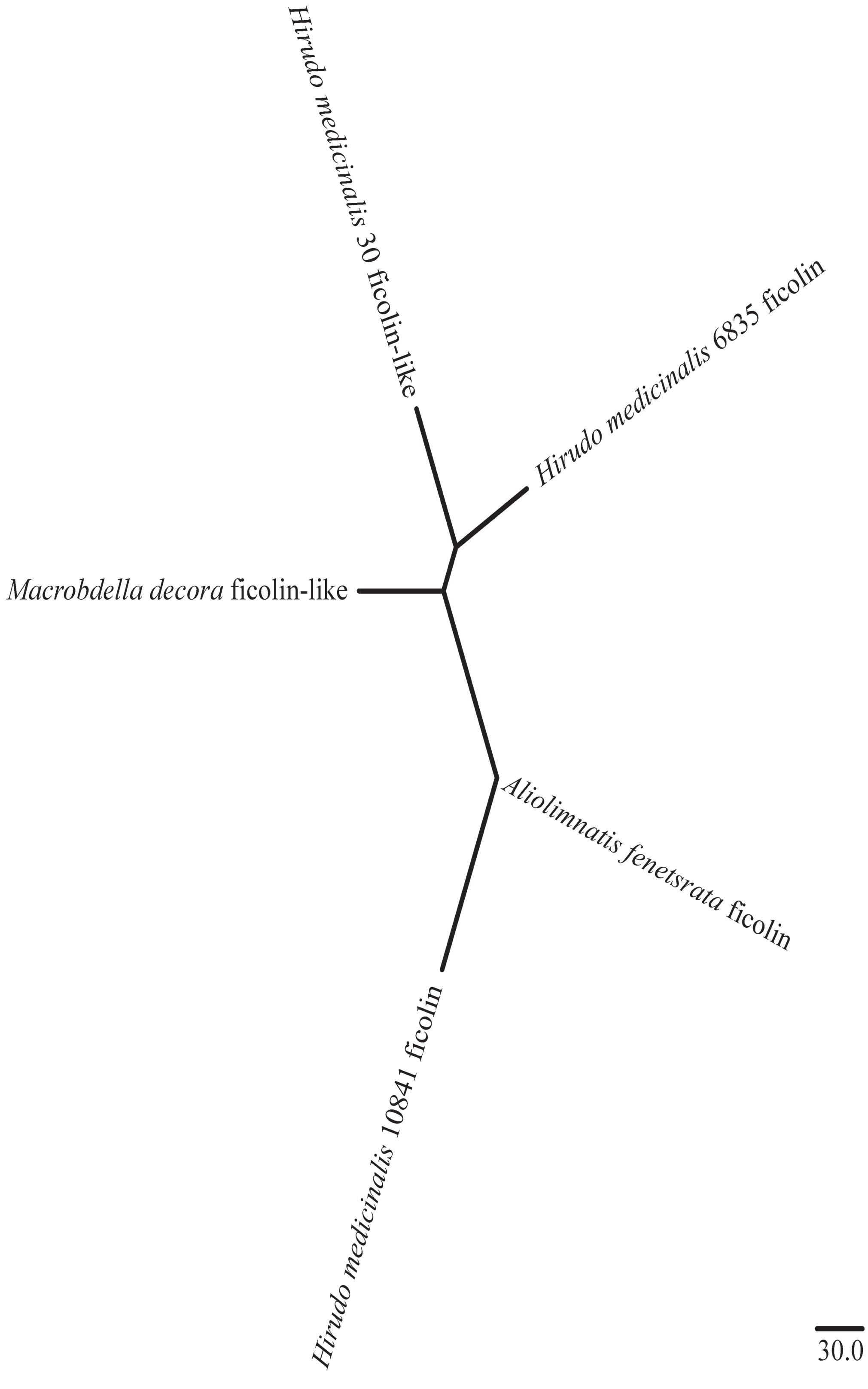
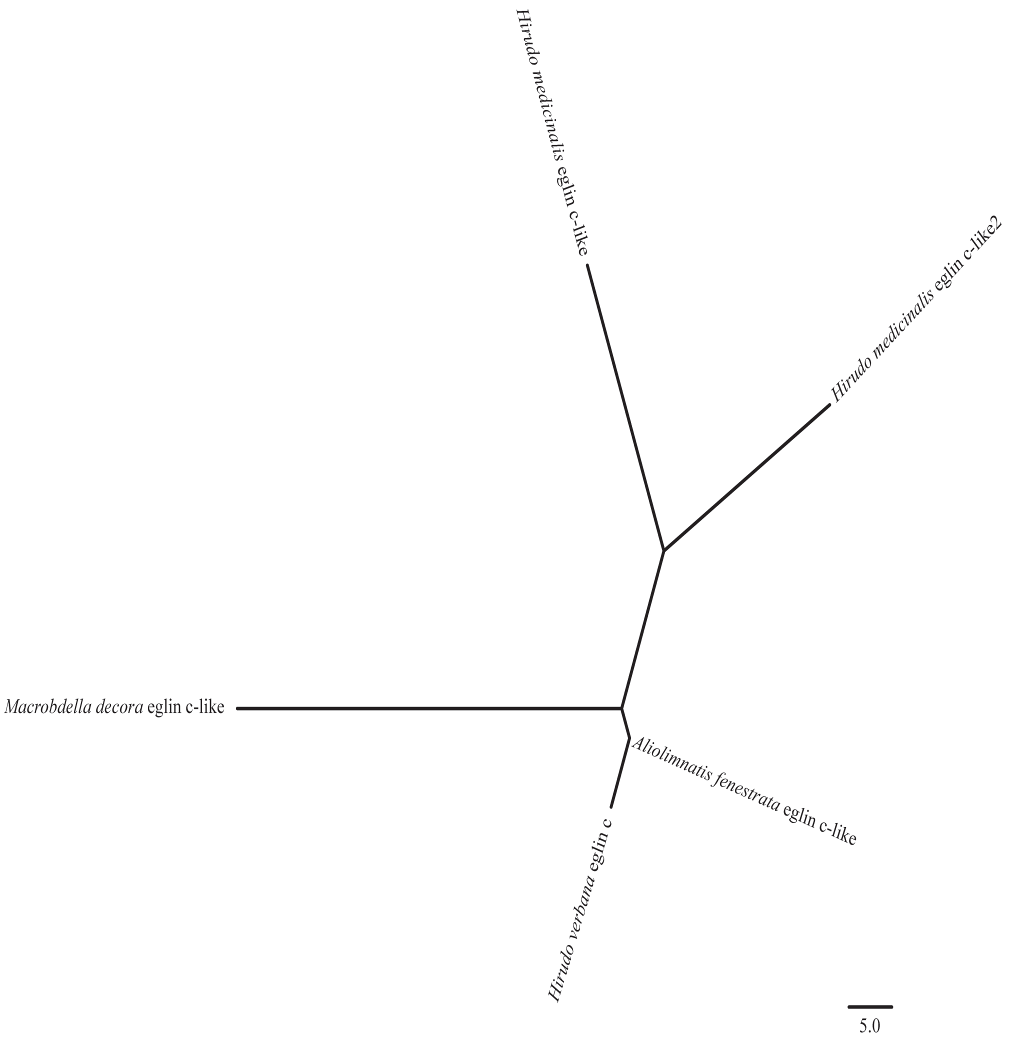

Hirudo medicinalis 30 ficolin-like
Hirudo medicinalis 6835 ficolin
Macrobdella decora ficolin-like
Aliolimnatis fenetsrata ficolin
Hirudo medicinalis 10841 ficolin
30.0

Fig. 4.9 Unrooted single most parsimonious tree recovered from analysis of the eglin c data set (L=321, CI=0.938, RI=0.512). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are drawn proportional to change.


Hirudo medicinalis eglin c-like
Hirudo medicinalis eglin c-like2
Macrobdella decora eglin c-like
Aliolimnatis fenestrata eglin c-like
Hirudo verbana eglin c
5.0

as sister to a cluster containing the remaining sequences from the three EST libraries of *M. decora*, *H. verbana* and *Aliolimnatis fenestrata*.

Selection pressures

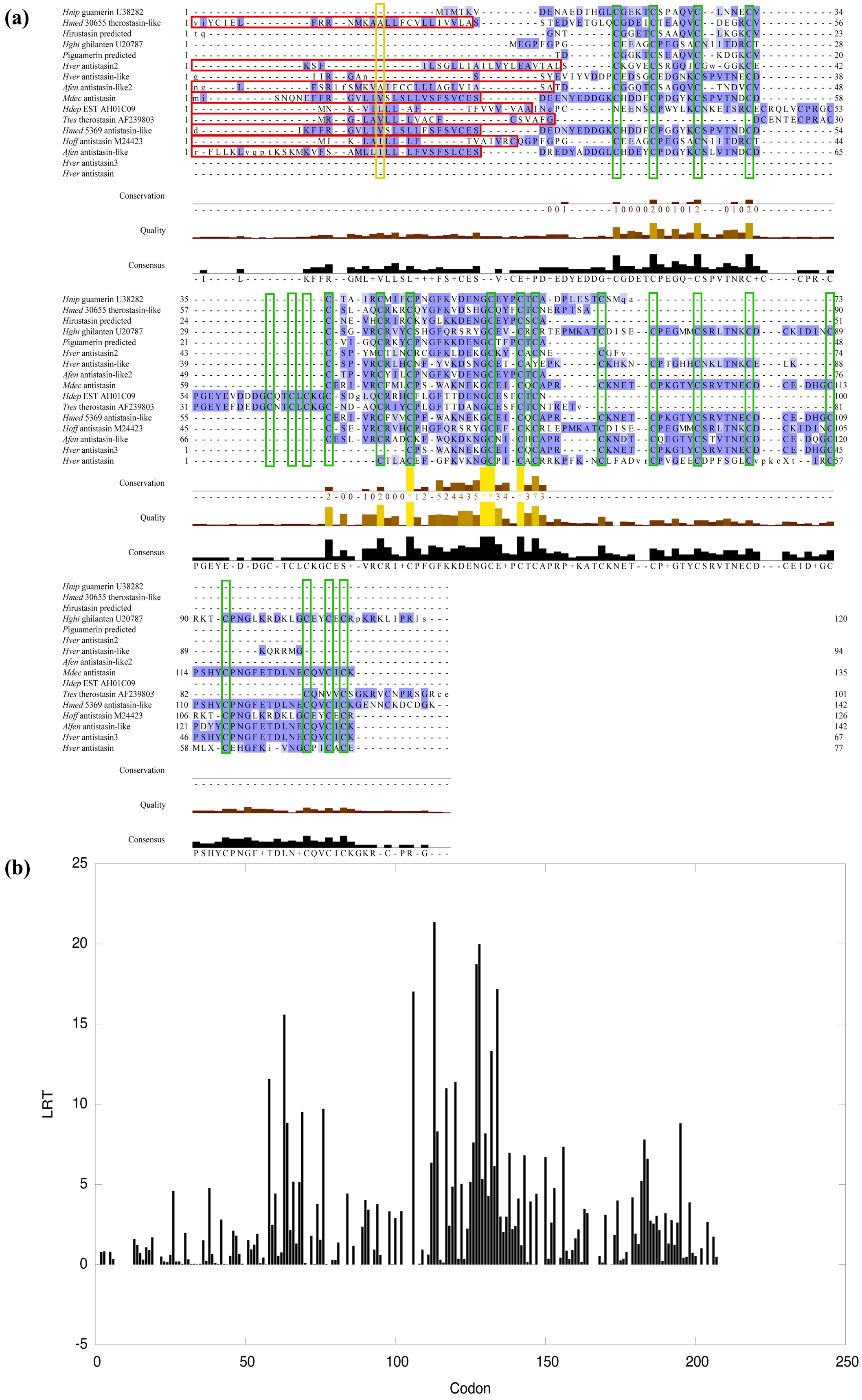
Complete data sets for each of the anticoagulants were subjected to selection pressure analyses. Most likely due to the low taxonomic coverage in the data sets, the REL method often predicted positive selection at invariable sites, i.e., sites with full conservation. Therefore, the results from the REL analyses were not considered herein. The results from the PARRIS, FEL and iFEL analyses are presented in Table 4.2. The analyses agreed on relatively high levels of purifying selection across the diversity of the anticoagulants, with only isolated sites showing signs of positive selection. By and large, both the FEL and iFEL analyses resulted in similar amounts of site-specific positive and negative selection (Table 4.2) and the PARRIS analysis showed no evidence of positive selection for any of the alignments.

The antistasin alignment included a single site under positive selection (codon 26 predicted by iFEL; Fig. 4.10a). However, this occurs within the predicted signal peptide region, thus outside of any expected active region (Nutt et al., 1988; Dunwiddie et al., 1989). The LRT-scores (recovered from the FEL analyses) show a conspicuous spike in the middle of the alignment (Fig. 4.10b), in a region that is transitively highly conserved. Besides high LRT scores (>10) for the purifying selection acting on disulphide-bond-forming cysteines, there is also high purifying selection acting on a glycine (Gly) residue at codon 127, immediately preceding a fully conserved cysteine.

Table 4.2. Number of sites found to be under positive or purifying selection in the FEL, iFEL, REL and PARRIS analyses ($P \leq 0.05$ in all cases).

Locus	#Aligned codons	Significant sites at $P \leq 0.05$					
		Positive selection			Purifying selection		
		FEL	iFEL	PARRIS	FEL	iFEL	PARRIS
Antistasin	209	1	0	0	44	30	N/A
Bdellin	80	0	1	0	18	16	N/A
Decorsin	70	0	0	0	2	2	N/A
Destabilase	151	1	0	0	32	27	N/A
EglinC	128	0	0	0	23	1	N/A
Ficolin	216	1	4	0	44	8	N/A
Hirudin	155	0	0	0	13	3	N/A
Manillase	419	1	3	0	40	20	N/A
Saratin	223	2	2	0	22	24	N/A

Fig. 4.10 Alignment of inferred amino acid sequences for antistasin-family putative orthologues from *Aliolimnatis fenestrate, Hirudo verbana* and *Macrobdella decora* together with the archetypal anticoagulants. (a) The full alignment of orthologues across the known taxonomic diversity. Red boxes denote the predicted signal peptide regions, green boxes denote fully conserved cysteines and yellow boxes denote sites predicted to be under positive selection by iFEL. Shading intensity corresponds to BLOSUM62 conservation. Afen, *Aliolimnatis fenetrata*; Hver, *Hirudo verbana*; Mdec, *Macrobdella decora*; Hdep, *Haementeria depressa*; Hoff, *Haementeria officinalis*; Hghi, *Haementeria ghilianii*; Hnip, *Hirudo nipponia*; Hmed, *Hirudo medicinalis*; Ttes, *Theromyzon tessulatum*. (b) Likelihood ratio test (LRT) scores for selection pressures at each site, plotted against codon position.


(a)
Hnip guamerin U38282
Hmed 30655 therostasin-like
Hirustasin predicted
Hghi ghilanten U20787
Piguamerin predicted
Hver antistasin2
Hver antistasin-like
Afen antistasin-like2
Mdec antistasin
Hdep EST AH01C09
Ttes therostasin AF239803
Hmed 5369 antistasin-like
Hoff antistasin M24423
Afen antistasin-like
Hver antistasin3
Hver antistasin
Conservation
Quality
Consensus
- I - - - L - - - - - - - - KFFR - - GML+VLLSL+++FS+CES - - V - CE+PD+EDYEDDG+CGDETCPEGQ+CSPVTNRC+C - - - - CPR - C
PGEYE - D - DGC - TCLCKGCES+ - VRCRI+CPFGFKKDENGCE+PCTCAPRP+KATCKNET - - CP+GTYCSRVTNECD - - - CEID+GC
Alfen antistasin-like
PSHYCPNGF+TDLN+CQVCICKGKR - C - PR - G - - -
(b)
LRT
Codon
25
20
15
10
5
0
-5
0
50
100
150
200
250

Three different codons in the saratin/LAPP alignment exhibited signs of positive selection: one by FEL (codon 147), one by iFEL (codon 121), and one predicted by both methods (codon 100) (Fig. 4.11a). The principle collagen-binding sites of saratin have been defined (Gronwald et al., 2008) and, interestingly, one of them (Tyr_{42}) is predicted to be under positive selection in the current alignment (this position is also involved in high exchange contributions to conformational motions; Gronwald et al., 2008). Interestingly, the three positively selected codons occur in regions with otherwise high prevalence of negative selection and with high accompanying LRT scores (Fig. 4.11b). There are particularly high levels of purifying selection in the second domain of the molecule, between codons 98-147, including two almost fully conserved lysines (Lys) between the second and third cysteines (Cys).

For hirudin, no positive selection was inferred from any of the three methods. In contrast, LRT scores are high for the prediction of purifying selection across the molecule (Fig. 4.12a), with two sites (codon 52 [Cys] and codon 53 [Glu/Gly]) showing LRT scores above 6 (Fig. 4.12b). Six cysteines, involved in three disulphide-bonds are conserved across the alignment, solidifying previous findings of the disulphide-folding pathway of hirudin (Chatrenet and Chang, 1993; Min et al., 2010).

In the bdellin alignment, orthologues show high amino acid conservation compared to the known sequence of the anticoagulant (AF223972), especially between codons 39-80 (corresponding to amino acid positions 23-59 in the archetypal anticoagulant; Fig. 4.13a). This includes full conservation of six cysteines, presumably involved in three disulfide-bonds. Positive selection was only predicted for one site (codon 26; FEL) and this was positioned inside the predicted signal peptide region. The

Fig. 4.11 Alignment of inferred amino acid sequences for saratin putative orthologues from *Aliolimnatis fenestrata* and *Macrobdella decora* together with the archetypal anticoagulants. (a) The full alignment of orthologues across the known taxonomic diversity. Red boxes denote the predicted signal peptide regions, green boxes denote fully conserved cysteines, black boxes denote sites predicted to be under positive selection by FEL and yellow boxes denote sites predicted to be under positive selection by iFEL. Shading intensity corresponds to BLOSUM62 conservation. Afen, *Aliolimnatis fenetrata*; Hver, *Hirudo verbana*; Mdec, *Macrobdella decora*; Hoff, *Haementeria officinalis*; Hrob, *Helobdella robusta*; Hmed, *Hirudo medicinalis*. (b) Likelihood ratio test (LRT) scores for selection pressures at each site, plotted against codon position.

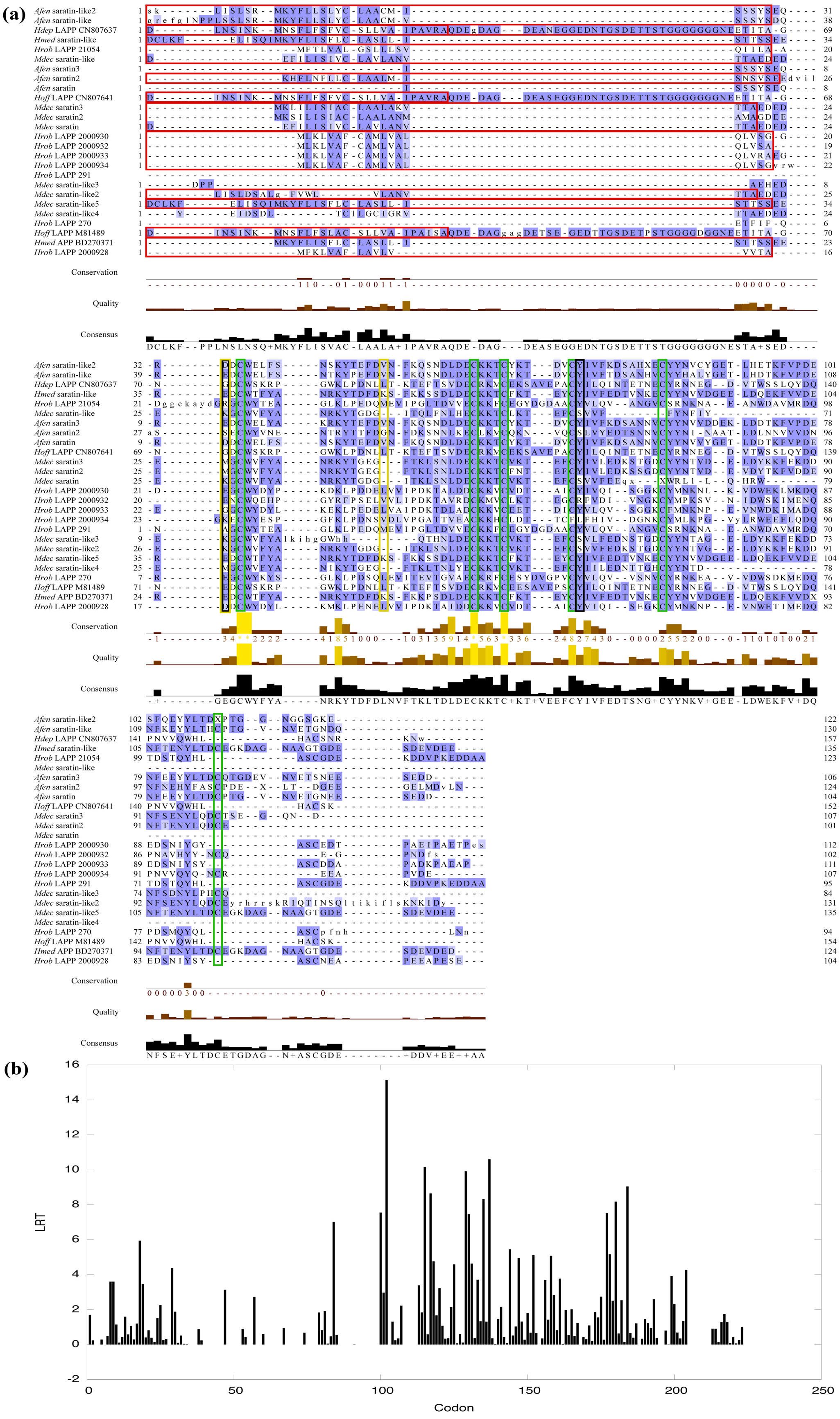

(a)
Afen saratin-like2
Afen saratin-like
Hdep LAPP CN807637
Hmed saratin-like
Hrob LAPP 21054
Mdec saratin-like
Afen saratin3
Afen saratin2
Afen saratin
Hoff LAPP CN807641
Mdec saratin3
Mdec saratin2
Mdec saratin
Hrob LAPP 2000930
Hrob LAPP 2000932
Hrob LAPP 2000933
Hrob LAPP 2000934
Hrob LAPP 291
Mdec saratin-like3
Mdec saratin-like2
Mdec saratin-like5
Mdec saratin-like4
Hrob LAPP 270
Hoff LAPP M81489
Hmed APP BD270371
Hrob LAPP 2000928
Conservation
Quality
Consensus
DCLKF--PPLNSLNSQ+MKYFLISVAC-LAALA+IPAVRAQDE-DAG---DEASEGGEDNTGSDETTSTGGGGGGGNESTA+SED----
-+------GEGCWYFYA-----NRKYTDFDLNVFTKLTDLDECKKTC+KT+VEEFCYIVFEDTSNG+CYYNKV+GEE-LDWEKFV+DQ
NFSE+YLTDCETGDAG--N+ASCGDE--------+DDV+EE++AA
(b)
LRT
Codon
16
14
12
10
8
6
4
2
0
-2
0
50
100
150
200
250

Fig. 4.12 Alignment of inferred amino acid sequences for hirudin putative orthologues from *Aliolimnatis fenestrata* and *Macrobdella decora* together with the archetypal anticoagulants. (a) The full alignment of orthologues across the known taxonomic diversity. Red boxes denote the predicted signal peptide region and green boxes denote fully conserved cysteines. Shading intensity corresponds to BLOSUM62 conservation. Afen, *Aliolimnatis fenetrata*; Mdec, *Macrobdella decora*; Hmed, *Hirudo medicinalis*; Hsyl, *Hirudinaria sylvestris*. (b) Likelihood ratio test (LRT) scores for selection pressures at each site, plotted against codon position.

