

## AN ABSTRACT OF THE THESIS OF

Erin K. van Olden for the degree of Master of Science in Genetics presented on June 6, 2003.

Title: Comparison of *Octopus dofleini* Hemocyanin Paralogues and Evidence for Recent Divergence.

Abstract approved: <sup>Redacted for Privacy</sup> ^

---

Kensal E. van Holde

Presented here is an in depth comparison of the *Octopus dofleini* hemocyanin paralogues, designated A and G. The protein sequences coded by these genes are each comprised of seven oxygen-binding units, or functional units (FU-a to FU-g). Each FU is approximately 400 amino acids in length and a linker region separates each of them. Within every linker, a phase one intron (found between the first and second base of an amino acid codon) was found which varies from 100-910 base pairs in length. Three FUs are split by internal introns, dividing FU-b, FU-e, and FU-f into two exons. These internal introns show no phase pattern like the linker introns and appear to represent later insertion events. The intron within FU-e houses a microsatellite region that contains di- and tetra- nucleotide tandem repeats. Variation between the two *Octopus* paralogues is greatest in this microsatellite region, however the similarity is still striking. Given the higher rate of mutation for microsatellites, due mainly to slippage, this similarity either lends support to the recent divergence of these two paralogues or implies some kind of secondary-structure related function that has slowed the divergence of these microsatellites. The high degree of similarity in the

coding regions, as well as the non-coding regions, of this gene is evidence that suggests these paralogues have recently diverged. Comparisons between the two *Octopus* sequences and comparisons to related hemocyanins will help uncover the evolutionary origins of this gene.

©Copyright by Erin K. van Olden  
June 6, 2003  
All Rights Reserved

Comparison of *Octopus dofleini* Hemocyanin Paralogues and Evidence for Recent Divergence

By  
Erin K. van Olden

A THESIS

Submitted to  
Oregon State University

in partial fulfillment of  
the requirements for the  
degree of

Master of Science

Presented June 6, 2003  
Commencement June 2004

Master of Science thesis of Erin K. van Olden presented on June 6, 2003.

APPROVED:

*Redacted for Privacy*

---

Major Professor, representing Genetics

*Redacted for Privacy*

---

Head of Genetics Program

*Redacted for Privacy*

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

*Redacted for Privacy*

---

Erin K. van Olden, Author

## ACKNOWLEDGEMENTS

In Honor of my grandfather, Murhl C. Turley, Jr. I got it all from his side.

I would like to express my thanks to Karen I. Miller and Ken E. van Holde, their enthusiasm for life and child-like wonder at the world and all its workings has been an inspiration to me. Their tireless patience and loving guidance have make these four years a blessing. Thank You.

To Sandy J. van Olden and Hans A. van Olden, my constant supporters no matter how hard I've had to lean. I love you, Mom and Dad.

Thank you, also, to Alex Vincent, Andrea Warrick, Frank Hays, and Kristin Rorrer. You all have made my lab experience a wonderful one.

And to Darren S. Brown, without whom I am only half a person.

## TABLE OF CONTENTS

	<u>Page</u>
Chapter 1: Introduction	
1.1 Origin of Oxygen-Binding Proteins	1
1.2 Molluscan versus Arthropod Hemocyanin	2
1.3 Molluscan Hemocyanin	4
Chapter 2: Materials & Methods	
2.1 Obtaining <i>Octopus dofleini</i> DNA and sequencing	9
2.2 Population Genetics Study	11
Chapter 3: Results & Discussion	
3.1 A- and G-types are Paralogous Genes	12
3.2 Microsatellites in the A- and G-type <i>Octopus dofleini</i> Hemocyanin	19
3.3 Evolution	22
Chapter 4: Conclusions	
4.1 Paralogous Genes	44
4.2 A- and G-type Microsatellites in <i>Octopus dofleini</i> Hemocyanin	44
4.3 Comparisons Among Corresponding Functional Units	45
Bibliography	47

## LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
1. Basic Structure of Arthropod and Molluscan Hemocyanin	3
2. <i>Octopus dofleini</i> Hemocyanin decamer	4
3. The <i>Octopus dofleini</i> Protein Subunit and Gene	5
4. High Pressure Liquid Chromatography (HPLC) results show two distinct peaks for <i>O. dofleini</i>	6
5. The <i>Haliotis tuberculata</i> Gene and the <i>Octopus dofleini</i> Gene	7
6. ABI Prism Staden File with overlapping sequence	12
7. Population Genetics PCR Primer Locations	16
8. Theoretical Gel Results for the PGen Primer Sets	16
9. Actual Gel versus Theoretical Gel for A- & G-type	17
10. <i>Octopus dofleini</i> microsatellites compared	20
11. A- and G-