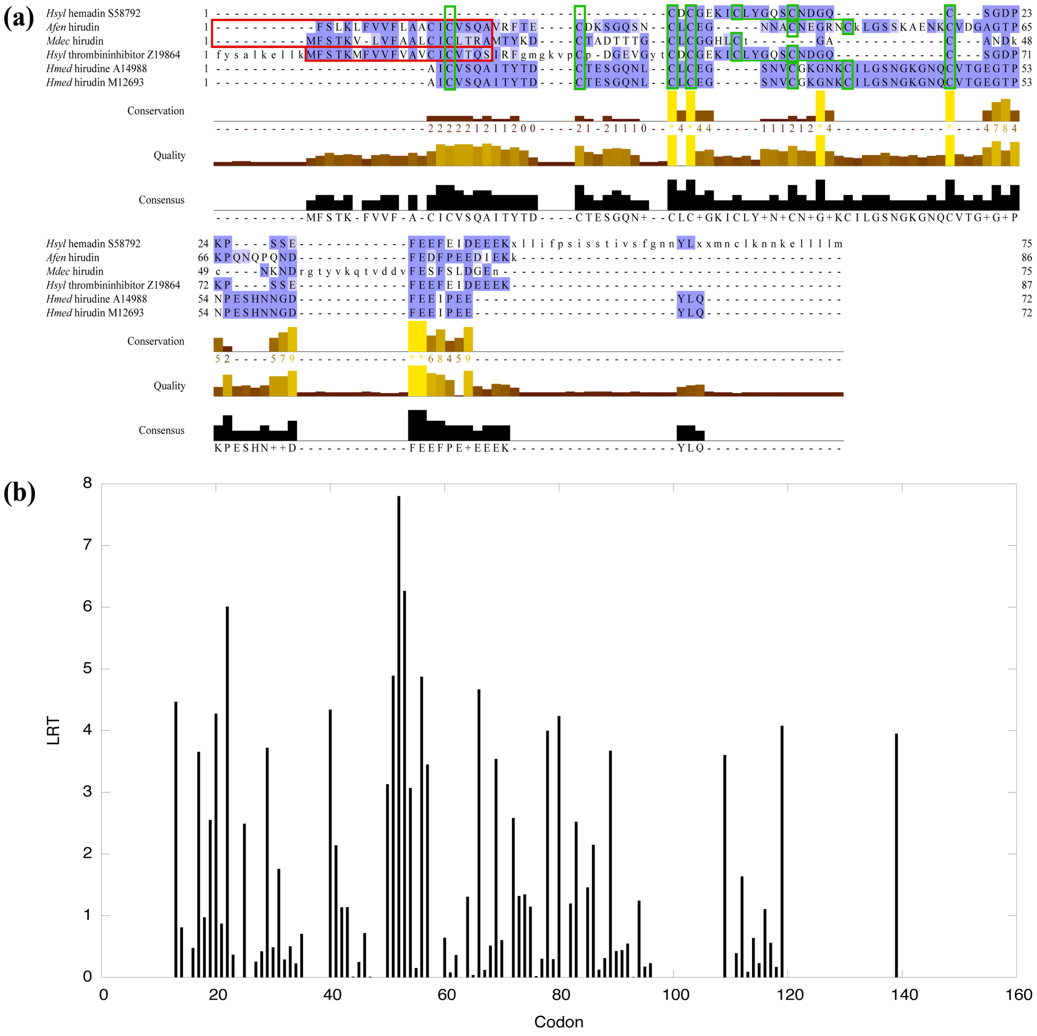
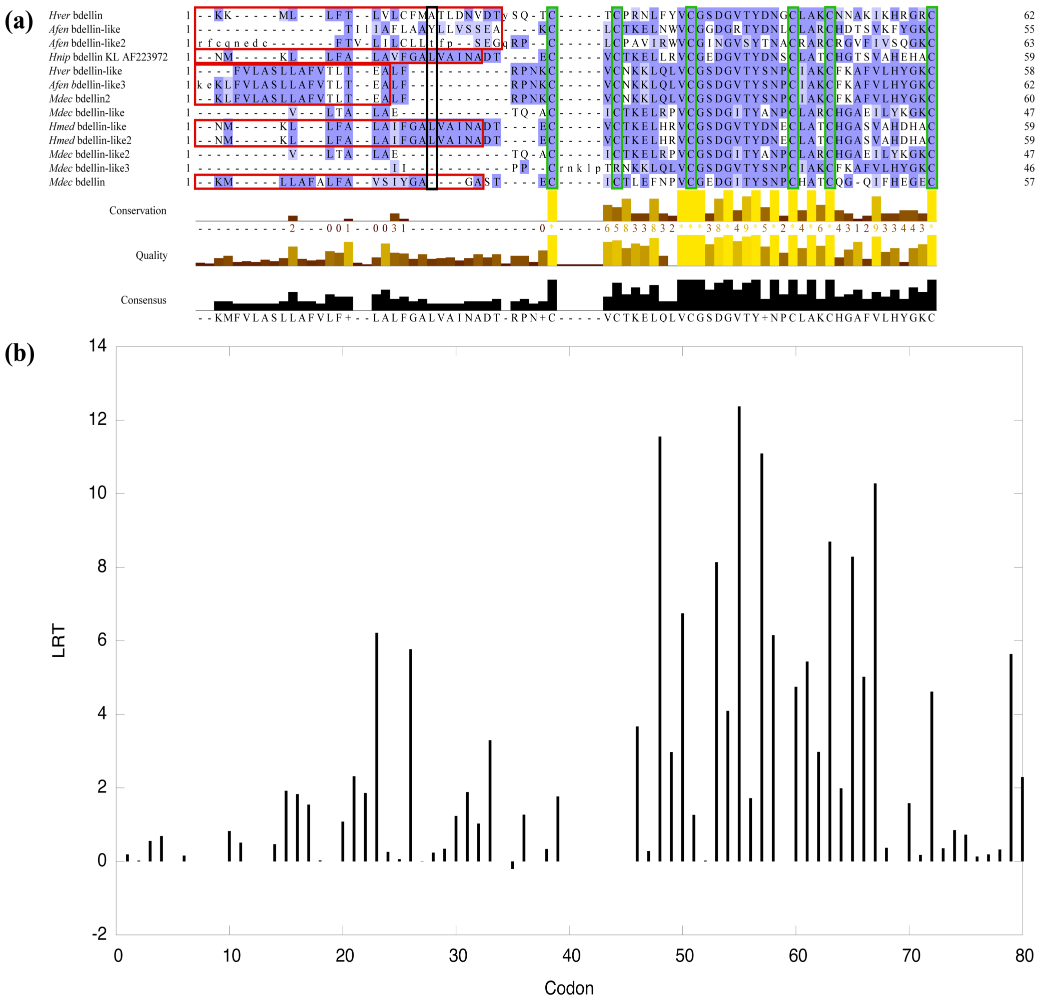

(a)
Hsyl hemadin S58792 1 ------------------------------------CDCGEKICLYGQSCNDGQ--------------C---SGDP 23
Afen hirudin 1 -----------FSLKLFVVFLAACICVSQAVRFTE----CDKSGQSN--CLCEG-----NNACNEGRNCkLGSSKAENKCVDGAGTP 65
Mdec hirudin 1 ----------MFSTKV-LVFAALCICLTRAMTYKD----CTADTTTG--CLCGGHLCt-------GA------------C---ANDk 48
Hsyl thrombininhibitor Z19864 1 fysalkellkMFSTKMFVVFVAVCICVTQSIRFgmgkvpCp-DGEVGytCDCGEKICLYGQSCNDGQ-------------C---SGDP 71
Hmed hirudine A14988 1 ---------------------AICVSQAITYTD----CTESGQNL--CLCEG-----SNVCGKGNKCILGSNGKGNQCVTGEGTP 53
Hmed hirudin M12693 1 ---------------------AICVSQAITYTD----CTESGQNL--CLCEG-----SNVCGKGNKCILGSNGKGNQCVTGEGTP 53
Conservation
----------------------2222212112 00----21-21110--*4*44-----111212*4------------*---4784
Quality
Consensus
----------MFSTK-FVVF-A-CICVSQAITYTD----CTESGQN+--CLC+GKICLY+N+CN+G+KCILGSNGKGNQCVTG+G+P
Hsyl hemadin S58792 24 KP----SSE-------------FEEFEIDEEEKxllifpsisstivsfgnnYLxxmnclknnkellllm 75
Afen hirudin 66 KPQNQPQND-------------FEDFPEEDIEKk----------------------------------- 86
Mdec hirudin 49 c----NKNDrgtyvkqtvddvFESFSLDGEn------------------------------------ 75
Hsyl thrombininhibitor Z19864 72 KP----SSE-------------FEEFEIDEEEK------------------------------------ 87
Hmed hirudine A14988 54 NPESHNNGD-------------FEEIPEE-------------------YLQ----------------- 72
Hmed hirudin M12693 54 NPESHNNGD-------------FEEIPEE-------------------YLQ----------------- 72
Conservation
52----579-------------**68459---
Quality
Consensus
KPESHN++D-------------FEEFPE+EEEK-----------------YLQ-----------------
(b)
LRT
Codon
0
1
2
3
4
5
6
7
8
0
20
40
60
80
100
120
140
160

Fig. 4.13 Alignment of inferred amino acid sequences for bdellin putative orthologues from *Aliolimnatis fenestrata, Hirudo verbana* and *Macrobdella decora* together with the archetypal anticoagulant. (a) The full alignment of orthologues across the known taxonomic diversity. Red boxes denote the predicted signal peptide region, green boxes denote fully conserved cysteines and black boxes denote sites predicted to be under positive selection by FEL. Shading intensity corresponds to BLOSUM62 conservation. Afen, *Aliolimnatis fenetrata*; Hver, *Hirudo verbana*; Mdec, *Macrobdella decora*; Hnip, *Hirudo nipponia*; Hmed, *Hirudo medicinalis*. (b) Likelihood ratio test (LRT) scores for selection pressures at each site, plotted against codon position.


(a)
Hver bdellin 1 --KK-----ML---LFT--LVLCFMATLDNVDTySQ-TC-----TCPRNLFYVCGSDGVTYDNGCLAKCNNAKIKHRGRC 62
Afen bdellin-like 1 ---------------TIlIAFLAAYLLVSSEA----KC-----LCTKELNWVCGGDGRTYDNLCLARCHDTSVKFYGKC 55
Afen bdellin-like2 1 rfcqnedc-------FTV-LILCLLtfp--SEGqRP--C-----LCPAVIRWVCGINGVSYTNACRARCRGVFIVSQGKC 63
Hnip bdellin KL AF223972 1 --NM-----KL---LFA--LAVFGALVAINADT----EC-----VCTKELLRVCGEDGVTYDNSCLATCHGTSVAHEHAC 59
Hver bdellin-like 1 ----FVLASLLAFVTLT--EALF-----------RPNKC-----VCNKKLQLVCGSDGVTYSNPCIAKCFKAFVLHYGKC 58
Afen bdellin-like3 1 keKLFVLASLLAFVTLT--EALF-----------RPNKC-----VCNKKLQLVCGSDGVTYSNPCIAKCFKAFVLHYGKC 62
Mdec bdellin2 1 --KLFVLASLLAFVTLT--EALF-----------RPNKC-----VCNKKLQLVCGSDGVTYSNPCIAKCFKAFVLHYGKC 60
Mdec bdellin-like 1 ----------V---LTA--LAE------------TQ-AC-----ICTKELRPVCGSDGITYANPCLARCHGAEILYKGKC 47
Hmed bdellin-like 1 --NM-----KL---LFA--LAIFGALVAINADT----EC-----VCTKELHRVCGSDGVTYDNECLATCHGASVAHDHAC 59
Hmed bdellin-like2 1 --NM-----KL---LFA--LAIFGALVAINADT----EC-----VCTKELHRVCGSDGVTYDNECLATCHGASVAHDHAC 59
Mdec bdellin-like2 1 ----------V---LTA--LAE------------TQ-AC-----ICTKELRPVCGSDGITYANPCLARCHGAEILYKGKC 47
Mdec bdellin-like3 1 ---------------------II-----------PP--CrnklpTRNKKLQLVCGSDGVTYSNPCIAKCFKAFVLHYGKC 46
Mdec bdellin 1 --KM-----LLAFALFA--VSIYGA----GAST----EC-----ICTLEFNPVCGEDGITYSNPCHATCQG-QIFHEGEC 57
Conservation
----------2---001--0031---------------0*-----65833832***38*49*5*2*4*6*4312933443*
Quality
Consensus
--KMFVLASLLAFVLF+--LALFGALVAINADT-RPN+C-----VCTKELQLVCGSDGVTY+NPCLAKCHGAFVLHYGKC
(b)
14
12
10
8
6
4
2
0
-2
LRT
0
10
20
30
40
50
60
70
80
Codon

highest LRT scores (Fig. 4.13b) were retrieved for codon 48 (Lys) and codon 55 (Gly). The identification of two proline (Pro) residues in bdellin from *M. decora* by Min et al. (2010), while surprising (see Fritz et al., 1971), is corroborated by equivalent residues in all variants of orthologues in both *H. verbana* and *A. fenestrata.*

For the site-specific analyses of the manillase alignment, both FEL and iFEL predicted positive selection at codon 358 (Asn), and iFEL alone predicted positive selection at codons 14 (Ala) and 370 (Ser) (Fig. 4.14a). High conservation occurs throughout the alignment; two fully conserved leucines (Leu) at codons 257 and 265 give rise to the two LRT scores above 10 (Fig. 4.14b).

There is no evidence of positive selection acting on the residues in the decorsin data set (Fig. 4.15a), which included orthologues from *Macrobdella decora* as well as the known decorsin sequence. In addition, whereas the entire alignment consists of almost fully conserved residues with LRT scores >3 (Fig. 4.15b), FEL and iFEL estimated only three of these as being under purifying selection: codons 51 (Pro), 60 (Arg) and 67 (Cys); both methods agree on the latter residue. The six cysteine residues are fully conserved.

A single site in the destabilase alignment was estimated to be under positive selection by FEL (codon 30 [Ser]). However, as with both antistasin and bdellin, this codon is within the signal peptide region, thus not in any region of the mature peptide (Fig. 4.16a). A disproportional spike in LRT score (>15), as compared to the remaining alignment, is calculated for the proline (Pro) present at codon 68 (Fig. 4.16b). Very high sequence conservation is present in the mature peptide (i.e., beyond the signal peptide

Fig. 4.14 Alignment of inferred amino acid sequences for heparanase-class endoglucuronidase (manillase) putative orthologues from *Aliolimnatis fenestrata, Hirudo verbana* and *Macrobdella decora* together with the archetypal anticoagulant. (a) The full alignment of orthologues across the known taxonomic diversity. Black boxes denote sites predicted to be under positive selection by FEL and yellow boxes denote sites predicted to be under positive selection by iFEL. Shading intensity corresponds to BLOSUM62 conservation. Afen, *Aliolimnatis fenetrata*; Hver, *Hirudo verbana*; Mdec, *Macrobdella decora*; Hmed, *Hirudo medicinalis*. (b) Likelihood ratio test (LRT) scores for selection pressures at each site, plotted against codon position.

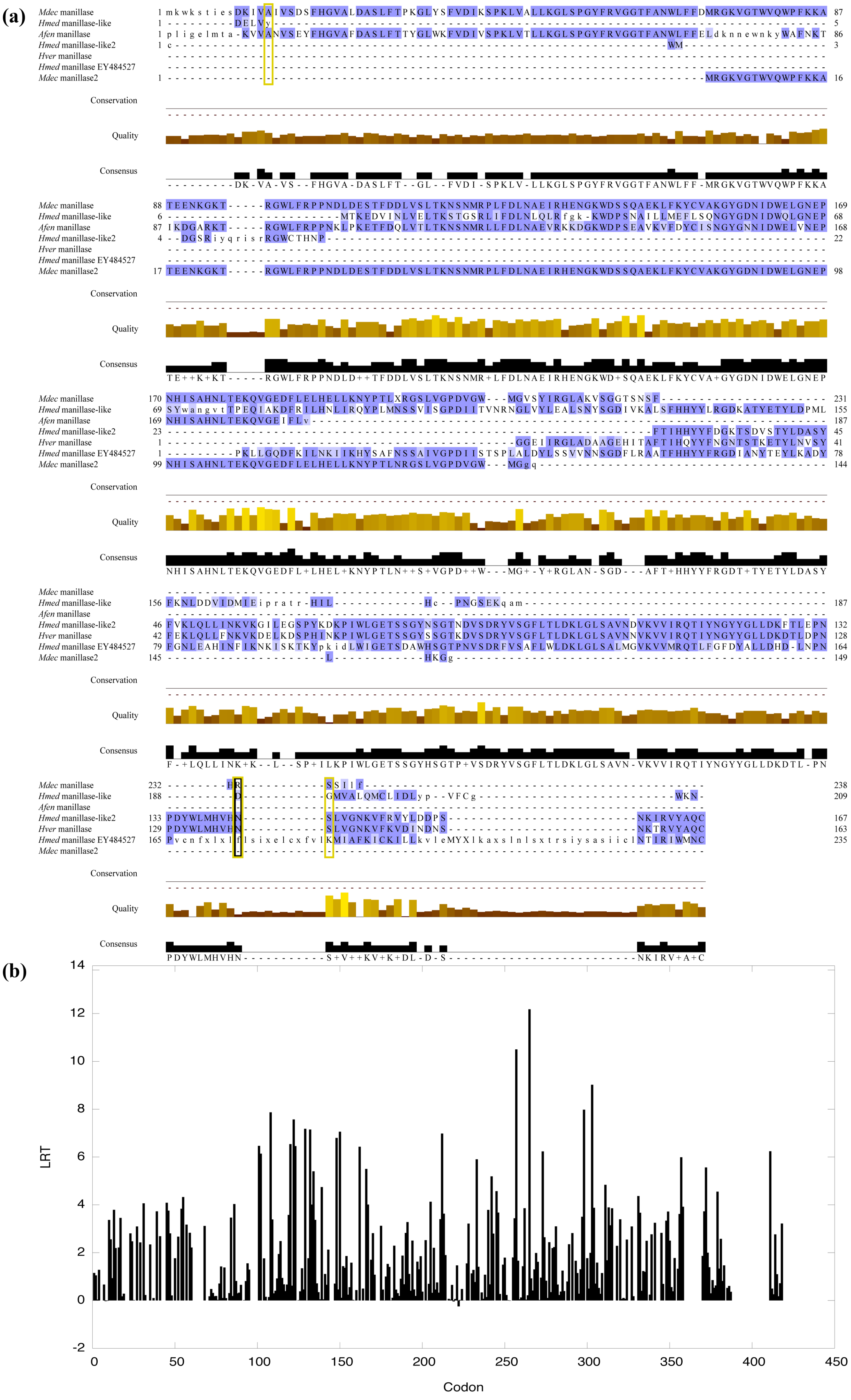

(a)
Mdec manillase
Hmed manillase-like
Afen manillase
Hmed manillase-like2
Hver manillase
Hmed manillase EY484527
Mdec manillase2
Conservation
Quality
Consensus
(b)
LRT
Codon
-2
0
2
4
6
8
10
12
14
0
50
100
150
200
250
300
350
400
450

Fig. 4.15 Alignment of inferred amino acid sequences for decorsin putative orthologues from *Macrobdella decora* together with the archetypal anticoagulants. (a) The full alignment of orthologues across the known taxonomic diversity. Red boxes denote the predicted signal peptide region and green boxes denote fully conserved cysteines. Shading intensity corresponds to BLOSUM62 conservation. Mdec, *Macrobdella decora.* (b) Likelihood ratio test (LRT) scores for selection pressures at each site, plotted against codon position.

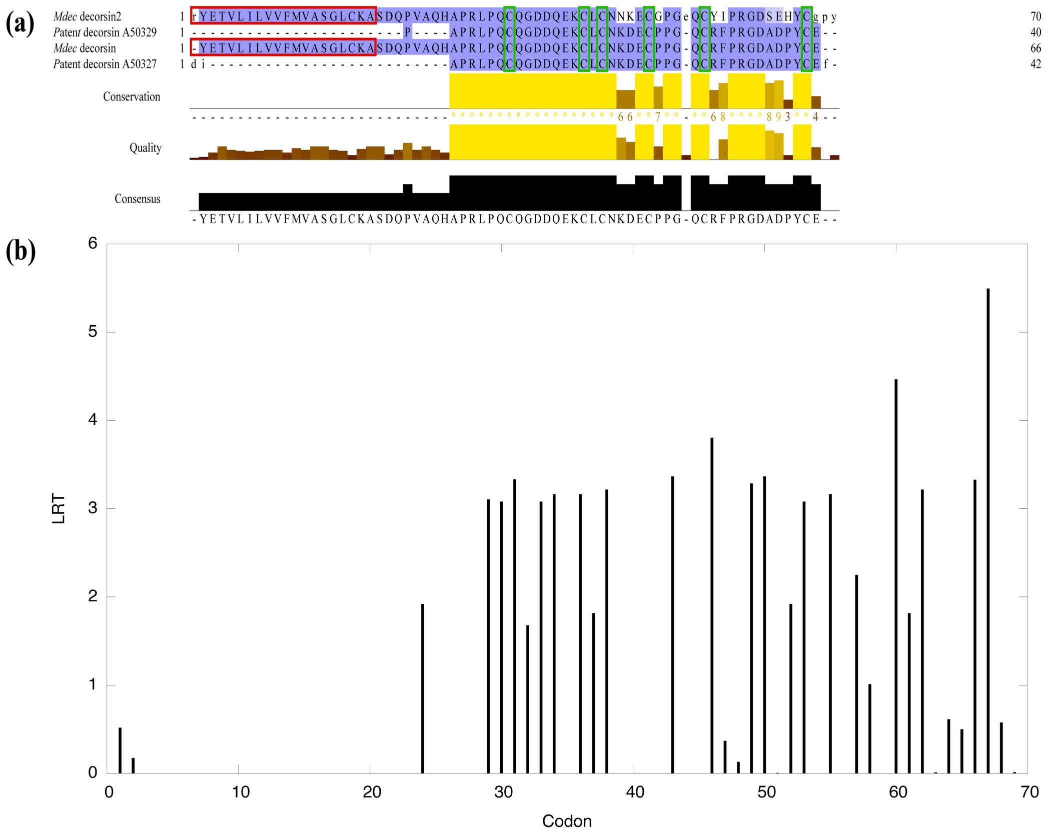
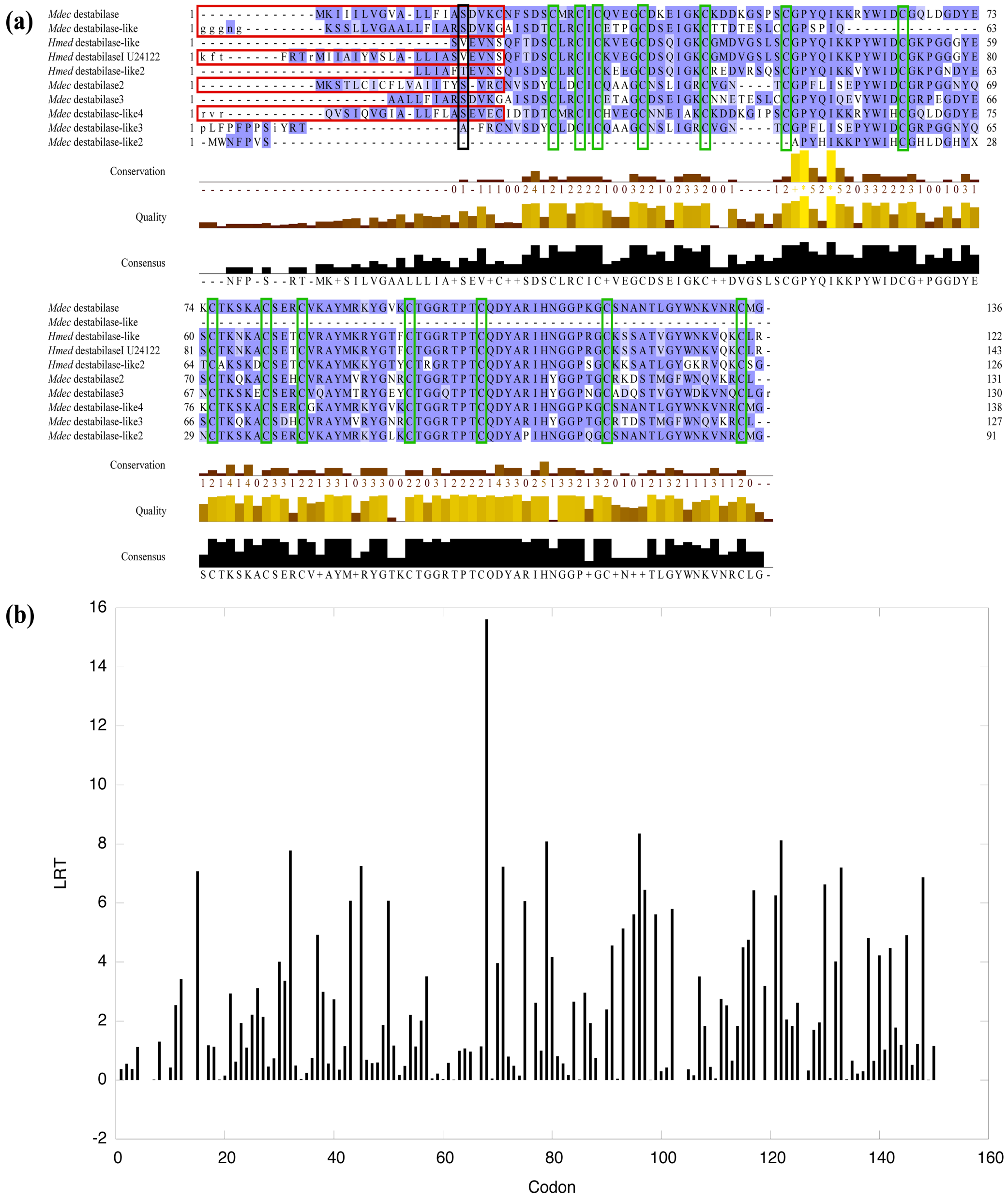

(a)
Mdec decorsin2 1 rYETVLILVVFMVASGLCKASDQPVAQHAPRLPQCQGDDQEKCLCNNKECGPGeQCYIPRGDSEHYCgpy 70
Patent decorsin A50329 1 ----------------------P----APRLPQCQGDDQEKCLCNKDECPPG-QCRFPRGDADPYCE-- 40
Mdec decorsin 1 -YETVLILVVFMVASGLCKASDQPVAQHAPRLPQCQGDDQEKCLCNKDECPPG-QCRFPRGDADPYCE-- 66
Patent decorsin A50327 1 di--------------------------APRLPQCQGDDQEKCLCNKDECPPG-QCRFPRGDADPYCEf- 42
Conservation
----------------------------*****************66**7**-**68****893**4--
Quality
Consensus
-YETVLILVVFMVASGLCKASDQPVAQHAPRLPQCQGDDQEKCLCNKDECPPG-QCRFPRGDADPYCE--
(b)
LRT
0
1
2
3
4
5
6
Codon
0
10
20
30
40
50
60
70

Fig. 4.16 Alignment of inferred amino acid sequences for destabilase putative orthologues from *Macrobdella decora* together with the archetypal anticoagulant. (a) The full alignment of orthologues across the known taxonomic diversity. Red boxes denote the predicted signal peptide region, green boxes denote fully conserved cysteines and black boxes denote sites predicted to be under positive selection by FEL. Shading intensity corresponds to BLOSUM62 conservation. Mdec, *Macrobdella decora*; Hmed, *Hirudo medicinalis*. (b) Likelihood ratio test (LRT) scores for selection pressures at each site, plotted against codon position.


(a)
Mdec destabilase 1 -------------MKIILVGVA-LLFIASDVKCNFSDSCMRCICQVEGCDKEIGKCKDDKGSPSCGPYQIKKRYWIDCGQLDGDYE 73
Mdec destabilase-like 1 gggng---------KSSLLVGAALLFIARSDVKGAISDTCLRCICETPGCDSEIGKCTTDTESLCCGPSPIQ----------------- 63
Hmed destabilase-like 1 -------------------------SVEVNSQFTDSCLRCICKVEGCDSQIGKCGMDVGSLSCGPYQIKKPYWIDCGKPGGGYE 59
Hmed destabilaseI U24122 1 kft------FRTrMIIAIYVSLA-LLIASVEVNSQFTDSCLRCICKVEGCDSQIGKCGMDVGSLSCGPYQIKKPYWIDCGKPGGGYE 80
Hmed destabilase-like2 1 -----------------------LLIAFTEVNSQISDSCLRCICKEEGCDSQIGKCREDVRSQSCGPYQIKKVYWIDCGKPGNDYE 63
Mdec destabilase2 1 -------------MKSTLCICFLVAIITYS-VRCNVSDYCLDCICQAAGCNSLIGRCVGN----TCGPFLISEPYWIDCGRPGGNYQ 69
Mdec destabilase3 1 ---------------------AALLFIARSDVKGAISDSCLRCICETAGCDSEIGKCNNETESLCCGPYQIQEVYWIDCGRPEGDYE 66
Mdec destabilase-like4 1 rvr-----------QVSIQVGIA-LLFLASEVECIDTDTCMRCICHVEGCNNEIAKCKDDKGIPSCGPYQIKKRYWIHCGQLDGDYE 75
Mdec destabilase-like3 1 pLFPFPPSiYRT-----------------A-FRCNVSDYCLDCICQAAGCNSLIGRCVGN----TCGPFLISEPYWIDCGRPGGNYQ 65
Mdec destabilase-like2 1 -MWNFPVS---APYHIKKPYWIHCGHLDGHYX 28
Conservation
---------------------------01-11100241212222100322102332001----12+*52*52033222310010 31
Quality
Consensus
---NFP-S--RT-MK+SILVGAALLLIA+SEV+C++SDSCLRCIC+VEGCDSEIGKC++DVGSLSCGPYQIKKPYWIDCG+PGGDYE
Mdec destabilase 74 KCTKSKACSERCVKAYMRKYGVKCTGGRTPTCQDYARIHNGGPKGCSNANTLGYWNKVNRCMG- 136
Mdec destabilase-like --
Hmed destabilase-like 60 SCTKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCKSSATVGYWNKVQKCLR- 122
Hmed destabilaseI U24122 81 SCTKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCKSSATVGYWNKVQKCLR- 143
Hmed destabilase-like2 64 TCAKSKDCSETCVKAYMKKYGTYCTRGRTPTCQDYARIHNGGPSGCKKSATLGYGKRVQKCSG- 126
Mdec destabilase2 70 SCTKQKACSEHCVRAYMVRYGNRCTGGRTPTCQDYARIHYGGPTGCRKDSTMGFWNQVKRCL-- 131
Mdec destabilase3 67 NCTKSKECSERCVQAYMTRYGEYCTGGQTPTCQDYARIHYGGPNGCADQSTVGYWDKVNQCLGr 130
Mdec destabilase-like4 76 KCTKSKACSERCGKAYMRKYGVKCTGGRTPTCQDYARIHNGGPKGCSNANTLGYWNKVNRCMG- 138
Mdec destabilase-like3 66 SCTKQKACSDHCVRAYMVRYGNRCTGGRTPTCQDYARIHYGGPTGCRTDSTMGFWNQVKRCL-- 127
Mdec destabilase-like2 29 NCTKSKACSERCVKAYMRKYGLKCTGGRTPTCQDYAPIHNGGPQGCSNANTLGYWNKVNRCMG- 91
Conservation
121414023312213310333002203122221433025133213201012132111311 20--
Quality
Consensus
SCTKSKACSERCV+AYM+RYGTKCTGGRTPTCQDYARIHNGGP+GC+N++TLGYWNKVNRCLG-
(b)
LRT
Codon
16
14
12
10
8
6
4
2
0
-2
0
20
40
60
80
100
120
140
160

region). In addition, the number of conserved cysteines (N=14) across the alignment agrees perfectly with previous findings in destabilase (Min et al., 2010).

For the ficolin alignment (Fig. 4.17a), iFEL predicts positive selection for four codons (3, 92, 135 and 165) whereas FEL predicts equivalent selection for only a single codon (146). Interestingly, both codon 135 and 165 occur in regions with otherwise high levels of purifying selection and high accompanying LRT scores (Fig. 4.17b); for both codons, the adjacent sites are fully conserved, codon 136 displaying the second highest LRT score (>12) across the alignment. Two cysteines show full conservation in the alignment; the tertiary structure for leech-derived ficolin has yet to be determined but it is likely that these cysteines form a disulphide-bridge.

The alignment of the eglin c-like orthologues shows high conservation, in particular between codons 38-101 (Fig. 4.18a) and FEL predicts 23 sites as being under purifying selection. The average LRT score across the alignment is 1.58 (Fig. 4.18b), the highest (8) occurring at codon 51 (Arg). No sites are estimated to be under positive selection by any of the methods and our results agree with previous studies (Min et al., 2010) in that no cysteines are present in the eglin c molecule.

Discussion

Our annotations of two newly prepared salivary transcriptome EST libraries from the European medicinal leech *Hirudo verbana* and the African medicinal leech *Aliolimnatis fenestrata* show that these leeches each produce pharmacological cocktails with similar anticoagulant-diversity to that of the North American medicinal leech *Macrobdella decora*, with the exception of decorsin and destabilase. In total, the *A.*

Fig. 4.17 Alignment of inferred amino acid sequences for ficolin putative orthologues from *Aliolimnatis fenestrata* and *Macrobdella decora* together with the archetypal anticoagulant. (a) The full alignment of orthologues across the known taxonomic diversity. Green boxes denote fully conserved cysteines, black boxes denote sites predicted to be under positive selection by FEL and yellow boxes denote sites predicted to be under positive selection by iFEL. Shading intensity corresponds to BLOSUM62 conservation. Mdec, *Macrobdella decora*; Afen, *Aliolimnatis fenetrata*; Hmed, *Hirudo medicinalis*. (b) Likelihood ratio test (LRT) scores for selection pressures at each site, plotted against codon position.

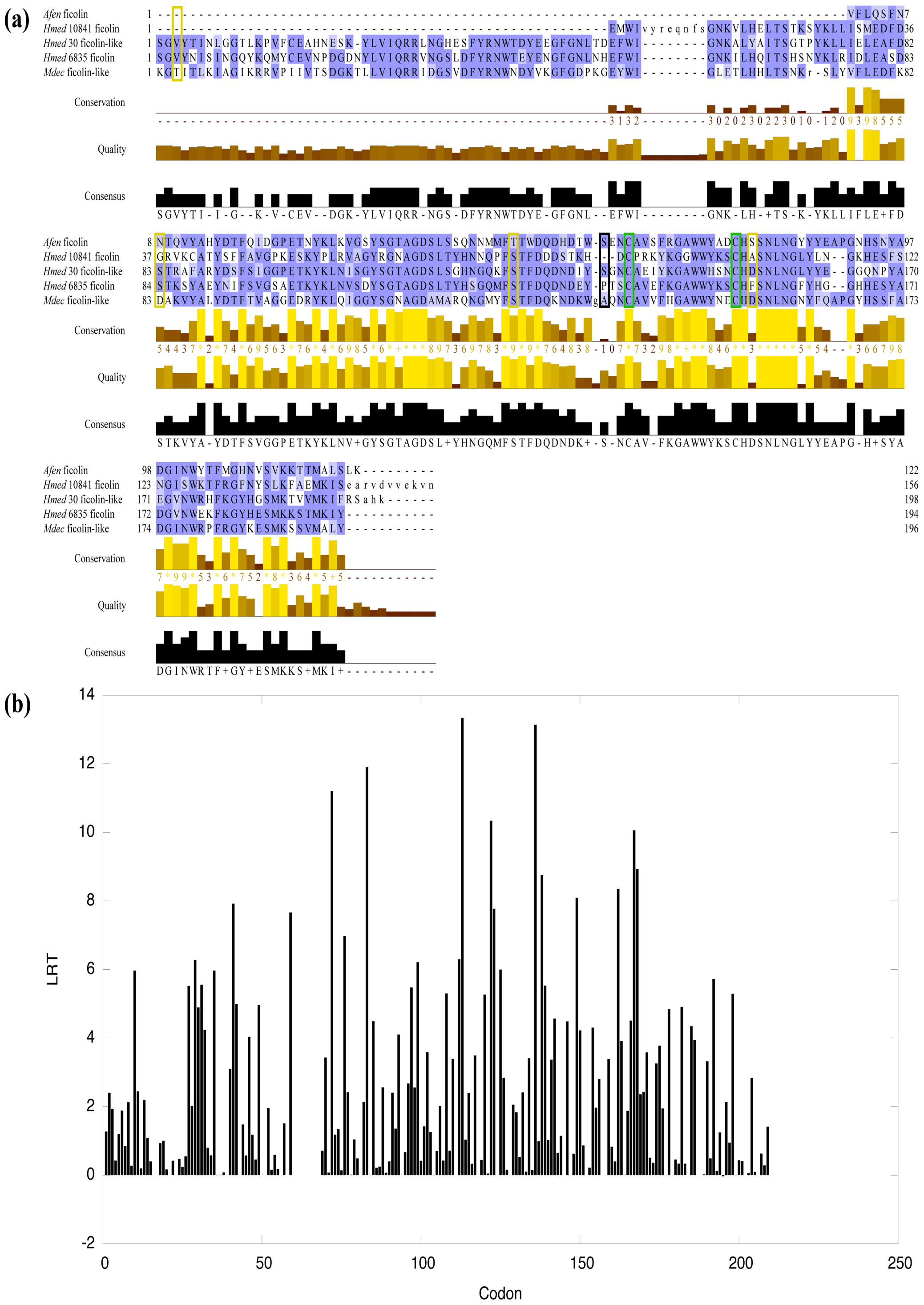
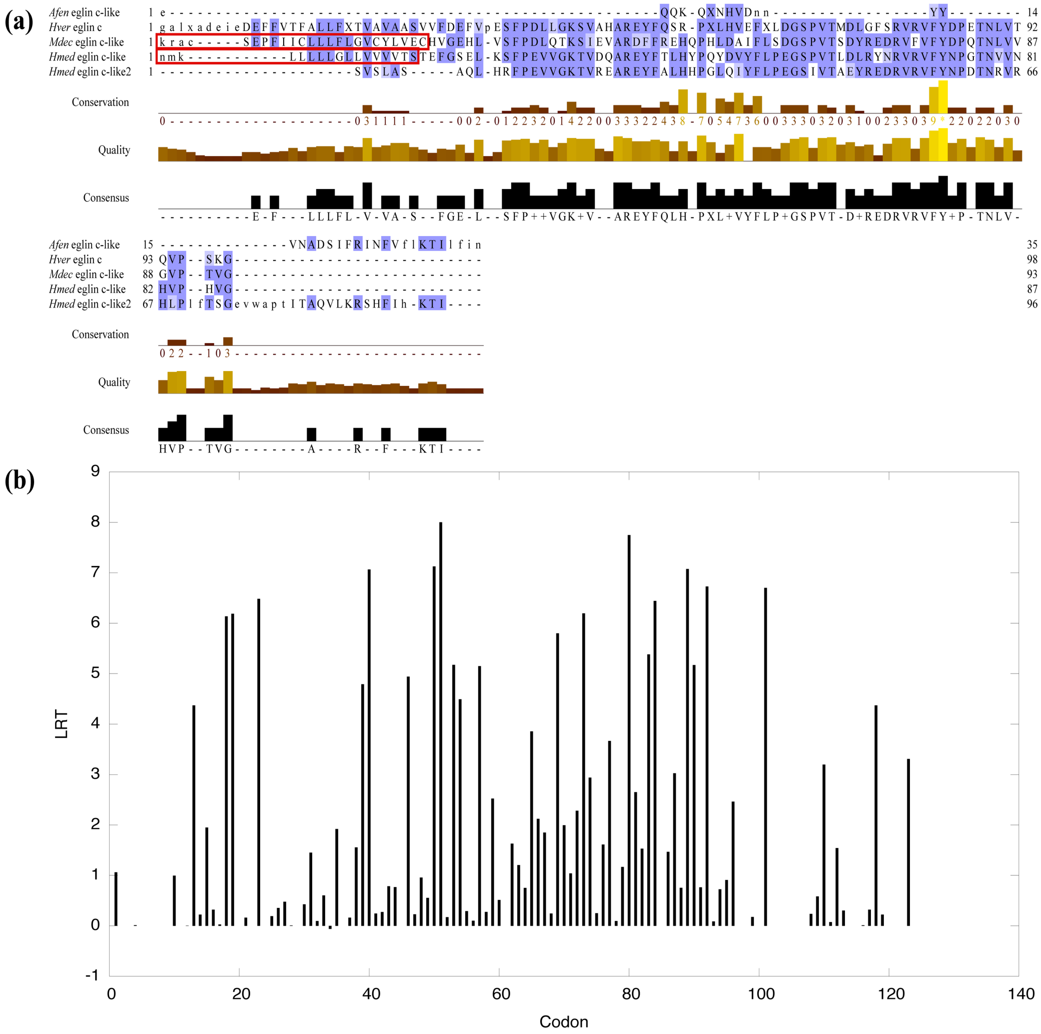

(a)
Afen ficolin 1 ---VFLQSFN7
Hmed 10841 ficolin 1 --EMWIvyreqnfsGNKVLHELTSTKSYKLLISMEDFD36
Hmed 30 ficolin-like 1 SGVYTINLGGTLKPVFCEAHNESK-YLVIQRRLNGHESFYRNWTDYEEGFGNLTDEFWI--------GNKALYAITSGTPYKLLIELEAFD82
Hmed 6835 ficolin 1 SGVYNISINGQYKQMYCEVNPDGDNYLVIQRRVNGSLDFYRNWTEYENGFGNLNHEFWI--------GNKILHQITSHSNYKLRIDLEASD83
Mdec ficolin-like 1 KGTITLKIAGIKRRVPIIVTSDGKTLLVIQRRIDGSVDFYRNWNDYVKGFGDPKGEYWI--------GLETLHHLTSNKr-SLYVFLEDFK82
Conservation
--3132---------302023022301 0-120939 8555
Quality
Consensus
SGVYTI-I-G--K-V-CEV--DGK-YLVIQRR-NGS-DFYRNWTDYE-GFGNL--EFWI--------GNK-LH-+TS-K-YKLLIFLE+FD
Afen ficolin 8 NTQVYAHYDTFQIDGPETNYKLKVGSYSGTAGDSLSSQNNMMFTTWDQDHDTW-SENCAVSFRGAWWYADCHSSNLNGYYYEAPGNHSNYA97
Hmed 10841 ficolin 37 GRVKCATYSFFAVGPKESKYPLRVAGYRGNAGDSLTYHNNQPFSTFDDDSTKH---DCPRKYKGGWWYKSCHASNLNGLYLN--GKHESFS122
Hmed 30 ficolin-like 83 STRAFARYDSFSIGGPETKYKLNISGYSGTAGDSLSGHNGQKFSTFDQDNDIY-SGNCAEIYKGAWWHSNCHDSNLNGLYYE--GGQNPYA170
Hmed 6835 ficolin 84 STKSYAEYNIFSVGSAETKYKLNVSDYSGTAGDSLTYHSGQMFSTFDQDNDEY-PTSCAVEFKGAWWYKSCHFSNLNGFYHG--GHHESYA171
Mdec ficolin-like 83 DAKVYALYDTFTVAGGEDRYKLQIGGYSGNAGDAMARQNGMYFSTFDQKNDKWgAQNCAVVFHGAWWYNECHDSNLNGNYFQAPGYHSSFA173
Conservation
54437*2*74*69563*76*4*6985*6*+***89736978 3*9*9*764838-107*73298*+**846**3*****5*54--*366798
Quality
Consensus
STKVYA-YDTFSVGGPETKYKLNV+GYSGTAGDSL+YHNGQMFSTFDQDNDK+-S-NCAV-FKGAWWYKSCHDSNLNGLYYEAPG-H+SYA
Afen ficolin 98 DGINWYTFMGHNVSVKKTTMALSLK--------- 122
Hmed 10841 ficolin 123 NGISWKTFRGFNYSLKFAEMKISearvdvvekvn 156
Hmed 30 ficolin-like 171 EGVNWRHFKGYHGSMKTVVMKIFRSahk------ 198
Hmed 6835 ficolin 172 DGVNWEKFKGYHESMKKSTMKIY----------- 194
Mdec ficolin-like 174 DGINWRPFRGYKESMKSSVMALY----------- 196
Conservation
7*99*53*6*752*8*364*5+5-----------
Quality
Consensus
DGINWRTF+GY+ESMKKS+MKI+-----------
(b)
LRT
14
12
10
8
6
4
2
0
-2
0
50
100
150
200
250
Codon

Fig. 4.18 Alignment of inferred amino acid sequences for eglin c putative orthologues from *Aliolimnatis fenestrata, Hirudo verbana* and *Macrobdella decora* together with the archetypal anticoagulants. (a) The full alignment of orthologues across the known taxonomic diversity. Red boxes denote the predicted signal peptide region and shading intensity corresponds to BLOSUM62 conservation. Hver, *Hirudo verbana*; Afen, *Aliolimnatis fenetrata*; Mdec, *Macrobdella decora*; Hmed, *Hirudo medicinalis*. (b) Likelihood ratio test (LRT) scores for selection pressures at each site, plotted against codon position.


(a)
Afen eglin c-like 1 e---QQK-QXNHVDnn-----------------YY-------- 14
Hver eglin c 1 galxadeieDEFFVTFALLFXTVAVAASVVFDEFVpESFPDLLGKSVAHAREYFQSR-PXLHVEFXLDGSPVTMDLGFSRVRVFYDPETNLVT 92
Mdec eglin c-like 1 krac-----SEPFIICLLLFLGVCYLVECHVGEHL-VSFPDLQTKSIEVARDFFREHQPHLDAIFLSDGSPVTSDYREDRVFVFYDPQTNLVV 87
Hmed eglin c-like 1 nmk-----------LLLLLGLLVVVVTSTEFGSEL-KSFPEVVGKTVDQAREYFTLHYPQYDVYFLPEGSPVTLDLRYNRVRVFYNPGTNVVN 81
Hmed eglin c-like2 1 ---------------------SVSLAS-----AQL-HRFPEVVGKTVREAREYFALHHPGLQIYFLPEGSIVTAEYREDRVRVFYNPDTNRVR 66
Conservation
0--------------------031111-----002-012232014220033322438-70547360033303203100233039*22022030
Quality
Consensus
---------E-F---LLLFL-V-VA-S--FGE-L--SFP++VGK+V--AREYFQLH-PXL+VYFLP+GSPVT-D+REDRVRVFY+P-TNLV-
Afen eglin c-like 15 --------------VNADSIFRINFVflKTIlfin 35
Hver eglin c 93 QVP--SKG--------------------------- 98
Mdec eglin c-like 88 GVP--TVG--------------------------- 93
Hmed eglin c-like 82 HVP--HVG--------------------------- 87
Hmed eglin c-like2 67 HLPlfTSGevwaptITAQVLKRSHFIh-KTI---- 96
Conservation
022--103---------------------------
Quality
Consensus
HVP--TVG--------A----R--F---KTI----
(b)
LRT
9
8
7
6
5
4
3
2
1
0
-1
0
20
40
60
80
100
120
140
Codon

fenestrata EST library included transcripts that showed BLASTx-orthology with 11 well-characterized leech anticoagulants and several other bioactive peptides, whereas the *H. verbana* library included transcripts orthologous to seven anticoagulants and several other peptides. Expressed salivary proteins in *M. decora* already have been shown to comprise an unprecedented diversity of anticoagulants (Min et al., 2010). The results presented here support the notion that such diversity is general to hirudinoid leeches. In addition, for the three taxa, the frequently predicted signal peptide-regions in orthologous sequences for each of antistasin-family proteins, bdellin, decorsin, destabilase, eglin c, hirudin and saratin/LAPP are indicative of the secretion, and ultimate use, of these anticoagulants by the leeches. In contrast to (e.g.,) the functionally diverse and highly prey-specific snake venoms, which commonly evolve under positive selection (Heatwole and Powell, 1998; Kordis et al., 2002; Gibbs and Rossiter, 2008), maintaining a wide variety of intact anticoagulation factors (mediated by strong purifying selection) would enable even an individual leech to feed on a wide array of prey.

As a first attempt to assess the type and level of selection acting on leech anticoagulants, here we show that purifying selection is extensive across the alignments of the various anticoagulants. In particular, the alignments for antistasin-family proteins, saratin/LAPP and destabilase show very high LRT scores for sites under purifying selection (Figs. 4.10b, 4.11b and 4.16b). Transitively, amino acid conservation is high throughout the alignments of all the anticoagulants, but polymorphisms and indels do exist and are especially conspicuous in the hirudin, manillase and eglin c alignments (Figs. 4.12a, 4.14a and 4.18a). Nonetheless, the importance of the anticoagulant proteins for the leeches seems to be manifested in the high levels of purifying selection, presumably acting on the genes in order to keep the ORF's intact. Common to most of the alignments is the fact that residues adjacent to disulphide-bond-forming cysteines often show high conservation, suggesting their importance to the structure and, by extension, function of the molecule. This is especially evident in the saratin/LAPP, bdellin, decorsin and destabilase alignments (Figs. 4.11a, 4.13a, 4.15a, 4.16a), as well as a particular region in the antistasin alignment (Fig. 4.10a).

By contrast, only isolated sites were predicted to be under positive selection, and these were commonly spread out across the alignments, as opposed to being concentrated in a particular region. Interestingly, for saratin/LAPP and ficolin, positive selection was estimated for sites situated in regions otherwise characterized by high levels of purifying selection (especially codons 100 and 147 in the saratin/LAPP alignment; Fig. 4.11a). If these polymorphic sites affect the structure-function relationship of the anticoagulants, this would allow each individual leech to simultaneously target a wide variety of factors in the coagulation cascades of their prey. In the specific case of saratin/LAPP, it has already been demonstrated that polymorphic orthologues occur as tandem arrays in the genome of the non-bloodfeeding leech *Helobdella robusta*, likely allowing the leeches to simultaneously target the wide assortment of collagen produced by their prey (Kvist et al., 2011). In the present study, we identified five different unigene transcripts for *Aliolimnatis fenestrata* and eight different unigene transcripts for *Macrobdella decora*. No saratin/LAPP orthologues were found in the *Hirudo verbana* EST library but because they occur in the closely related *Hirudo medicinalis*, this is likely contingent on the specifics of the EST library creation rather than *H. verbana* not possessing the anticoagulant.

On the one hand, the methods used here to assign selection pressures to codons overcome the lack of statistical power inherent in simple counting strategies and are less prone to false positives but, on the other hand, their performances are reliant on the size of the data sets (Kosakovsky Pond & Frost, 2005). Due to both the paucity of comparative data for leech anticoagulant repertoires and the short sequence nature of the peptides, some of the data sets used here may not possess the inherent degrees of freedom needed to accurately infer evolutionary selection pressures. In order to promote a fuller understanding of the selective pressures and diversities of anticoagulation factors (and other bioactive salivary peptides) across the evolutionary history of leeches, future studies should, importantly, appoint both target taxa that widen the scope of diversity and full transcriptome sequencing.

Decorsin, destabilase and hirudin: low or transient expression?

Most of the anticoagulants found in the present study are common to all three taxa. However, decorsin and destabilase were only found in *Macrobdella decora* and hirudin was not recovered in the *Hirudo verbana* EST library. This is notwithstanding that *H. verbana* has been the model for biomedical studies of hirudin for the last 20 years (Salzet, 2001; Siddall et al., 2007; Mamelak et al., 2010; Porshinsky et al., 2011; Gröbe et al., 2012). Whereas the absence of these anticoagulants can be taken as a sign that these are exclusive to certain taxa, it is also possible that they are transiently expressed in the salivary glands. In more specific terms, the leech may need excessive or particular stimuli in order to commence the secretion of these proteins or they may only be expressed after a certain time-period of bloodfeeding. An improbable alternative explanation to this involves the impediment of the expression of the anticoagulants by the use *Escherichia coli* as a vector (see Sudbery, 1996; Tan et al., 2007). However, by virtue of the identification of these anticoagulants in at least one taxon, it seems that using *E. coli* has no bearing on the expression levels.

Evolution of anticoagulants

Determining the genealogical relationships of the anticoagulation factors is important for understanding how evolutionary change within each molecule proceeds over time. For most of the anticoagulants, because of the often-paraphyletic nature of the orthologues from a single taxon (e.g., Fig. 4.4), the analyses performed here may suggest that these sequences represent several different (both paralogous and orthologous?) loci. This is further corroborated by the irreconcilability of several different transcripts within each taxon.

In addition to the phylogeny of the anticoagulants, the leech phylogeny adds a historical correlative framework for inferences on the evolution of the specific leech-derived proteins (Siddall et al., 2011). Overall, however, there is little concordance between the evolutionary histories of the anticoagulants (Figs. 4.1-4.9) and previous hypotheses of the leech phylogeny (Min et al., 2010; Siddall et al. 2011). In the leech hypotheses, glossiphoniid leeches are recovered at the base of the tree (see also Light and Siddall, 1999; Siddall et al., 2005) and *Aliolimnatis* and *Hirudo* are more closely related to each other than either is to *Macrobdella* (see also Phillips and Siddall, 2009).

The anticoagulant-trees shown in the present study often show little structure in terms of monophyly of orthologues from any single species. One exception to this, however, is the tree derived from the saratin/LAPP data set, which, if rooted at an orthologue from a glossiphoniid taxon, is largely congruent with the reigning leech phylogenetic hypothesis. The presence of different protein-variants within a single leech (Mason et al., 2004; Faria et al., 2005; Kvist et al., 2011) may be the cause for the lack of concordance between the anticoagulant trees and the leech phylogeny. Regardless of this, our analyses suggest that, much like Siddall et al. (2011) predicted, the origins of each of hirudin, bdellin, various antistasin-family proteins, eglin c and endoglucuronidases predate the origins of the medicinal leeches considered here.

References

Alaama, M., Alnajjar, M., Abdualkader, A.M., Mohammad, A., Merzouk, A. 2011. Isolation and analytical characterization of local Malaysian leech salivary extracts. IIUM Engineering Journal 12: 51-59.

Averbeck, M., Gebhardt, C.A., Voigt, S., Beilharz, S., Anderegg, U., Termeer, C.C., Sleeman, J.P., Simon, J.C. 2007. Differential regulation of hyaluronic acid metabolism in the epidermal and dermal compartments of human skin by UV irradiation. J. Invest. Dermatol. 127: 687-697.

Chatrenet, B., Chang, J-Y. 1993. The disulphide folding pathway of hirudin elucidated by stop/go folding experiments. J. Biol. Chem. 268: 20988-20996.

Connolly, T.M., Jacobs, J.W., Condra, C. 1992. An inhibitor of collagen-stimulated platelet activation from the salivary glands of the *Haementeria officinalis* leech. J. Biol. Chem. 10: 6893-6898.

Dabb, R.W., Malone, J.M., Leveret, L.C. 1992. The use of medicinal leeches in the salvage of flaps with venous congestion. Ann. Plast. Surg. 29: 250-256.

Dunwiddie, C., Thornberry, N.A., Bull, H.G., Sardana, M., Friedman, P.A., Jacobs, J.W., Simpson, E. 1989. Antistasin, a leech-derived inhibitor of factor Xa. Kinetic analysis of enzyme inhibition and identification of the reactive site. J. Biol. Chem. 264: 16694-16699.

Faria, F., Junqueira-de-Azevedo, I.L.M., Ho, P.L., Sampaio, M.U., Chudzinski-Tavassi, A.M. 2005. Gene expression in the salivary complexes from *Haementeria depressa* leech through the generation of expressed sequence tags. Gene 349: 173-185.

Fritz, H., Gebhardt, M., Meister, R., Fink, E. 1971. Trypsin-plasmin inhibitors from leeches – isolation, amino acid composition, inhibitory characteristics. In: Fritz, H., Tschesche, H. (Eds.), Proceedings of the International Research Conference on Proteinase Inhibitors, Walter de Gruyter, Berling, Germany.

Gibbs, H.L., Rossiter, W. 2008. Rapid evolution by positive selection and gene gain and loss: PLA_2 venom genes in closely related *Sistrurus* rattlesnakes with divergent diets. J. Mol. Evol. 66: 151-166.

Goloboff, P.A., Farris, J.S., Nixon, K.C. 2008. TNT, a free program for phylogenetic analysis. Cladistics 24: 774-786.

Greinacher, A., Völpel, H., Janssens, U., Hach-Wunderle, V., Kemkes-Matthes, B., Eichler, P., Mueller-Velten, H.G., Pötzsch, B. 1999. Recombinant hirudin

(Lepirudin) provides safe and effective anticoagulation in patients with heparin-induced thrombocytopenia. Circulation 99: 73-80.

Gröbe, A., Michaelsen, A., Hanken, H., Schmelzle, R., Heiland, M., Blessman, M. 2012. Leech therapy in reconstructive maxillofacial surgery. J. Oral Maxillofac. Surg. 70: 221-227.

Gronwald, W., Bomke, J., Maurer, T., Domogalla, B., Huber, F., Schumann, F., Kremer, W., Fink, F., Rysiok, T., Frech, M., Kalbitzer, H.R. 2008. Structure of the leech protein saratin and characterization of its binding to collagen. J. Mol. Biol. 381: 913-927.

Hardwick, C., Hoare, K., Owens, R., Hohn, H.P., Hook, M., Moore, D., Cripps, V., Austen, L., Nance, D.M., Turley, E.A. 1992. Molecular cloning of a novel hyaluronic acid receptor that mediates tumor cell motility. J. Cell Biol. 6: 1343-1350.

Heatwole, H., Powell, J. 1998. Resistance of eels (Gymnothorax) to the venom of sea kraits (*Laticauda colubrina*): a test of coevolution. Toxicon 36: 619-625.

Hovingh, P., Linker, A. 1999. Hyaluronidase activity in leeches (Hirudinea). Comp. Biochem. Phys. B 124: 319-326.

Itano, N., Kimata, K. 1996. Molecular cloning of human hyaluronic acid synthase. Biochem. Biophys. Res. Commun. 222: 816-820.

Jacoby, C. 1904. Uber Hirudin. Deut. Med. Wochenschr. 30: 1786.

Kordis, D., Bdolah, A., Gubensek, F. 1998. Positive Darwinian selection in *Vipera palestinae* phospholipase genes is unexpectantly limited to the third exon. Biochem. Biophys. Res. Commun. 251: 613-619.

Kosakovsky Pond, S.L., Frost, S.D.W. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22: 1208-1222.

Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S.V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676-679.

Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M. J., Bustamante, C.D., Nielsen, R., Siepel, A. 2008. Patterns of positive selection in six mammalian genomes. PLoS Genetics 4, e1000144 doi:10.1371/journal.pgen.1000144.

Kvist, S., Sarkar, I.N., Siddall, M.E. 2011. Genome-wide search for leech antiplatelet proteins in the non-bloodfeeding leech *Helobdella robusta* (Rhyncobdellida:

Glossiphoniidae) reveals evidence of secreted anticoagulants. Inv. Biol. 130: 344-350.

Light, J.E., Siddall, M.E. 1999. Phylogeny of the leech family Glossiphoniidae based on mitochondrial gene sequences and morphological data. J. Parasitol. 85:815-823.

Linker, A., Hoffman, P., Meyer, K. 1957. The hyaluronidase of the leech: an endoglucuronidase. Nature 180: 810-811.

Mamelak, A.J., Jackson, A., Nizamani, R., Amon, O., Liegeois, N.J., Redett, R.J., Byrne, P.J. 2010 Leech therapy in cutaneous surgery and disease. J. Drugs Dermatol. 9: 252-257.

Mason, T.A., McIlroy, P.J., Shain, D.J. 2004. A cysteine-rich protein in the *Theromyzon* (Annelida: Hirudinea) cocoon membrane. FEBS Lett. 561: 167-172.

Min, G-S., Sarkar, I.N., Siddall, M.E. 2010. Salivary transcriptome of the North American medicinal leech, *Macrobdella decora*. J. Parasitol. 96: 1211-1221.

Nutt, E., Gasic, T., Rodkey, J., Gasic, G.J., Jacobs, J.W., Friedman, P.A., Simpson, E. 1988. The amino acid sequence of antistasin. A potent inhibitor of factor Xa reveals a repeated internal structure. J. Biol. Chem. 263: 10162-10167.

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8: 785-786.

Phillips, A.J., Siddall, M.E. 2009. Poly-paraphyly of hirudinidae: many lineages of medicinal leeches. BMC Evol. Biol. 9: 246.

Porshinsky, B.S., Saha, S., Grossman, M.D., Beery II, P.R. Stawicki, S.P.A. 2011. Clinical uses of the medicinal leech: a practical review. J. Postgrad. Med. 57: 65-71.

Rydel, T.J., Ravichandran, K.G., Tulinsky, A., Bode, W., Huber, R., Roitsch, C., Fenton II, J.W. 1990. The structure of a complex of recombinant hirudin and human alpha-thrombin. Science 249: 277-280.

Salzet, M. 2001. Anticoagulants and inhibitors of platelet aggregation derived from leeches. FEBS Lett. 429: 187-192.

Scheffler, K., Martin, D.P., Seoighe, C. 2006. Robust inferences of positive selection from recombining coding sequences. Bioinformatics 22: 2493-2499.

Shain DH 2009. Annelids in modern biology. John Wiley, Sons, New York, USA.

Siddall, M.E., Budinoff, R.B., Borda, E. 2005. Phylogenetic evaluation of systematics and biogeography of the leech family Glossiphoniidae. Invertebr. Syst. 19:105-112.

Siddall, M.E., Burreson, E.M. 1995. Phylogeny of the Euhirudinea: independent evolution of blood feeding by leeches? Can. J. Zool. 73: 1048-1064.

Siddall, M.E., Burreson, E.M. 1996. Leeches (Oligochaeta?: Euhirudinea), their phylogeny and the evolution of life history strategies. Hydrobiologia 334: 277-285.

Siddall, M.E., Min, G-S., Fontanella, F.M., Phillips, A.J.P., Watson, S.C. 2011. Bacterial symbiont and salivary peptide evolution in the context of leech phylogeny. Parasitology 138: 1815-1827.

Siddall, M.E., Trontelj, P., Utevsky, S.Y., Nkamany, M. Macdonald III, K.S. 2007. Diverse molecular data demonstrate that commercially available medicinal leeches are not *Hirudo medicinalis*. Proc. R. Soc. Lond. B 274: 1481-1487.

Soucacos, P.N., Beris, A.E., Malizos, K.N., Kabani, C.T., Pakos, S. 1994. The use of medicinal leeches, *Hirudo medicinalis*, to restore venous circulation in trauma and reconstructive microsurgery. 13: 251-258.

Spicer, A.P., McDonald, J.A. 1998. Characterization and molecular evolution of a vertebrate hyaluronic acid synthase gene family. J. Biol. Chem. 273: 1923-1932.

Subramanian, A.R., Weyer-Menkhoff, J., Kaufmann, M., Morgenstern, B. 2005. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics 6: 66 doi: 10.1186/1471-2105-6-66.

Sudbery, P.E. 1996. The expression of recombinant proteins in yeasts. Curr. Opin. Biotech. 7: 517-524.

Tan, S., Wu, W., Li, X., Cui, L., Li, B., Ruan, Q. 2007. Enhanced secretion of adhesive recognition sequence containing hirudin III mutein in *E. coli.* Mol. Biotechnol. 36: 1-8.

Trontelj, P., Sket, B., Steinbrück, G. 1999. Molecular phylogeny of leeches: congruence of nuclear and mitochondrial rDNA data sets and the origin of bloodsucking. J. Zool. Syst. Evol. Research 37: 141-147.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J. 2009. Jalview version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

Wernersson, R., Pedersen, A.G. 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucl. Acids Res. 31: 3537-3539.

Whitaker, I.S., Izadi, D., Oliver, D.W., Monteath, G.M., Butler, P.E. 2004. *Hirudo Medicinalis* and the plastic surgeon. Brit. J. Plast. Surg. 57: 348-353.

CHAPTER V

PHYLOGENOMICS OF *REICHENOWIA PARASITICA*, AN ALPHAPROTEOBACTERIAL ENDOSYMBIONT OF THE FRESHWATER LEECH *PLACOBDELLA PARASITICA*

(Adapted from: Kvist, S., Narechania, A., Oceguera-Figueroa, A., Fuks, B., Siddall, M.E. 2011. Phylogenomics of *Reichenowia parasitica*, an alphaproteobacterial endosymbiont of the freshwater leech *Placobdella parasitica*. PLoS One 6: e28192. doi: 10.1371/journal.pone.0028192)

Abstract

Although several commensal alphaproteobacteria form close relationships with plant hosts where they aid in (e.g.,) nitrogen fixation and nodulation, only a few inhabit animal hosts. Among these, *Reichenowia picta, R. ornata* and *R. parasitica*, are currently the only known mutualistic, alphaproteobacterial endosymbionts to inhabit leeches. These bacteria are harbored in the epithelial cells of the mycetomal structures of their freshwater leech hosts, *Placobdella* spp., and these structures have no other obvious function than housing bacterial symbionts. However, the function of the bacterial symbionts has remained unclear. Here, we focused both on exploring the genomic makeup of *R. parasitica* and on performing a robust phylogenetic analysis, based on more data than previous hypotheses, to test its position among related bacteria. We sequenced a combined pool of host and symbiont DNA from 36 pairs of mycetomes and performed an *in silico* separation of the different DNA pools through subtractive scaffolding. The bacterial contigs were compared to 50 annotated bacterial genomes and the genome of the freshwater leech *Helobdella robusta* using a BLASTn protocol. Further, amino acid sequences inferred from the contigs were used as queries against the 50 bacterial genomes to establish orthology. A total of 358 orthologous genes were used for the phylogenetic analyses. In part, results suggest that *R. parasitica* possesses genes coding for proteins related to nitrogen fixation, iron/vitamin B translocation and plasmid survival. Our results also indicate that *R. parasitica* interacts with its host in part by transmembrane signaling and that several of its genes show orthology across Rhizobiaceae. The phylogenetic analyses support the nesting of *R. parasitica* within the Rhizobiaceae, as sister to a group containing *Agrobacterium* and *Rhizobium* species.

Introduction

Hematophagous leeches (Hirudinida) of the family Glossiphoniidae posses specialized organs related to the esophagous whose only known function is to house intracellular bacterial symbionts [1-3]. These structures, known as mycetomes or bacteriomes, show high morphological plasticity across the family ranging from granular tube-like structures circumscribing the esophagous in the genus *Placobdelloides* to distinct spheroid structures in the genus *Haementeria* [1]. In the genus *Placobdella*, the mycetomes are arranged as a pair of blind caeca about half-way down the esophagous [1,4]. Notably, mycetomes and the associated symbionts are completely absent from those leeches in Glossiphoniidae that have given up blood-feeding entirely (e.g., species of *Glossiphonia* and *Helobdella*). Because of the retention of these organs in hematophagous glossiphoniid leeches, the bacterial symbionts likely play an important role for the hosts. It has been hypothesized that the lack of essential nutrients, such as vitamins and enzymes, brought by the leeches' restricted diet of vertebrate blood [5], is ameliorated by the provision of nutrients by bacterial symbionts housed in the mycetomes [6]. Commonly in both plants and animals, obligate bacterial symbionts (primary symbionts) are housed in a distinct set of host-cells, known as bacteriocytes, and are strongly associated with these cells, to the point that they cannot invade unspecialized tissues [7]. The importance of the leech bacterial symbionts is also suggested by their vertical transovarial transmission [4].

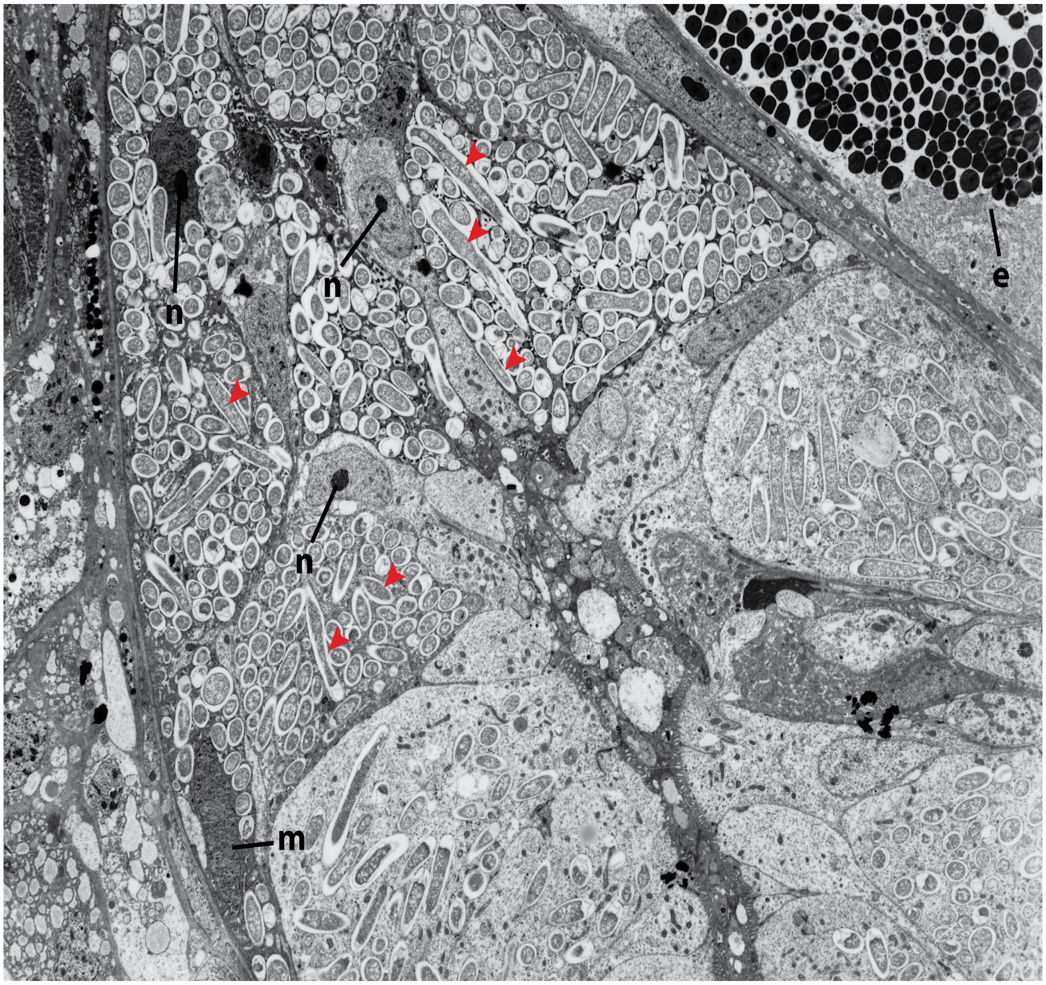
Although symbiotic associations between bacteria and leeches are well-documented [1,4,6,8,9], several questions concerning the details of the symbioses still remain. In particular, neither the function of the bacterial symbionts nor their putative "symbiont syndrome" has been clearly determined. The symbiont syndrome is a collective term for a set of features that are characteristic of intracellular bacterial symbionts [10,11]. These include a reduction in genome size, A-T bias, rapid sequence evolution and frequent gene rearrangements.

Siddall et al. [4] described the alphaproteobacterium *Reichenowia parasitica* from the mycetomes of its freshwater leech host, *Placobdella parasitica*, and, hitherto, the genus *Reichenowia* (three species; *R. picta, R. ornata* and *R. parasitica*) contain the only known mutualistic, endosymbiotic Rhizobiaceae that inhabit animal hosts. Other mutualistic alphaproteobacteria inhabit plants (e.g., *Rhizobium, Agrobacterium*) and most of those that infect animals (e.g., *Brucella* spp.) are parasitic [12; and references therein). Among other functions, bacterial plant-symbionts aid in nitrogen fixation and nodulation in the plants, allowing for more effective nutrient uptake and rapid growth

[13]. Moreover, the nitrogen fixation capability of prokaryotes has been highly studied because of its large impact on the ecosystem [14-16].

Using phylogenetic analyses, Siddall et al. [4] recovered *R. parasitica* within the family Rhizobiaceae but with low resolution concerning the internal placement of the species within this group. Moreover, for Siddall et al. [4], all attempts at culturing the bacteria, using various media, were unsuccessful, suggesting that the symbiont has a reciprocally obligate relationship with the host. Unculturable bacteria represent the majority of life forms [17]; many of these are endosymbionts of animal hosts and are vertically transmitted from parent to offspring, like *R. parasitica*. Taking into consideration that these bacteria prove refractory to culturing, direct and simultaneous sequencing of both associates is one of the few ways to obtain genetic material from the endosymbiont. It then becomes important to understand the diversity of the bacterial symbionts in the host. Primary evidence suggests that *R. parasitica* is the only bacterial symbiont to inhabit the mycetomal structures of the leech *Placobdella parasitica*. Several independent forms of data support this: first, multiple sequencing efforts of the mycetomes, using bacterial-specific primers for the 16S rDNA region, resulted in only a single bacterial haplotype [4]; second, fluorescent *in situ* hybridization of the mycetomes, using both alpha- and gammaproteobacterial probes, shows that only alphaproteobacteria are present and that these are found exclusively in the epithelial cell layer surrounding the sac-like structure [4] such that no contaminants would stem from intraluminal endosymbionts; third, transmission electron microscopy of the epithelial cells shows the presence of only one bacterial morphotype [4]. Interestingly, *R. parasitica* maintains a rod-shaped morphology (Fig. 5.1), common in free-living

Fig. 5.1. Transmission electron micrograph showing the rod-shaped morphology and several cross-sections of *Reichenowia parasitica*. The micrograph shows the inside of an epithelial cell of the mycetome from *Placobdella parasitica* at 5640x magnification, with some bacterial cells (red arrowheads), secretory esophageal cells (e), nuclei (n) and a mitochondrion (m) marked.


n
n
e
n
m

bacteria [18]. However, a rod-shaped morphology has been described also for endosymbiotic bacteria [19,20] and it is known that conversions from a rod shape to a sphere (but not the opposite) occur in single bacterial cultures [21,22].

Advances in sequencing technology allow for high-throughput and high-coverage sequencing of bacterial symbionts without the need to culture the bacteria [23]. We sought to characterize and annotate a large subset of the genome of *R. parasitica* in an attempt to investigate how the symbiont may affect the host and to assess the symbiont's phylogenetic position among a wide range of bacteria, with much greater genetic coverage than that of previous phylogenetic hypotheses.

Material and Methods

Leech Collection and Dissection

A total of 39 specimens of *Placobdella parasitica* were collected in Algonquin Park, Ontario, Canada in July 2009. All necessary collection permits were obtained from Ontario Parks, Canadian Ministry of Natural Resources. Most specimens were found attached to and feeding on hosts, specifically painted turtles (*Chrysemys picta*) and snapping turtles (*Chelydra serpentina*). Specimens were also collected by hand from under rocks, submerged wood and the underside of canoes. Specimens were brought back to the lab where they were dissected using a Nikon SMZ645 stereomicroscope. A total of 72 mycetomes (36 pairs) were removed from the leeches and directly transferred to Buffer AL (Qiagen Ltd.).

DNA Extraction, Amplification and Pyrosequencing

From the mycetomes, total combined genomic DNA from both the host and the bacterial associate was extracted using DNeasy Blood and Tissue Kit (Qiagen Ltd.) following the manufacturer's protocol with the addition of 1 µl of ribonuclease in order to increase the DNA/RNA ratio (i.e., 260/280 ratio). Due to the high amount of DNA required for pyrosequencing (10 µg), the extracted DNA was subjected to whole-genome amplification using REPLI-G UltraFast Mini Kit (Qiagen Ltd.). The amount of DNA was calculated by fluorometry to be in excess of 10 µg using Quant-iT PicoGreen Kit (Invitrogen). A GS Titanium Shotgun sequence library was prepared and massively parallel pyrosequencing of the amplicon was performed on the GS/FLX Titanium Shotgun XLR sequencing platform at SUNY Buffalo's Center for Excellence in Bioinformatics and Life Sciences (Buffalo, New York).

Assembly, Subtractive Scaffolding and Orthologue Recovery

The combined pool of host and symbiont DNA fragments from the FLX run were jointly assembled into contigs using Newbler ver. 2.3 (454 Life Sciences) and employing the "-large" option.

To separate the host and symbiont DNA, contigs were subjected to subtractive scaffolding: they were used as queries against 40 selected alphaproteobacterial target genomes and 10 non-alphaproteobacterial genomes (Beta-, Gamma-, Delta-, and Epsilonproteobacteria, as well as Firmicutes, Aquificae, Bacteroidetes and Cyanobacteria), both from endosymbiotic and free-living bacteria, and with largely varying genome sizes (see Supplementary Table S5.1). Alphaproteobacteria were over-represented because of previous phylogenetic hypotheses placing *R. parasitica* within this class [4,6]. Moreover, the contigs were queried against the only sequenced leech genome, *Helobdella robusta* (family Glossiophoniidae), which coincidentally is in the same taxonomic family as *Placobdella parasitica* [24,25]. The leech genome is available at the DOE Joint Genome Institute portal website (http://genome.jgi-psf.org/Helro1/Helro1.home.html). Two local searches were performed using the BLASTn protocol applying default settings, one with a cut-off expectation value of $1E^{-5}$ and the other with $1E^{-2}$. All contigs simultaneously matching both associates using the $1E^{-2}$ cut-off rate were also deleted from the $1E^{-5}$ data set. The criteria were asymmetric in order to enrich for bacterial sequences in our retained DNA-pool; the purpose being to completely purge the leech DNA from the data set, including putative chimeric sequences resulting from the nested assembly of both associates. With these criteria, each of the retained hits necessarily had a three orders of magnitude lower e-value when queried against bacteria than when queried against leech. Annotations of the *R. parasitica* sequences follow the GenBank annotations of the 50 bacterial genomes and inferences of molecular function follow information from UniProt and appropriate references.

Retained bacterial contigs also were subjected to gene prediction using GeneMark ver. 2.4 [26], which employs both ORF's (Open Reading Frames) and hidden Markov models for prediction, and using *Sinorhizobium meliloti* as a scaffold genome. This species was chosen by virtue of previous phylogenetic hypotheses showing a close relationship between *R. parasitica* and *S. meliloti* [6]. Resulting nucleotide sequences of putative genes were translated into stop-codon-free amino acid sequences by GeneMark and these were then queried against the 50 bacterial proteomes downloaded from GenBank. Orthologues were recovered employing a tree-based approach as implemented in OrthologID [27]. A 70% similarity cut-off rate and a lower e-value limit of $1E^{-10}$ were employed. OrthologID was also used to align the amino acid sequences using multiple sets of alignment parameters and employing the MAFFT L-INS-i algorithm [28].

Clusters of Orthologous Groups (COG's)

The predicted *Reichenowia parasitica* genes as well as genes from species of *Agrobacterium*, *Mesorhizobium*, *Wigglesworthia*, *Buchnera* and *Escherichia* each were compared to the NCBI COG database (http://www.ncbi.nlm.nih.gov/COG/) by in-house scripting. The species were chosen with respect to their phylogenetic placement and life history strategies (see Results). A *ruby* script was run locally to compare each of the genes against the database and only the best hit for each gene was retained.

Phylogenetic Analyses

The matrix of the aligned amino acid orthologues recovered by OrthologID was subjected to parsimony analysis using TNT [29] and likelihood analysis using RAxML ver. 7.2.8 [30]. In TNT, a New Technology search was conducted employing sectorial searching, with the tree fusing and ratcheting algorithms turned on. Trees were retrieved by a driven search using 100 initial addition sequences and requiring that the minimum length tree be found a total of 10 times. All characters were equally weighted and non-additive, and gaps were treated as missing data. Support values for nodes were also calculated in TNT through both standard bootstrap resampling and partition bootstrapping [31] using the *blockboot.run* script available on the TNT Wiki site (http://tnt.insectmuseum.org/index.php/Manual) for the latter. Both bootstrap analyses employed 100 iterations, each subjected to ten iterations of ratcheting and three rounds of tree fusing after an initial five rounds of Wagner tree building. To examine the relative support of each separate locus predicted by GeneMark for the tree obtained from all of the data, constrained analyses were employed in PAUP* ver. 4.0b10 [32].

For the likelihood analyses, a heuristic search was performed under both PROTCATJTTF and PROTGAMMAJTTF models of protein evolution, treating the blocks as a single set. Runs were performed for 100 iterations with an initial 25 CAT rate categories and final optimization with 4 gamma shape categories. Bootstrap analysis employed the PROTGAMMAJTTF model for 100 pseudoreplicates with a random starting-tree for each replicate.

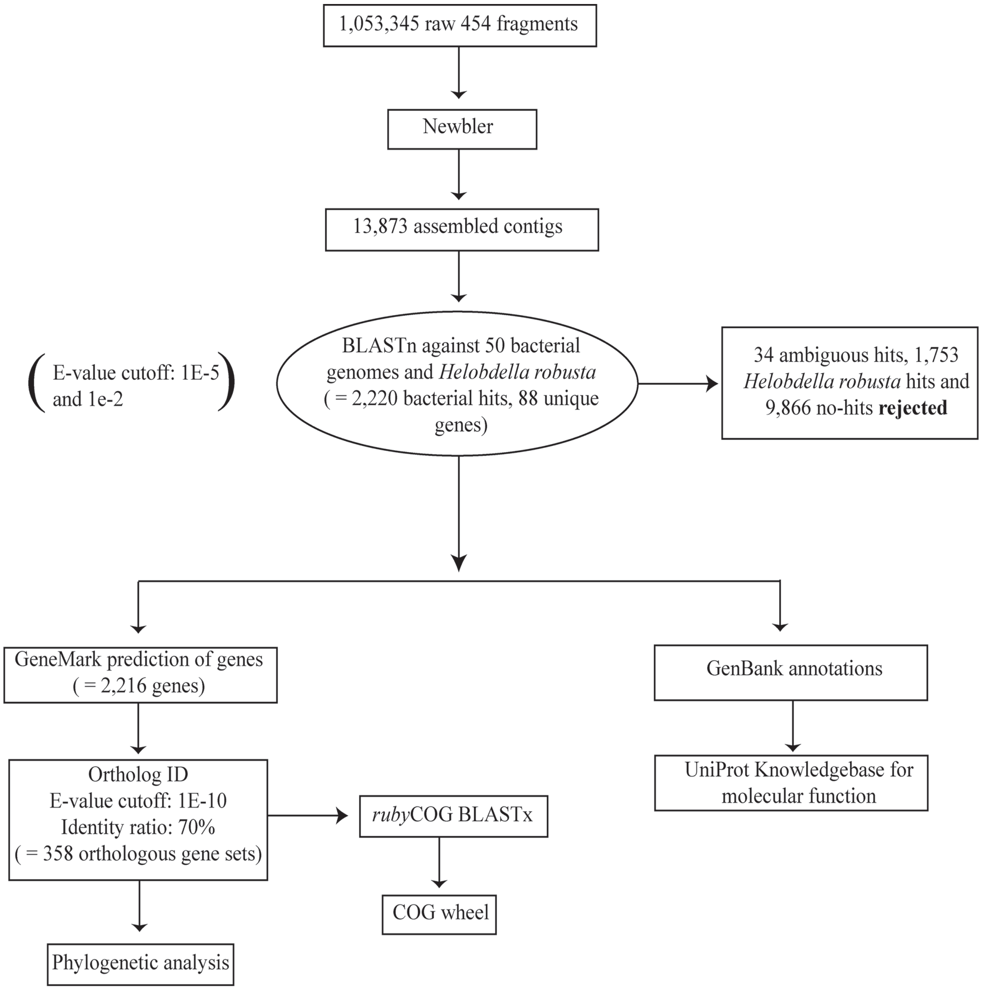
The trees were rooted with *Aquifex aeolicus* (Aquificae), following the hypotheses of Snel et al. [33].

Results

Sequence Analysis

The main workflow of this study is presented in Figure 5.2. The pyrosequencing returned 1,053,345 fragments of mixed host and symbiont DNA (GenBank Sequence Read Archive [SRA] accession number SRA030522.3) and these were assembled into 13,873 contigs by Newbler. The BLASTn search using a cutoff e-value of $1E^{-5}$ resulted

Fig. 5.2. Main workflow followed in this study


1,053,345 raw 454 fragments
Newbler
13,873 assembled contigs
BLASTn against 50 bacterial genomes and *Helobdella robusta* (= 2,220 bacterial hits, 88 unique genes)
(E-value cutoff: 1E-5 and 1e-2)
34 ambiguous hits, 1,753 *Helobdella robusta* hits and 9,866 no-hits **rejected**
GeneMark prediction of genes (= 2,216 genes)
GenBank annotations
Ortholog ID E-value cutoff: 1E-10 Identity ratio: 70% (= 358 orthologous gene sets)
*ruby*COG BLASTx
UniProt Knowledgebase for molecular function
COG wheel
Phylogenetic analysis

in 2,247 of the contigs hitting bacteria alone, 1,753 contigs hitting leech alone, seven contigs hitting both the 50 bacterial genomes and the leech and 9,866 contigs not hitting either of these (Table 5.1). Among the seven ambiguous contigs, four hit bacteria with very low e-values ($1E^{-37}$-$1E^{-175}$) while, at the same time, showing high e-values for the leech hit ($1E^{-6}$-10^{-10}). The remaining three hits showed the reverse scenario with low e-values for leech hits and high e-values for bacterial hits, implying that these seven contigs are not shared by the leech and bacterial genomes but, rather, are artifacts of the protocol used for the BLAST search. The second BLASTn search ($1E^{-2}$) resulted in 2,611 of the contigs hitting bacteria alone, 4,553 contigs hitting leech alone, 207 contigs hitting both bacteria and leech and 6,502 contigs hitting neither bacteria nor leech (Table 5.1). From the resulting 2,247 contigs matching bacteria at $1E^{-5}$, 27 out of the total 207 contigs matching both associates at $1E^{-2}$ were removed. The remaining 180 ambiguous contigs were predicted leech hits at $1E^{-5}$ and also hit bacteria with marginal e-values at $1E^{-2}$; these were already removed from the data set after the $1E^{-5}$ search. After pruning, 2,220 definitive bacterial contigs were retained. The 2,220 contigs, in turn, pertained to 88 uniquely annotated genes among the 50 bacterial genomes and 39 of these were hit with a perfect e-value (0). As was expected, most of the bacterial contigs hit multiple times for the same annotated locus but with differing e-values and starting/stopping points for a total of 42,025 hits stemming from the 2,220 *R. parasitica* contigs. The most frequently found annotations of the *R. parasitica* contigs, in terms of representation, seem to relate to two biological processes: transportation and catalytic activity of various components. Other rather highly represented biological processes among the contig matches were DNA transcription and metabolic processes, and for several of the hit-

Table 5.1. Distribution of leech and bacterial BLASTn hits among the 13,873 contigs assembled from the 454 pyrosequencing reads. Ambiguous hits indicate those contigs that matched both leech and bacteria simultaneously. In the 1E^{-5} protocol, both the leech and ambiguous hits, as well as the contigs without match were deleted from the data set. Also, 27 out of the 207 contigs matching both associates at 1E^{-2} (the remaining 180 hits were predicted leech hits at 1E^{-5}) were deleted from the 2,247 contigs matching bacteria at 1E^{-5} resulting in 2,220 definitively bacterial hits in the data set. See the results section for further discussion.

Cutoff e-value	# Bacterial hits	# Leech hits	# Ambiguous hits	# No match hits
1E^{-5}	2,247	1,753	7	9,866
1E^{-2}	2,611	4,553	207	6,502

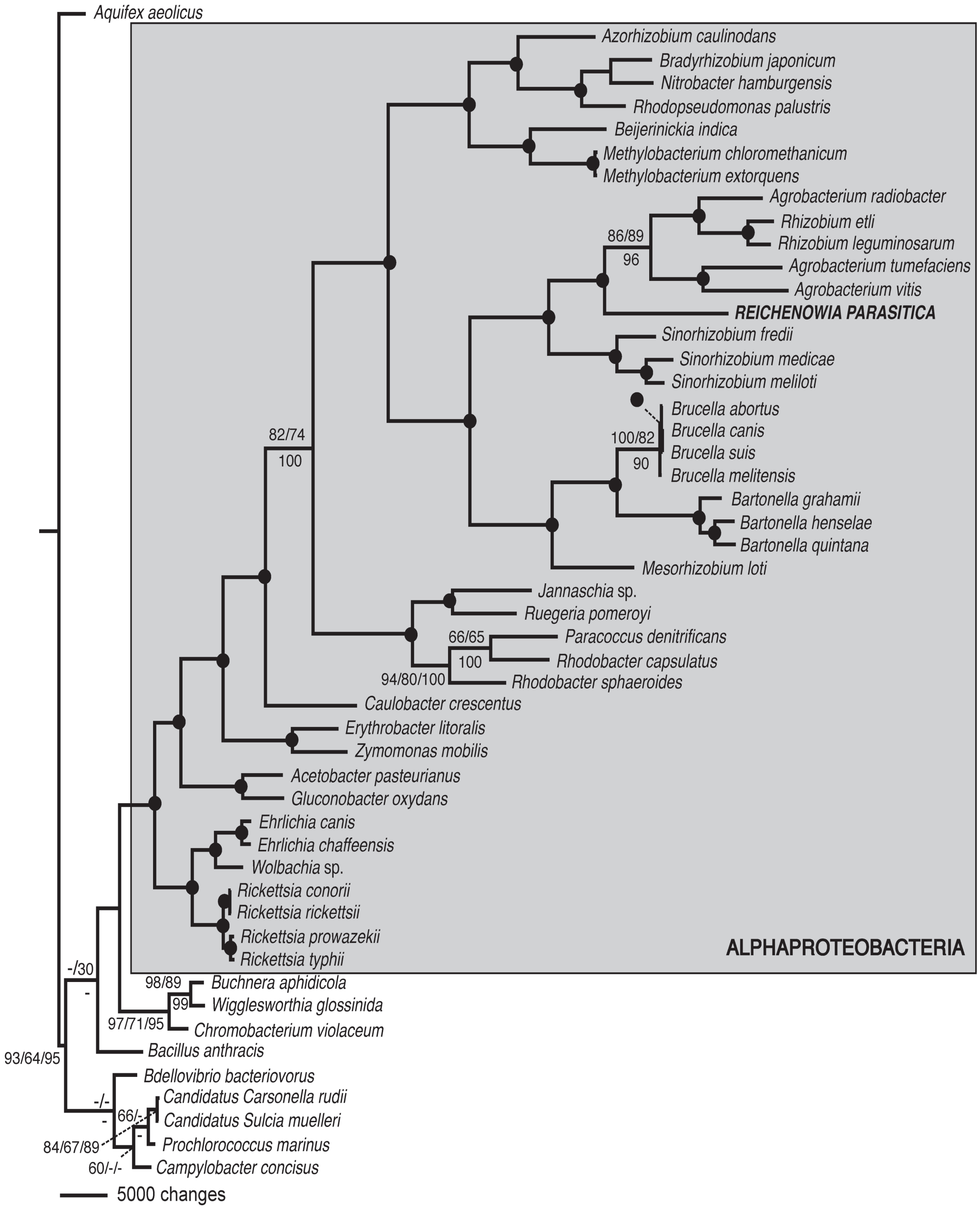
descriptions of our contig matches there is little or no information in the Protein Knowledgebase, UniProtKB (e.g., polyhydroxyalkonate synthesis repressor; 1975 hits).

Gene Prediction and Phylogeny

Among the 2,220 *R. parasitica* contigs, GeneMark predicted 2,916 genes for a total of 1,785,377 basepairs. The G+C content pertaining to these was 62.78%. OrthologID identified a total of 9,135 orthologous genes among the 51 (including *R. parasitica*) genomes, 358 of which included an *R. parasitica* orthologue (3.9% of the total gene-groups). That is, among the 2,916 *R. parasitica* genes predicted, 358 were found orthologous to any of the genes in the 50 bacterial genomes. These orthologues accounted for 181,848 aligned amino acids sites, and these were jointly submitted to TNT and RAxML for phylogenetic analyses. The percentage of missing data amounted to ~55% within the total data set, due to numerous instances of gene loss, common in bacterial genomes and anticipated to be even more so in endosymbionts [34,35].

Out of the 181,848 aligned amino acid sites, 58,887 were parsimony informative. Each of the retained gene groups containing an *R. parasitica* orthologue (n=358) was used as an independent block both for the partition bootstrapping and the partition congruence test. The TNT run and both RAxML runs (using PROTCATJTTF and PROTGAMMAJTTF models of evolution) returned optimal trees with identical topologies; a single most parsimonious tree with a length of 408,192 steps for the TNT run and a tree with an ln *L* of -2,262,856.651 for the RAxML run using the PROTGAMMAJTTF model. In the tree (Figure 5.3), the alphaproteobacteria, as well as each of the families contained therein were recovered as monophyletic, and 33 out of the

Fig. 5.3. Single most parsimonious tree (length= 408,192 steps, consistency index=0.647 and retention index=0.648) recovered from the phylogenetic analysis of the 358 orthologues across 51 taxa. The topology is identical to the maximum likelihood tree recovered by RAxML. Values above the nodes are standard bootstrap re-sampling and partition bootstrap values, respectively, and below the nodes are likelihood bootstrap values. Solid black circles denote nodes with bootstrap support ≥ 90% for all three support measures.


Aquifex aeolicus
Azorhizobium caulinodans
Bradyrhizobium japonicum
Nitrobacter hamburgensis
Rhodopseudomonas palustris
Beijerinckia indica
Methylobacterium chloromethanicum
Methylobacterium extorquens
Agrobacterium radiobacter
Rhizobium etli
Rhizobium leguminosarum
Agrobacterium tumefaciens
Agrobacterium vitis
86/89
96
REICHENOWIA PARASITICA
Sinorhizobium fredii
Sinorhizobium medicae
Sinorhizobium meliloti
Brucella abortus
Brucella canis
Brucella suis
Brucella melitensis
100/82
90
82/74
100
Bartonella grahamii
Bartonella henselae
Bartonella quintana
Mesorhizobium loti
Jannaschia sp.
Ruegeria pomeroyi
66/65
100
Paracoccus denitrificans
Rhodobacter capsulatus
94/80/100
Rhodobacter sphaeroides
Caulobacter crescentus
Erythrobacter litoralis
Zymomonas mobilis
Acetobacter pasteurianus
Gluconobacter oxydans
Ehrlichia canis
Ehrlichia chaffeensis
Wolbachia sp.
Rickettsia conorii
Rickettsia rickettsii
Rickettsia prowazekii
Rickettsia typhii
ALPHAPROTEOBACTERIA
-/30
-
98/89
Buchnera aphidicola
99
Wigglesworthia glossinida
97/71/95
Chromobacterium violaceum
Bacillus anthracis
93/64/95
Bdellovibrio bacteriovorus
-/-
-
Candidatus Carsonella rudii
66/-
Candidatus Sulcia muelleri
-
84/67/89
Prochlorococcus marinus
60/-/-
Campylobacter concisus
5000 changes

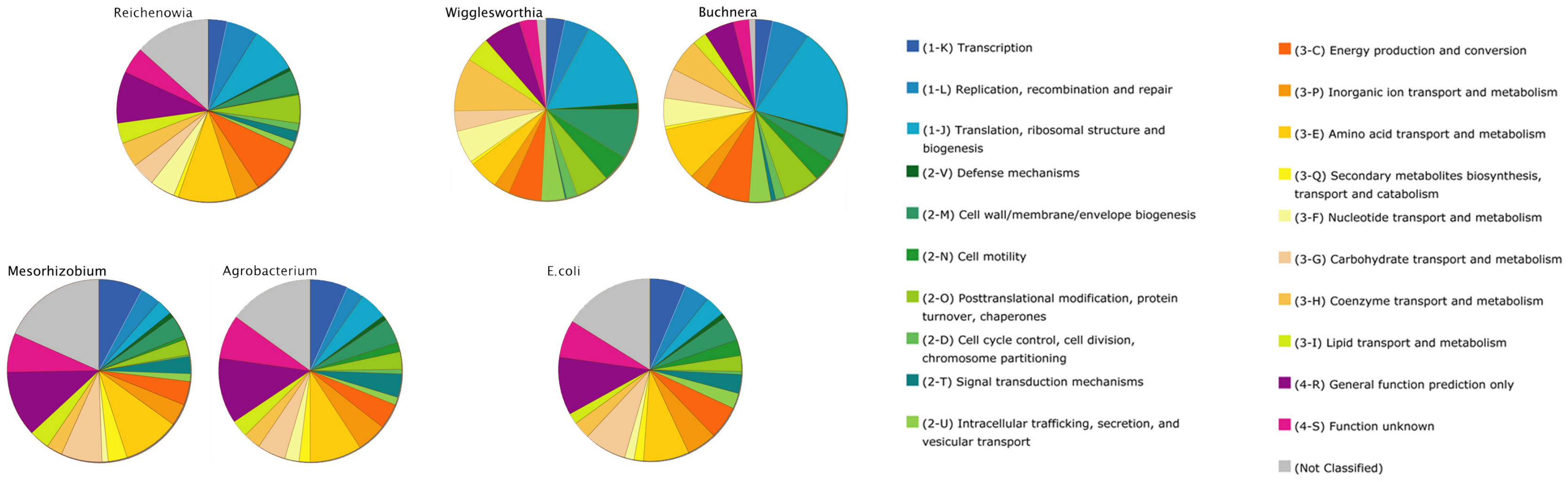
48 nodes show high support for all three support measures (>90% parsimony bootstrap support: bs; parsimony partitioned bootstrap support: pbs; likelihood bootstrap support: lbs). *Reichenowia parasitica* was recovered nested within the Rhizobiaceae (100% bs; 100% pbs; 100% lbs), as sister to a monophyletic cluster consisting of *Agrobacterium* and *Rhizobium* species (86% bs; 89% pbs; 96% lbs), and this group in turn placed as sister to the *Sinorhizobium* species (100%bs; 100% pbs; 100% lbs). Rhizobiaceae (the genera mentioned above) was recovered as sister to a larger assemblage containing species of the families Brucellaceae, Bartonellaceae and Phyllobacteriaceae (100% bs; 97% pbs; 100% lbs). In addition, relative support conferred by each locus (n=358), for the placement of *R. parasitica* within Rhizobiaceae was assessed by employing constraint trees in PAUP* (under the parsimony criterion). That is, for each locus, two values were found: one constraining to include *R. parasitica* in Rhizobiaceae, and another excluding it from Rhizobiaceae (but imposing no other relationship constraints on taxa). In the combined analysis, the number of extra steps incurred by combining the partitions was 9,867 and the difference in length between the best trees constraining *R. parasitica* to be inside and outside of Rhizobiaceae was 232 steps (~2.4% of the total incongruence). A total of 206 loci (58%) support the placement of *R. parasitica* inside of Rhizobiaceae, whereas only 45 partitions (13%) do not support its placement inside the family. The sum of the number of extra steps from partitions that do not support *R. parasitica* inside of Rhizobiaceae is 371. However, 1057 extra steps are required to remove *R. parasitica* from the family. In other words, there is almost three times as much information supporting the placement of *R. parasitica* inside of Rhizobiaceae, as opposed to outside the family. Though none of the 45 partitions individually place *R.*

parasitica in Rhizobiaceae, even combining these 45 loci again places the species inside of the family.

COG Analyses

Alternative to the annotations of the contig matches mentioned above, using the amino acids of the predicted genes (employing the 358 partitions as separate loci) as queries against the COG database unveiled three main functional groups to which the *R. parasitica* genes seem to be related. These were: (i) information storage and processing, (ii) cellular processes and signaling and (iii) metabolism. The distributions and subdivisions of these COG-groups are presented in Figure 5.4. The *R. parasitica* COG-groups were also compared to those of the closely related plant-inhabiting *Agrobacterium* and *Mesorhizobium* to investigate for patterns in the devotion of the genome to particular processes as a result of a change in endosymbiotic lifestyle from a plant to an animal host. To corroborate the findings, the COG's of *Wigglesworthia* and *Buchnera*, both gammaproteobacterial endosymbionts of animals, were compared to the ubiquitous gammaproteobacterium, *Escherichia coli*. The patterns within and across these largely separate phylogenetic clusters were then investigated. When compared to related non-animal endosymbionts, *Reichenowia, Wigglesworthia* and *Buchnera* all show a decrease in the proportion of genes devoted to transcriptional processes (Figure 5.4; dark blue field 1-K). Furthermore, they all show an increase in proportional gene-devotion to each of translation, ribosomal structure and biogenesis (Figure 5.4; light blue field 1-J), posttranslational modification, protein turnover and chaperones (Figure 5.4;

Fig. 5.4. Comparison of Clusters of Orthologous Groups (COG's) between animal and non-animal endosymbionts. The 358 *R. parasitica* orthologues, as well as the genomes of species of *Agrobacterium, Mesorhizobium, Wigglesworthia, Buchnera* and *Escherichia* were used as queries against the database. The different colors denote separate functional groups to which the genes are linked. In both of the phylogenetically related groups (left: *Reichenowia, Agrobacterium* and *Mesorhizobium*, and right: *Wigglesworthia, Buchnera* and *Escherichia*) the topmost wheels represents animal-inhabiting endosymbionts, whereas the bottommost wheels represent non-animal endosymbionts. When compared to the non-animal endosymbionts, the animal endosymbionts each show a decrease in the proportion of genes related to 1-K (transcription), and an increase in the proportion of genes related to 1-J (translation, ribosomal structure and biogenesis), 2-O (posttranslational modification, protein turnover, chaperones), and 3-F (nucleotide transport and metabolism).


Reichenowia
Wigglesworthia
Buchnera
Mesorhizobium
Agrobacterium
E.coli
(1-K) Transcription
(1-L) Replication, recombination and repair
(1-J) Translation, ribosomal structure and biogenesis
(2-V) Defense mechanisms
(2-M) Cell wall/membrane/envelope biogenesis
(2-N) Cell motility
(2-O) Posttranslational modification, protein turnover, chaperones
(2-D) Cell cycle control, cell division, chromosome partitioning
(2-T) Signal transduction mechanisms
(2-U) Intracellular trafficking, secretion, and vesicular transport
(3-C) Energy production and conversion
(3-P) Inorganic ion transport and metabolism
(3-E) Amino acid transport and metabolism
(3-Q) Secondary metabolites biosynthesis, transport and catabolism
(3-F) Nucleotide transport and metabolism
(3-G) Carbohydrate transport and metabolism
(3-H) Coenzyme transport and metabolism
(3-I) Lipid transport and metabolism
(4-R) General function prediction only
(4-S) Function unknown
(Not Classified)

light green field 2-O), and nucleotide transport and metabolism (Figure 5.4; light yellow field 3-F).

Discussion

Beyond corroborating and solidifying the hypothesis that *Reichenowia parasitica*, a mutualistic, intracellular bacterial symbiont of the fresh-water leech *Placobdella parasitica*, places phylogenetically among the alphaproteobacterial Rhizobiaceae, the present study also reveals several interesting features of the genomic makeup of the bacterium. Some of the BLAST-based hits, e.g., histidine ammonia-lyase (1100 hits among the *R. parasitica* contigs; Supplementary Table S5.2) are fairly common across prokaryotes and eukaryotes alike [36] while other loci are more elusive, making them of special interest based on our, albeit limited, knowledge of the biology of the symbiont. Some of these loci are discussed below (see Supplementary Table S5.2 for the full list of hits) and a broad phylogenetic discussion is presented. Insofar as the *R. parasitica* genome was only partially sequenced, no examination of the functional consequences of the lack of genes can be definitively performed.

Cation Pump Membrane Proteins (Nitrogen Fixation)

Because of the close relationship of *R. parasitica* to each of *Rhizobium* and *Sinorhizobium*, it is likely that these taxa share genes by virtue of having a rather recent common ancestor. Both of the mentioned genera have been frequently studied for their established symbiosis with legumes, and in particular for their nitrogen fixation capabilities [37]. Already, Siddall et al. [4] noted that *Reichenowia* species are especially interesting because of their putative role in nitrogen metabolism in the leech hosts. Here, we identified 34 contigs that show high sequence similarity to the cation pump membrane proteins of *Rhizobium etli*, and 6 contigs that show similarity to potassium ion transmembrane transporter proteins from *Sinorhizobium medicae* (Supplementary Table S5.2). Cation pump membrane proteins, such as FixG, FixH, FixI or Na^+/K^+ ATPase, are required for symbiotic nitrogen fixation and it has been hypothesized that these genes are private (i.e., present only in a specific group, but not necessarily in all members of that group) to symbiotic bacteria, as they do not hybridize well with DNA from free-living bacteria [38-40]. Notwithstanding the K^+ ion transporters, it is unclear which type of cation pump membrane protein the *R. parasitica* contigs are related to but, regardless, they may be involved in nitrogen metabolism in the host. In addition, cation pumps have been shown to be coupled with redox processes [38,41] and numerous *R. parasitica* contigs show sequence similarity to known oxyreductase proteins (e.g., NuoK2 NADH: quinone oxidoreductase in *Sinorhizobium meliloti*, XoxF in *Methylobacterium extorquens* and oxidoreductase in *Agrobacterium vitis*; see Supplementary Table S5.2), providing a possibility for coupling of cation pumps and redox systems in the bacteria.

Nitrogen fixation is vital for biosynthesis of amino acids in plants and has been coupled with metabolic processes in animals. For example, in the shipworm *Lyrodus pedicellatus* (Bivalvia), nitrogen fixation by the endosymbiotic gammaproteobacteria *Teredinibacter turnerae* has enabled the shipworm to survive and grow on a nitrogen-poor diet [42]. That is, the *L. pedicellatus* - *T. turnerae* system is an example of a symbiosis, in which atmospheric nitrogen is converted into animal biomass. To this end, the leech host, *Placobdella parasitica*, may increase its growth due to the increase in organic nitrogen provided by the bacteria. Moreover, the leech may be alleviated from costly inorganic nitrogen excretion due to the conversion of inorganic to organic nitrogen by the bacteria.

Iron Siderophore/Cobalamin (Vitamin B_{12}) ABC Transporters

ATP binding cassette (ABC) transmembrane transporters consist of two membrane-spanning domains, which form a translocation pathway, and two cytoplasmic ABC domains, which power the transport process [43]. In prokaryotes, ABC transporters are chiefly devoted to the export and import of essential nutrients, such as iron and vitamin B_{12} in *E. coli* [44]. Several nutrients, including vitamin B_{12}, are low in vertebrate blood such that hematophagous parasites must rely on a symbiotic organism that has the capability of synthesizing and transporting them to the host [45]. Dietary supplementation experiments have shown that endosymbiotic bacteria (*Wigglesworthia*) in the bloodfeeding tsetse fly play a role in vitamin B metabolism [5,46]. The primary diet for *Placobdella parasitica* is poikilothermic vertebrate blood, which is low on vitamin B_{12}. Therefore, it would make sense for the leech to harbor bacterial symbionts with the capacity for synthesizing and transporting vitamin B_{12} across cell membranes to host receptors. An iron siderophore/cobalamin (vitamin B_{12}) ABC transporter from *Rhodobacter capsulatus* significantly matched 19 contigs in *R. parasitica*, putatively indicating that, as speculated by Perkins et al. [6], the bacteria supply essential nutrients to the leech host.

Prevent-Host-Death (phd) *Family Proteins*

Escherichia coli is the most well known symbiont to exhibit plasmid addiction. Plasmid-encoded addiction genes are thought to be involved in conserving low-copy bacterial plasmids by selectively killing cells that have lost a plasmid. For the prevent-host-death system, this entails two genetic markers: the toxin (death-on-curing; *doc*) and the antitoxin (*phd*). Functionally, in cells that posses the focal low-copy plasmid, *phd* must be maintained at a sufficient level to inhibit the function and/or synthesis of the toxin, allowing survival of plasmid-possessing cell-lines and ultimately the plasmids themselves [47]. Because of the high energy-expenditure involved in producing antitoxins by the cells only to maintain status quo, plasmid addiction has been referred to as a Red Queen-type system. In total, 26 *R. parasitica* contigs were matched with DNA sequences annotated as *phd*-type proteins from *Methylobacterium chloromethanicum* (alphaproteobacteria) (Supplementary Table S5.2). As our knowledge of the plasmid set-up for *R. parasitica* is virtually non-existent, this finding at least indicates that the bacteria posses plasmids (although some bacterial toxin-antitoxin systems are chromosomally encoded; e.g., [48]). A more in-depth study of the plasmid addiction associates would be beneficial as it would allow for an understanding of the plasmid count, composition and expression levels in the bacterial symbiont, as well as the underlying survival techniques of the plasmids.

Antirestriction Family Proteins

Antirestriction family proteins are commonly involved in overcoming restriction barriers during establishment after conjugative transfer. For example, in *E.*

coli antirestriction proteins of type Ard (Alleviation of Restriction of DNA) specifically affect the restriction activity of type I restriction-modification systems, and may be involved in the regulation of gene transfer between bacterial genomes [49]. Moreover, the restriction-modification system is important in limiting the transfer of genetic elements responsible for bacterial resistance to antibiotics [50], making the inhibition of this system by the antirestriction proteins of human concern.

The BLASTn search, performed here, recovered 125 *R. parasitica* contigs with low e-values when compared to genes annotated for antirestriction family proteins (Supplementary Table S5.2). As with the *phd* family proteins (see above), this result indicates that *R. parasitica* does possess plasmids, unlike several other bacterial symbionts [51]. In regards to the function, it is still unclear if *R. parasitica* uses the putative genes for any of the reasons mentioned above. When compared to the protein sequence of annotated antirestriction proteins from *Agrobacterium vitis* (GenBank Protein ID: YP_002551430.1), one of the contigs shows 27% conservation (for shared amino acid positions). At this stage, we cannot conclude that the putative antirestriction protein present in *R. parasitica* does not function in the same way as in other prokaryotes, as a counteract against the restriction-modification system ultimately allowing foreign DNA to enter the cell. However, without performing functional analyses (such as mutagenesis), it would be premature to infer that these proteins are functionally related.

Autoaggregation Proteins

Autoaggregation proteins share homology with adhering proteins of e.g., *Rhizobium* species [52]. Adhering proteins are calcium-binding proteins that recognize receptors on the bacterial surface, leading to congregation of cells. In plant associated symbionts, it is thought that the proteins are involved in the attachment process to plant lectins [53]. For many animal pathogens (e.g., *Bartonella* spp.), an important factor for virulence is that the bacteria can adhere to the host-cell surface or the extracellular matrix components. It is likely that *R. parasitica* uses these putative adhesion proteins in much the same way. By sticking to the mycetomal cell walls, and to each other, the bacteria can maintain their position in the cell. In fact, transmission electron microscopy has shown that the cytoplasmic space of epithelial cells in the mycetomes of *Placobdella* species are almost completely filled with bacteria [4], suggesting the need for adhesion to the host-cell walls. A total of 1972 *R. parasitica* contigs hit autoaggregation protein (adhering protein from *Rhizobium etli* CFN 42) with significant e-values (Supplementary Table S5.2).

Phylogeny

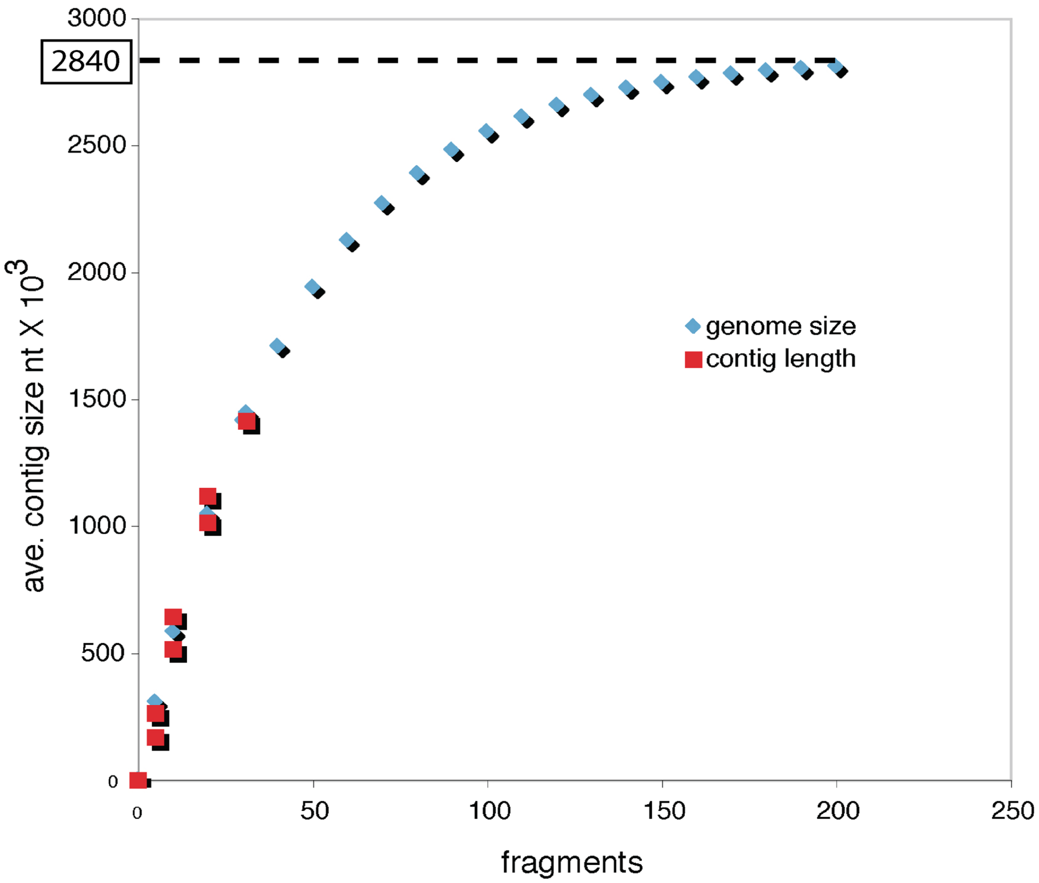
Based on both parsimony and likelihood algorithms, Siddall et al. [4] performed a phylogenetic analysis of three *Reichenowia* species using 16S and 23S ribosomal RNA. That study, corroborated by the present study (see Figure 5.3), recovered *R. parasitica* among the Rhizobiaceae as sister to a group including the *Rhizobium* and *Agrobacterium* species. Later, Perkins et al. [6] recovered the same three species as sister to a group containing *Sinorhizobium meliloti* (with an unresolved position), *Brucella melitensis* and *Brucella henselae*. In the analysis performed by Perkins et al. [6], the *Agrobacterium* species and the *Rhizobium* species were recovered as consecutive sister-groups to this larger group. From a biological standpoint, and because contemporary bacterial taxonomy and phylogenetics focuses largely on 16S and 23S rDNA [54-56], it is comforting to know that the phylogenetic signal present in 16S or 23S alone is rather concordant with that of the 358 orthologues used here.

The well-supported plant-symbiont affiliation of *R. parasitica* raises some interesting questions concerning the evolutionary history of the bacteria. Because of the basal position of the *Sinorhizobium* species in the phylogenetic hypothesis presented here, the ancestral life history trait of the Rhizobiaceae seems to be plant symbiosis, with *R. parasitica* showing a host switch from plant to leech. This is further supported by the finding of several plant-associated genes, such as phosphatase, in the genome of *R. parasitica* (Supplementary Table S5.2). Out of the 358 orthologues detected among the *R. parasitica* contigs, several were private to *Rhizobium, Agrobacterium, Sinorhizobium* and *Reichenowia*, possibly indicating common ancestry among these genera. However, it is also possible that the ancestor of the *R. parasitica* was free-living by virtue of the rod-shape of the bacterium, a shape common in several other free-living taxa [57], and it is possible that the same free-living ancestor also evolved into the plant-symbiotic bacteria that we see today. A more taxonomically rich study of the alphaproteobacteria as a whole will likely shed light on the ancestral life-history strategy of the Rhizobiaceae.

The phylogenetic hypothesis also enables some inferences regarding the currently unknown genome size of *R. parasitica*. Among other things, an understanding of the genome size of the symbiont may guide future sequencing efforts of its entire genome. The size of the chromosomal genomes of the *Agrobacterium* and *Rhizobium* species (sister to *Reichenowia*) used here range between 5.66-7.42 megabasepairs (Mbp), whereas the *Sinorhizobium* species (basal to *Reichenowia*) possess chromosomal genomes in the range of 6.71-6.89 Mbp. By extension, it is probable that the genome size of *R. parasitica* is somewhere in the vicinity of that of its closest relatives, between 5.66-7.42 Mbp. However, we also performed a genome-size calculation based on statistical inferences. We examined the trend using average, not total, contig length (fragments assembled using EGassembler [58]) for 16.5%, 33%, 66% and 100% of the total bacterial pyrosequencing fragment pool with the asymptotic end-point being predictive of full-genome size using Newton-Rhapson estimation on a non-linear general logistic equation [GENOME*(1-(1/e$^{(\mathrm{obs*CONSTANT})}$))]. The resulting predicted genome size of *R. parasitica* was 2.84 Mbp (Figure 5.5). This value corresponds with the reduced genomes evident in several other animal endosymbionts and would imply that *R. parasitica* displays at least one feature of the symbiont syndrome.

Sequencing the entire genome of *R. parasitica* should be the focus of future studies as it would also allow for insights into the full genomic makeup of the symbiont, including the functional consequences of the absence of genes, and the potential finding of more genes related to the endosymbiotic lifestyle of this non-parasitic, animal-inhabiting alphaproteobacterium.

Fig. 5.5. Estimation of the genome size of *Reichenowia parasitica* based on Newton-Rhapson estimation on a non-linear general logistic equation. Blue diamonds denote the general logistic equation with the asymptotic end-point being predictive of full genome size. Red squares denote the average contig size at 16.5%, 33%, 66% and 100% of the total bacterial pyrosequencing fragment pool, respectively. The estimated end-point and thus the full genome size is predicted at 2.84 Mbp.


3000
2840
2500
2000
1500
1000
500
0
ave. contig size nt X 10^3
0
50
100
150
200
250
fragments
◆ genome size
■ contig length

There are, of course, numerous ways of assembling and managing short sequence reads. Although the methods and results conveyed here are straight-forward, only a small subset of the bacterial contigs (n=358) were analyzed. We are currently exploring different, and possibly more efficient, ways of assembling the fragments andanalyzing the data, chiefly to identify the origin of the 9,866 contigs that did not have a match. However, it is our hope that the preliminary data shown here will serve as a stepping-stone for future studies of this intriguing symbiosis.

References

1. Graf J, Kikuchi Y, Rio RVM (2006) Leeches and their microbiota: naturally simple symbiosis models. Trends Microbiol 14: 365-371.

2. Reichenow E (1921) Über intrazelluläre Symbionten bei Blutsaugern. Arch Schiffs-u Tropen-Hyg 25.

3. Reichenow E (1922) Intrazelluläre Symbionten bei blutsaugenden Milben und Egeln Arch Protistenk 45.

4. Siddall ME, Perkins SL, Desser SS (2004) Leech mycetome endosymbionts are a new lineage of alphaproteobacteria related to the Rhizobiaceae. Mol Phylogenet Evol 30: 178-186.

5. Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, et al. (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinida*. Nat Genet 32: 402-407.

6. Perkins SL, Budinoff RB, Siddall ME (2005) New Gammaproteobacteria associated with blood-feeding leeches and a broad phylogenetic analysis of leech endosymbionts. Appl Environ Microbiol 71: 5219-5224.

7. Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet 42: 165-190.

8. Kikutchi Y, Fukatsu T (2002) Endosymbiotic bacteria in the esophageal organ of glossiphoniid leeches. Appl Environ Microbiol 68: 4637-4641.

9. Siddall ME, Min G-S, Fontanella FM, Phillips AJ, Watson SC (2011) Bacterial symbiont and salivary peptide evolution in the context of leech phylogeny. Parasitol 138: 1815-1827.

10. Moran NA, Wernegreen JJ (2000) Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol Evol 15: 321-326.

11. Andersson SGE, Kurland CG (1998) Reductive evolution of resident genomes. Trends Microbiol 6: 263-268.

12. Moreno E (1998) Genome evolution within the alpha *Proteobacteria*: why do some bacteria not possess plasmids and others exhibit more than one different chromosome? FEMS Micribiol Rev 22: 255-275.

13. Fischer HM (1994) Genetic regulation of nitrogen fixation in rhizobia. Microbiol Mol Biol Rev 58: 352-386.

14. Townsend AR, Howarth RW, Bazzaz FA, Booth MS, Cleveland CC, et al. (2003) Human health effects of a changing global nitrogen cycle. Frontiers Ecol Environ 1: 240-246.

15. Carpenter SR, Caraco NF, Correll DL, Howarth RW, Sharpley AN, et al. (1998) Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecol Appl 8: 559-568.

16. Howarth RW, Marino R, Cole JJ (1988) Nitrogen fixation in freshwater, estuarine, and marine ecosystems. 2. Biogeochemical controls. Limnol Oceanogr 33: 688-701.

17. Moran NA, Baumann P (2000) Bacterial endosymbionts in animals. Curr Opin Microbiol 3: 270-275.

18. Tamames J, González-Moreno M, Mingorance J, Valencia A, Vicente M (2001) Bringing gene order into bacterial shape. Trends Genet 17: 124-126.

19. Lefèvre C, Charles H, Vallier A, Delobel B, Farrell B, et al. (2004) Endosymbiont phylogenesis in the Dryophtoridae weevils: evidence for bacterial replacement. Mol Biol Evol 21: 965-973.

20. van Borm S, Buschinger A, Boomsma JJ, Billen J (2002) *Tetraponera* ants have gut symbionts related to nitrogen-fixing root-nodule bacteria. Proc R Soc Lond B 269: 2023-2027.

21. Fontana R, Canepari P, Satta G (1979) Alterations in peptidoglycan chemical composition associated with rod-to-sphere transition in a conditional mutant of *Klebsiella pneumoniae*. J Bacteriol 139: 1028-1038.

22. Henning U, Rehn K, Braun V, Hohn B (1972) Cell envelope and shape of *Escherichia coli* K12. Properties of a temperature-sensitive rod mutant. Eur J Biochem 26: 570-586.

23. Rogers GB, Carroll MP, Bruce KD (2009) Studying bacterial infections through culture-independent approaches. J Med Microbiol 58: 1401-1418.

24. Siddall ME, Budinoff RB, Borda E (2005) Phylogenetic evaluation of systematics and biogeography of the leech family Glossiphoniidae. Invertebr Syst 19: 105-112.

25. Light JE, Siddall ME (1999) Phylogeny of the leech family Glossiphoniidae based on mitochondrial gene sequences and morphological data. J Parasitol 85: 815-823.

26. Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucl Acids Res 26: 1107-1115.

27. Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, et al. (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics 22: 699-707.

28. Katoh K, Kuma K-I, Toh H, Miyata T (2005) MAFFT verison 5: improvement in accuracy of multiple sequence alignment. Nucl Acids Res 33: 511-518.

29. Goloboff PA, Farris JS, Nixon KC (2008) TNT, a free program for phylogenetic analysis. Cladistics 24: 774-786.

30. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

31. Siddall ME (2010) Unringing a bell: metazoan phylogenomics and the partition bootstrap. Cladistics 26: 444-452.

32. Swofford D (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods), ver. 4.0b10. Sunderland, MA: Sinauer Associates.

33. Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nat Genet 21: 108-110.

34. Ochman H, Moran NA (2001) Genes lost and found: evolution of bacterial pathogenesis and symbiosis. Science 292: 1096-1099.

35. Casjens S (1998) The diverse and dynamic structure of bacterial genomes. Annu Rev Genet 32: 339-377.

36. Röther D, Poppe L, Viergutz S, Langer B, Rétey J (2001) Characterization of the active site of histidine ammonia-lyase from *Pseudomonas putida*. Eur J Biochem 268: 6011-6019.

37. Weidner S, Pühler A, Küster H (2003) Genomic insights into symbiotic nitrogen fixation. Curr Opin Biotechnol 14: 200-205.

38. Kahn D, David M, Domergue O, Daveran ML, Ghai J, et al. (1989) *Rhizobium meliloti fixGHI* sequence predicts involvement of a specific cation pump in symbiotic nitrogen fixation. J Bacteriol 171: 929-939.

39. Batut J, Bostard P, Debelle F, Denarie J, Ghai J, et al (1985a) Developmental biology of the *Rhizobium meliloti*-alfalfa symbiosis: a joint genetic and cytological approach. In: Evans HJ, Bottomley PJ, Newton WE, editors. Nitrogen fixation research progress. Dordrecht: Martinus Nijhoff. pp. 109-115.

40. Batut J, Terzaghi B, Ghérardi M, Huguet M, Terzaghi E, et al. (1985b) Localization of a symbioticfix region on *Rhizobium meliloti* pSym megaplasmid more than 200 kilobases from the *nod-nif* region. Mol Gen Genet 19: 232-239.

41. Rubinstein B, Stern AI (1986) Relationship of transplasmalemma redox activity to proton and solute transport by roots of *Zea mays*. Plant Physiol 80: 805-811.

42. Lechene CP, Luyten Y, McMahon G, Distel DL (2007) Quantitative imaging of nitrogen fixation by individual bacteria within animal cells. Science 317: 1563-1566.

43. Higgins CF (2001) ABC transporters: physiology, structure and mechanism - an overview. Res Microbiol 152: 205-210.

44. Borths EL, Locher KP, Lee AT, Rees DC (2002) The structure of *Escherichia coli* BtuF and binding to its cognate ATP binding cassette transporter. Proc Natl Acad Sci U S A 99: 16642-16647.

45. Nogge G (1981) Significance of symbionts for the maintenance of an optimal nutritional state for successful reproduction in hematophagous arthropods. Parasitol 82: 101-104.

46. Nogge G (1976) Sterility in tsetse flies (*Glossina morsitans* Westwood) caused by loss of symbionts. Experientia 32: 995-996.

47. Lehnerr H, Yarmolinsky MB (1995) Addiction protein Phd of plasmid prophage P1 is a substrate of the ClpXP serine protease of *Escherichia coli*. Proc Natl Acad Sci U S A 92: 3274-3277.

48. Engelberg-Kulka H, Hazan R, Amitai S (2005) *mazEF*: a chromosomal toxin-antitoxin module that triggers programmed cell death in bacteria. J Cell Sci 118: 4327-4332.

49. Nekrasov SV, Agafonova OV, Belogurova NG, Delver EP, Belogurov AA (2007) Plasmid-encoded antirestriction protein ArdA can descriminate Type I methyltransferase and complete restriction-modification system. J Mol Biol 365: 284-297.

50. Kennaway CK, Obarska-Kosinska A, White JH, Tuszynska I, Cooper LP, et al. (2009) The structure of M.EcoKI Type I DNA methyltransferase with a DNA mimic antirestriction protein. Nucl Acids Res 37: 762-770.

51. Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, et al. (2002) Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. DNA Res 9: 189-197.

52. Spaepen S, Das F, Luyten E, Michiels J, Vanderleyden J (2009) Indole-3-acetic acid-regulated genes in *Rhizobium etli* CNPAF512. FEMS Microbiol Lett 291: 195-200.

53. Ausmees N, Jacobsson K, Lindberg M (2001) A unipolarly located, cell-surface-associated agglutinin, RapA, belongs to a family of Rhizobium-adhering proteins (Rap) in *Rhizobium leguminosarum* bv. *trifolii.* Microbiol-SGM 147: 549-559.

54. Bouchon D, Rigaud T, Juchault P (1998) Evidence for widespread *Wolbachia* infection in isopod crustaceans: molecular identification and host feminization. Proc R Soc Lond B 265: 1081-1090.

55. Burnett WJ, McKenzie JD (1997) Subcuticular bacteria from the brittle star *Ophiactis balli* (Echinodermata: Ophiuroidea) represent a new lineage of extracellular marine symbionts in the a subdivision of the class Proteobacteria. Appl Environ Microbiol 63: 1721-1724.

56. Manz W, Amann R, Ludwig W, Vancanneyt M, Schleifer K-H (1996) Application of a suite of 16S rRNA-specific oligonucleotide probes designed to investigate bacteria of the phylum cytophaga-flavobacter-bacteroides in the natural environment. Micriobiology 142: 1097-1106.

57. van Brussel AAN, Planqué K, Quispel A (1977) The wall of *Rhizobium leguminosarum* in bacteroid and free-living forms. J Gen Microbiol 101: 51-56.

58. Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, et al. (2006) EGassembler: online bioinformatics service for large-scale processing, clustering and assembling EST's and genomic DNA fragments. Nucl Acids Res 34 (suppl. 2):W459-W462.

CHAPTER VI

GENERAL CONCLUSIONS

Whereas each chapter in this dissertation contains several results of a peripheral nature, which are not discussed here, in the following I outline the main general conclusions concerning each chapter:

- Using the most comprehensive molecular character set to date, parsimony analysis concerning the phylogenetic relationships within Annelida and between the phylum and its constituent taxa are at odds with previous Bayesian and Maximum Likelihood estimations. The newly re-erected taxa Sedentaria and Errantia are not recovered as monophyletic by parsimony methods and the taxonomy should be re-evaluated in light of these new findings.
- Insofar as this large data set will likely be used in future studies of evolutionary histories of annelid worms, users should note that substantial redundancy does exist in the data set, possibly giving rise to artificial clades or artificial support for clades.
- The glossiphoniid leech *Helobdella robusta* does possess leech antiplatelet protein despite of its non-bloodfeeding nature. Whereas the function of these proteins can only be speculated upon, several of the orthologous *H. robusta*-sequences possess a signal peptide-region, indicating their secretion by the leech. Because the family Glossiphoniidae is commonly recovered at the base of the leech phylogeny, it is likely

that *H. robusta* possesses anticoagulants by virtue of their presence in a common ancestor of all leeches.

- The diversity of anticoagulation factors expressed in the salivary glands of the European medicinal leech *Hirudo verbana* and the African medicinal leech *Aliolimnatis fenestrata* is equivalent to that previously found in the North American medicinal leech *Macrobdella decora*. In total, *A. fenestrata* was shown to possess 11 anticoagulants in addition to several other bioactive peptides and *H. verbana* was shown to also possess seven of these, in addition to other bioactive peptides.

- Selection pressures acting on the anticoagulants within the salivary transcriptomes of *Hirudo verbana, Aliolimnatis fenestrata* and *Macrobdella decora* are mostly of a purifying nature. The importance of the anticoagulant proteins for the leeches seems to be manifested in the high levels of purifying selection, presumably acting on the genes in order to keep them intact.

- Overall, there is little concordance between the evolutionary histories of the anticoagulants and previous hypotheses of the leech phylogeny. Regardless of this, however, the origins of each of hirudin, bdellin, various antistasin-family proteins, eglin c and endoglucuronidases predate the origins of the medicinal leeches considered here.

- The leech-endosymbiotic alphaproteobacterium *Reichenowia parasitica* possesses genes coding for proteins related to nitrogen fixation, iron/vitamin B translocation and plasmid survival. Our results also indicate that *R. parasitica* interacts with its host in part by transmembrane signaling.

- Phylogenetic analyses of 358 loci across Bacteria support the nesting of *R. parasitica* within the Rhizobiaceae, as sister to a group containing *Agrobacterium* and *Rhizobium* species.

APPENDIX A:

SUPPLEMENTARY MATERIAL FOR CHAPTER V:

PHYLOGENOMICS OF *REICHENOWIA PARASITICA*, AN ALPHAPROTEOBACTERIAL ENDOSYMBIONT OF THE FRESHWATER LEECH *PLACOBDELLA PARASITICA*

Supplementary Table S5.1. List of species used for subtractive scaffolding, orthologue recovery and phylogenetic analysis. Bold font denotes the non-alphaproteobacterial species. GenBank RefSeq refers to the submission inclusive of the entire genome.

Species	**Class**	**Family**	**GenBank RefSeq**	**Estimated Genome Size (Mbp)**
Acetobacter pasteurianus	Alphaproteobacteria	Acetobacteraceae	NC_013209.1	3.328
Agrobacterium radiobacter	Alphaproteobacteria	Rhizobiaceae	NC_011985.1	7.26491
Agrobacterium tumefaciens	Alphaproteobacteria	Rhizobiaceae	NC_00306.2	5.66716
Agrobacterium vitis	Alphaproteobacteria	Rhizobiaceae	NC_011989.1	6.33537
Aquifex aeolicus	Aquificae	Aquificaceae	NC_000918.1	1.59079
Azorhizobium caulinodans	Alphaproteobacteria	Xanthobacteraceae	NC_009937.1	5.36977
Bacillus anthracis	Firmicutes	Bacillaceae	NC_007530.2	5.50242
Bartonella grahamii	Alphaproteobacteria	Bartonellaceae	NC_012846.1	2.36933
Bartonella henselae	Alphaproteobacteria	Bartonellaceae	NC_005956.1	1.93105
Bartonella quintana	Alphaproteobacteria	Bartonellaceae	NC_005955.1	1.58138
Bdellovibrio bacteriovorus	Deltaproteobacteria	Bdellovibrionaceae	NC_005363.1	3.8
Beijerinckia indica	Alphaproteobacteria	Beijerinickiaceae	NC_01058.1	4.41715
Bradyrhizobium japonicum	Alphaproteobacteria	Bradyrhizobiaceae	NC_004463.1	9.1
Brucella abortus	Alphaproteobacteria	Brucellaceae	NC_006932.1	3.28645
Brucella canis	Alphaproteobacteria	Brucellaceae	NC_010103.1	3.3
Brucella melitensis	Alphaproteobacteria	Brucellaceae	NC_003317.1	3.27779
Brucella suis	Alphaproteobacteria	Brucellaceae	NC_004310.3	3.3
Buchnera aphidicola	Gammaproteobacteria	Enterobacteriaceae	NC_002528.1	0.655086
Campylobacter concisus	Epsilonproteobacteria	Campylobacteraceae	NC_009802.1	2.09901
Candidatus Carsonella rudii	Gammaproteobacteria	-	NC_008512.1	0.16
Candidatus Sulcia muelleri	Bacteroidetes	-	NC_014004.1	0.24
Caulobacter crescentus	Alphaproteobacteria	Caulobacteraceae	NC_002696.2	4
Chromobacterium violaceum	Betaproteobacteria	Neisseriaceae	NC_005085.1	4.75108
Ehrlichia canis	Alphaproteobacteria	Anaplasmataceae	NC_007354.1	1.3
Ehrlichia chaffeensis	Alphaproteobacteria	Anaplasmataceae	NC_007799.1	1.17625
Erythrobacter litoralis	Alphaproteobacteria	Erythrobacteraceae	NC_007722.1	3.0524
Gluconobacter oxydans	Alphaproteobacteria	Acetobacteraceae	NC_006677.1	2.92021
Jannaschia sp.	Alphaproteobacteria	Rhodobacteraceae	NC_007802.1	4.386
Mesorhizobium loti	Alphaproteobacteria	Phyllobacteriaceae	NC_002678.1	7.5963
Methylobacterium chloromethanicum	Alphaproteobacteria	Methylobacteriaceae	NC_011757.1	6.18091
Methylobacterium extorquens	Alphaproteobacteria	Methylobacteriaceae	NC_012808.1	6.86846
Nitrobacter hamburgensis	Alphaproteobacteria	Bradyrhizobiaceae	NC_007964.1	5
Paracoccus denitrificans	Alphaproteobacteria	Rhodobacteraceae	NC_008686.1	5.23238
Prochlorococcus marinus	Cyanobacteria	Prochlorococcaceae	NC_009091.1	1.64188

Supplementary Table S5.1 (Continued)

Rhizobium etli	Alphaproteobacteria	Rhizobiaceae	NC_010994.1	6.44
Rhizobium leguminosarum	Alphaproteobacteria	Rhizobiaceae	NC_008380.1	7.74714
Rhodobacter capsulatus	Alphaproteobacteria	Rhodobacteraceae	NC_014034.1	3.83
Rhodobacter sphaeroides	Alphaproteobacteria	Rhodobacteraceae	NC_007494.1	4.607
Rhodopseudomonas palustris	Alphaproteobacteria	Bradyrhizobiaceae	NC_008435.1	5.5
Rickettsia conorii	Alphaproteobacteria	Rickettsiaceae	NC_003103.1	1.26876
Rickettsiaprowazekii	Alphaproteobacteria	Rickettsiaceae	NC_000963.1	1.1
Rickettsia ricketsii	Alphaproteobacteria	Rickettsiaceae	NC_009882.1	1.25771
Rickettsia typhi	Alphaproteobacteria	Rickettsiaceae	NC_006142.1	1.1115
Ruegeria pomeroyi	Alphaproteobacteria	Rhodobacteraceae	NC_003911.11	4.59
Sinorhizobium fredii	Alphaproteobacteria	Rhizobiaceae	NC_012587.1	6.89574
Sinorhizobium medicae	Alphaproteobacteria	Rhizobiaceae	NC_009636.1	6.83636
Sinorhizobium meliloti	Alphaproteobacteria	Rhizobiaceae	NC_003047.1	6.70836
Wigglesworthia glossinida	Gammaproteobacteria	Enterobacteriaceae	NC_004344.2	0.7053
Wolbachia endosymbiont of D. melanogaster	Alphaproteobacteria	Anaplasmataceae	NC_002978.6	1.26778
Zymomonas mobilis	Alphaproteobacteria	Sphingomonadaceae	NC_006526.2	2.1986

Supplementary Table S5.2. Description of BLASTn Hits Encountered using the *Reichenowia parasitica* Contigs as Queries Against 50 Selected Bacterial Genomes. All hits matched at 1E^{-5} or lower. Hit descriptions follow the GenBank annotations for the genes, and the hit-taxon is shown in brackets.

No. of hits in *R. parasitica*	**E-value range**	**Hit description**
1156	0-1E^{-5}	(2Fe-2S)-binding domain protein [Methylobacterium chloromethanicum CM4]
120	1E^{-128}-4E^{-6}	2-hydroxychromene-2-carboxylate isomerase protein [Rhizobium etli CFN 42]
1182	0-1E^{-5}	2-oxoisovalerate dehydrogenase beta subunit [Bradyrhizobium japonicum USDA 110]
1308	0-1E^{-5}	3-demethylubiquinone-9 3-methyltransferase [Azorhizobium caulinodans ORS 571]
274	0-9E^{-6}	ABC transporter membrane spanning protein (dipeptide) [Agrobacterium vitis S4]
311	1E^{-136}-9E^{6}	aliphatic sulphonate ABC transporter [Agrobacterium radiobacter K84]
938	0-1E^{-5}	amidophosphoribosyltransferase [Brucella canis ATCC 23365]
67	4E^{-40}-1E^{-5}	amino-acid ABC transporter binding protein [Bartonella quintana str. Toulouse]
1160	0-1E^{-5}	aminotransferase, class I [Brucella abortus bv. 1 str. 9-941]
125	1E^{-155}-1E^{-6}	antirestriction protein [Agrobacterium vitis S4]
63	6E^{-33}-8E^{-6}	apolipoprotein N-acyltransferase [Bartonella grahamii as4aup]
957	0-1E^{-5}	AraC family transcriptional regulator [Brucella melitensis bv. 1 str. 16M]
551	1E^{-166}-9E^{-6}	asparagine synthetase [glutamine-hydrolyzing] [Gluconobacter oxydans 621H]
1972	0-9E^{-6}	autoaggregation protein (adhering protein) [Rhizobium etli CFN 42]
386	4E^{-74}-1E^{-5}	bacteriophage tail fiber protein [Chromobacterium violaceum ATCC 12472]
1039	0-1E^{-5}	branched-chain amino acid ABC transporter periplasmic branched-chain amino acid-binding protein LivK [Rhodobacter capsulatus SB 1003]
34	0-4E^{-6}	cation pump membrane (nitrogen fixation)protein [Rhizobium etli CFN 42]
342	0-1E^{-5}	co-chaperonin GroES [Brucella abortus bv. 1 str. 9-941]
8	5E^{-20}-6E^{-6}	copper tolerance protein [Agrobacterium vitis S4]
99	2E^{-57}-9E^{-6}	cystathionine beta-lyase [Zymomonas mobilis subsp. mobilis NCIB 11163]
1	9E^{-6}	cytochrome c oxidase assembly protein [Wolbachia endosymbiont of Drosophila melanogaster]
4	4E^{-31}-7E^{-6}	cytochrome c oxidase, subunit I [Nitrobacter hamburgensis X14]
138	1E^{-154}-7E^{-6}	cytochrome O ubiquinol oxidase, subunit III protein [Rhizobium etli CFN 42]
265	0-1E^{-5}	D-amino acid dehydrogenase small subunit [Brucella suis 1330]
178	0-8E^{-6}	deoxyribose-phosphate aldolase/phospho-2-dehydro-3-deoxyheptonate aldolase [Rhizobium leguminosarum bv. trifolii WSM1325]
314	0-9E^{-6}	endoglucanase precursor [Rhizobium sp. NGR234]
871	1E^{-169}-1E^{-5}	extracellular solute-binding protein [Beijerinckia indica subsp. indica ATCC 9039]

Supplementary Table S5.2 (Continued)

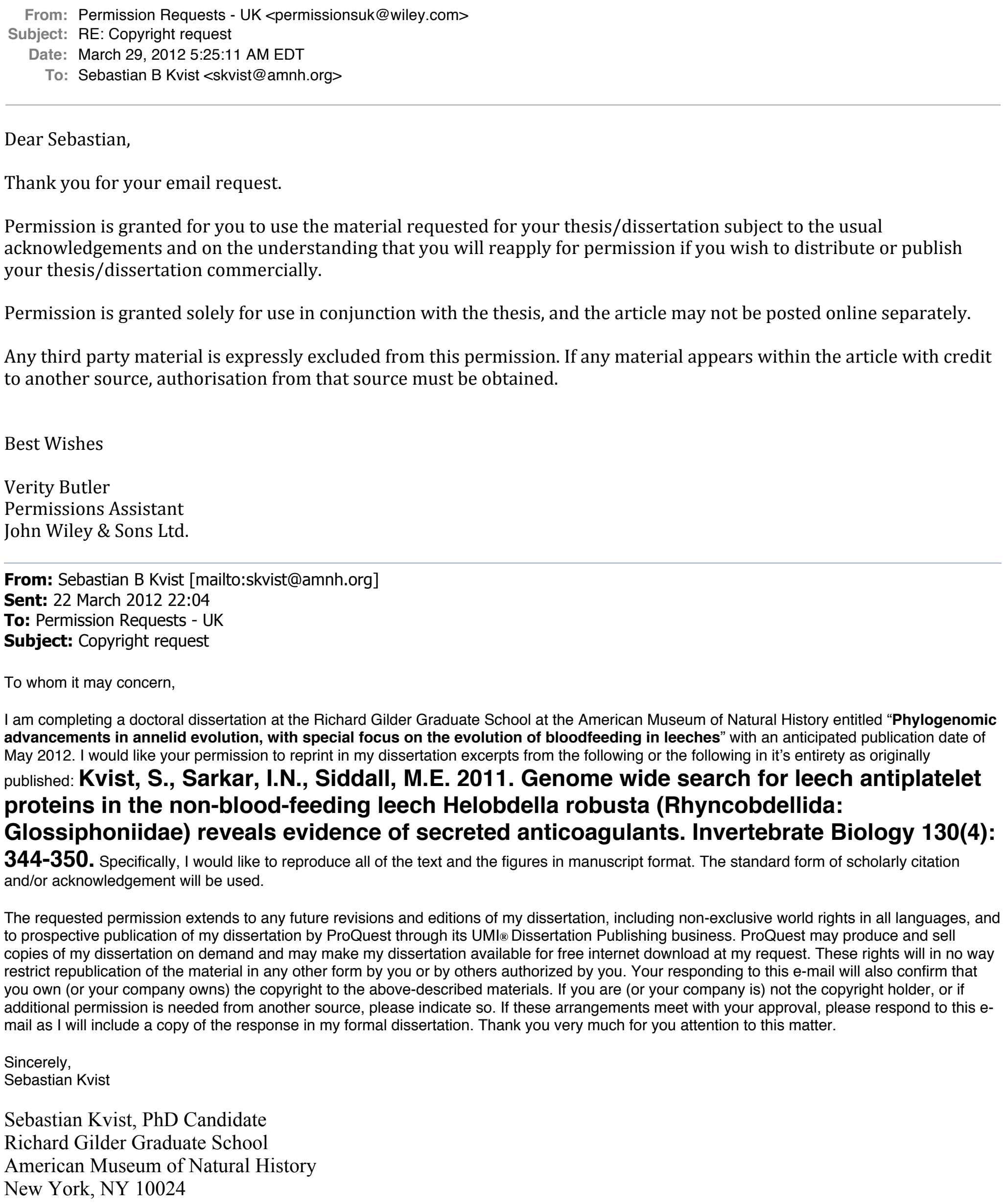
296	0-1E^{-5}	flagellar biosynthesis protein FliR [Brucella canis ATCC 23365]
267	0-9E^{-6}	flagellar biosynthesis protein FliR [Brucella melitensis bv. 1 str. 16M]
2003	0-1E^{-5}	glucokinase [Agrobacterium radiobacter K84]
35	0-5E^{-6}	glycosy hydrolase family protein [Mesorhizobium loti MAFF303099]
91	2E^{-59}-9E^{-6}	GntR family transcriptional regulator [Ruegeria pomeroyi DSS-3]
180	1E^{-111}-9E^{-6}	hemolysin-type calcium-binding region, RTX [Rhodobacter sphaeroides 2.4.1]
1100	0-1E^{-5}	histidine ammonia-lyase [Rhodobacter sphaeroides 2.4.1]
1	3E^{-6}	H-NS family DNA-binding protein [Rhodobacter sphaeroides 2.4.1]
88	7E^{-77}-5E^{-6}	hypothetical 22.1 kDa periplasmic protein [Rhizobium sp. NGR234]
4	6E^{-8}-3E^{-6}	hypothetical protein A1G_03905 [Rickettsia rickettsii str. 'Sheila Smith']
6	1E^{-16}-1E^{-6}	hypothetical protein aq_1163 [Aquifex aeolicus VF5]
123	1E^{-172}-3E^{-6}	hypothetical protein Arad_12331 [Agrobacterium radiobacter K84]
483	0-1E^{-5}	hypothetical protein Atu4528 [Agrobacterium tumefaciens str. C58]
45	2E^{-32}-2E^{-6}	hypothetical protein Atu6049 [Agrobacterium tumefaciens str. C58]
90	7E^{-50}-8E^{-6}	hypothetical protein Atu8047 [Agrobacterium tumefaciens str. C58]
1654	0-1E^{-5}	hypothetical protein Atu8164 [Agrobacterium tumefaciens str. C58]
1790	0-1E^{-5}	hypothetical protein Avi_2578 [Agrobacterium vitis S4]
3	1E^{-144}-4E^{-6}	hypothetical protein Avi_9567 [Agrobacterium vitis S4]
36	4E^{-40}-9E^{-6}	hypothetical protein BH11390 [Bartonella henselae str. Houston-1]
31	1E^{-11}-8E^{-6}	hypothetical protein Bind_3812 [Beijerinckia indica subsp. indica ATCC 9039]
607	0-1E^{-5}	hypothetical protein ELI_07840 [Erythrobacter litoralis HTCC2594]
797	1E^{-155}-1E^{-5}	hypothetical protein Jann_4099 [Jannaschia sp. CCS1]
58	1E^{-108}-6E^{-6}	hypothetical protein MexAM1_META2p0295 [Methylobacterium extorquens AM1]
1985	0-1E^{-5}	hypothetical protein mll4271 [Mesorhizobium loti MAFF303099]
3	1E^{-29}-3E^{-6}	hypothetical protein mll9560 [Mesorhizobium loti MAFF303099]
1962	0-1E^{-5}	hypothetical protein NGR_c14070 [Rhizobium sp. NGR234]
392	1E^{-173}-9E^{-6}	hypothetical protein Pden_3905 [Paracoccus denitrificans PD1222]
165	1E^{-78}-8E^{-6}	hypothetical protein Pden_4702 [Paracoccus denitrificans PD1222]
43	5E^{-60}-2E^{-6}	hypothetical protein Rleg_5777 [Rhizobium leguminosarum bv. trifolii WSM1325]
155	1E^{-133}-5E^{-6}	hypothetical protein Rleg_6295 [Rhizobium leguminosarum bv. trifolii WSM1325]
18	2E^{-25}-8E^{-6}	hypothetical protein RSP_3910 [Rhodobacter sphaeroides 2.4.1]
291	0-9E^{-6}	hypothetical protein SM_b20011 [Sinorhizobium meliloti 1021]
1895	0-1E^{-5}	hypothetical protein Smed_0776 [Sinorhizobium medicae WSM419]
948	0-1E^{-5}	hypothetical protein SPO3085 [Ruegeria pomeroyi DSS-3]
83	1E^{125}-5E^{-6}	inner-membrane translocator [Rhizobium leguminosarum bv. trifolii WSM1325]

Supplementary Table S5.2 (Continued)

19	$3E^{-16}$-$4E^{-6}$	iron siderophore/cobalamin ABC transporter periplasmic iron siderophore/cobalamin-binding protein [Rhodobacter capsulatus SB 1003]
1	$7E^{-15}$	isochorismatase family protein [Bacillus anthracis str. A0248]
6	$3E^{-41}$-$1E^{-10}$	K potassium transporter [Sinorhizobium medicae WSM419]
553	0-$9E^{-6}$	LacI family transcription regulator [Caulobacter crescentus CB15]
7	$9E^{-13}$-$8E^{-6}$	LysR family transcriptional regulator [Agrobacterium radiobacter K84]
162	0-$9E^{-6}$	NuoK2 NADH:quinone oxidoreductase subunit 11 (chain K) [Sinorhizobium meliloti 1021]
56	$3E^{-25}$-$1E^{-5}$	oxidoreductase [Agrobacterium vitis S4]
1975	0-$1E^{-5}$	polyhydroxyalkonate synthesis repressor, PhaR [Rhizobium leguminosarum bv. trifolii WSM1325]
26	$2E^{-21}$-$1E^{-5}$	prevent-host-death family protein [Methylobacterium chloromethanicum CM4]
1167	0-$1E^{-5}$	putative glycohydrolase [Rhodopseudomonas palustris BisA53]
116	$1E^{-156}$-$8E^{-6}$	putative glyoxalase protein [Rhizobium etli CFN 42]
1896	0-$9E^{-6}$	putative hydantoin racemase protein [Sinorhizobium meliloti 1021]
169	$1E^{-31}$-$9E^{-6}$	putative phosphatase [Bdellovibrio bacteriovorus HD100]
122	$1E^{-140}$-$1E^{-5}$	quinolinate synthetase complex, A subunit [Rhizobium leguminosarum bv. trifolii WSM1325]
915	0-$1E^{-5}$	resolvase family site-specific recombinase [Brucella suis 1330]
3	$4E^{-7}$-$3E^{-6}$	ribonuclease PH [Rickettsia conorii str. Malish 7]
2	$1E^{-130}$-$6E^{-6}$	serine hydroxymethyltransferase [Rickettsia typhi str. Wilmington]
2	$2E^{-6}$-$3E^{-6}$	single-strand binding protein [Nitrobacter hamburgensis X14]
192	0-$6E^{-6}$	sugar ABC transporter, ATP binding protein [Rhizobium etli CFN 42]
710	0-$1E^{-5}$	TonB-dependent receptor, plug [Paracoccus denitrificans PD1222]
10	$2E^{-26}$-$4E^{-6}$	transposase [Agrobacterium vitis S4]
920	0-$9E^{-6}$	transposase IS3/IS911 [Nitrobacter hamburgensis X14]
28	$9E^{-83}$-$9E^{-6}$	universal stress protein [Rhodobacter sphaeroides 2.4.1]
253	$1E^{-179}$-$9E^{-6}$	uracil-xanthine permease [Sinorhizobium medicae WSM419]
110	0-$9E^{-6}$	UvrD/REP helicase [Sinorhizobium medicae WSM419]
4	$1E^{-19}$-$3E^{-9}$	virulence VirF1 protein [Rhizobium etli CFN 42]
1010	0-$9E^{-6}$	XoxF, PQQ-linked dehydrogenase of unknown function [Methylobacterium extorquens AM1]

APPENDIX B:

COPYRIGHT RELEASE FORMS FOR ALREADY PUBLISHED MATERIAL



From: Permission Requests - UK <permissionsuk@wiley.com>
Subject: RE: Copyright request
Date: March 29, 2012 5:25:11 AM EDT
To: Sebastian B Kvist <skvist@amnh.org>

Dear Sebastian,

Thank you for your email request.

Permission is granted for you to use the material requested for your thesis/dissertation subject to the usual acknowledgements and on the understanding that you will reapply for permission if you wish to distribute or publish your thesis/dissertation commercially.

Permission is granted solely for use in conjunction with the thesis, and the article may not be posted online separately.

Any third party material is expressly excluded from this permission. If any material appears within the article with credit to another source, authorisation from that source must be obtained.

Best Wishes

Verity Butler
Permissions Assistant
John Wiley & Sons Ltd.

From: Sebastian B Kvist [mailto:skvist@amnh.org]
Sent: 22 March 2012 22:04
To: Permission Requests - UK
Subject: Copyright request

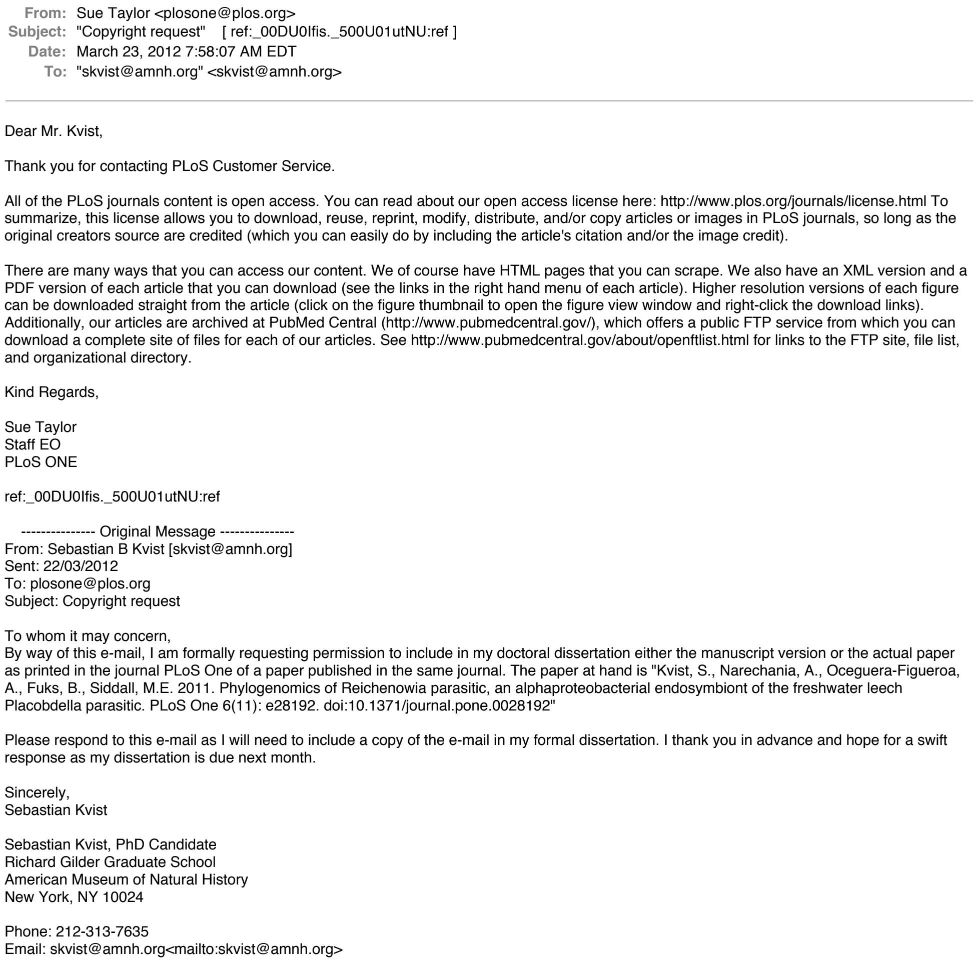
To whom it may concern,

I am completing a doctoral dissertation at the Richard Gilder Graduate School at the American Museum of Natural History entitled "**Phylogenomic advancements in annelid evolution, with special focus on the evolution of bloodfeeding in leeches**" with an anticipated publication date of May 2012. I would like your permission to reprint in my dissertation excerpts from the following or the following in it's entirety as originally published: **Kvist, S., Sarkar, I.N., Siddall, M.E. 2011. Genome wide search for leech antiplatelet proteins in the non-blood-feeding leech Helobdella robusta (Rhyncobdellida: Glossiphoniidae) reveals evidence of secreted anticoagulants. Invertebrate Biology 130(4): 344-350.** Specifically, I would like to reproduce all of the text and the figures in manuscript format. The standard form of scholarly citation and/or acknowledgement will be used.

The requested permission extends to any future revisions and editions of my dissertation, including non-exclusive world rights in all languages, and to prospective publication of my dissertation by ProQuest through its UMI® Dissertation Publishing business. ProQuest may produce and sell copies of my dissertation on demand and may make my dissertation available for free internet download at my request. These rights will in no way restrict republication of the material in any other form by you or by others authorized by you. Your responding to this e-mail will also confirm that you own (or your company owns) the copyright to the above-described materials. If you are (or your company is) not the copyright holder, or if additional permission is needed from another source, please indicate so. If these arrangements meet with your approval, please respond to this e-mail as I will include a copy of the response in my formal dissertation. Thank you very much for you attention to this matter.

Sincerely,
Sebastian Kvist

Sebastian Kvist, PhD Candidate
Richard Gilder Graduate School
American Museum of Natural History
New York, NY 10024



From: Sue Taylor <plosone@plos.org>
Subject: "Copyright request" [ref:_00DU0Ifis._500U01utNU:ref]
Date: March 23, 2012 7:58:07 AM EDT
To: "skvist@amnh.org" <skvist@amnh.org>

Dear Mr. Kvist,

Thank you for contacting PLoS Customer Service.

All of the PLoS journals content is open access. You can read about our open access license here: http://www.plos.org/journals/license.html To summarize, this license allows you to download, reuse, reprint, modify, distribute, and/or copy articles or images in PLoS journals, so long as the original creators source are credited (which you can easily do by including the article's citation and/or the image credit).

There are many ways that you can access our content. We of course have HTML pages that you can scrape. We also have an XML version and a PDF version of each article that you can download (see the links in the right hand menu of each article). Higher resolution versions of each figure can be downloaded straight from the article (click on the figure thumbnail to open the figure view window and right-click the download links). Additionally, our articles are archived at PubMed Central (http://www.pubmedcentral.gov/), which offers a public FTP service from which you can download a complete site of files for each of our articles. See http://www.pubmedcentral.gov/about/openftlist.html for links to the FTP site, file list, and organizational directory.

Kind Regards,

Sue Taylor
Staff EO
PLoS ONE

ref:_00DU0Ifis._500U01utNU:ref

--------------- Original Message ---------------
From: Sebastian B Kvist [skvist@amnh.org]
Sent: 22/03/2012
To: plosone@plos.org
Subject: Copyright request

To whom it may concern,
By way of this e-mail, I am formally requesting permission to include in my doctoral dissertation either the manuscript version or the actual paper as printed in the journal PLoS One of a paper published in the same journal. The paper at hand is "Kvist, S., Narechania, A., Oceguera-Figueroa, A., Fuks, B., Siddall, M.E. 2011. Phylogenomics of Reichenowia parasitic, an alphaproteobacterial endosymbiont of the freshwater leech Placobdella parasitic. PLoS One 6(11): e28192. doi:10.1371/journal.pone.0028192"

Please respond to this e-mail as I will need to include a copy of the e-mail in my formal dissertation. I thank you in advance and hope for a swift response as my dissertation is due next month.

Sincerely,
Sebastian Kvist

Sebastian Kvist, PhD Candidate
Richard Gilder Graduate School
American Museum of Natural History
New York, NY 10024

Phone: 212-313-7635
Email: skvist@amnh.org<mailto:skvist@amnh.org>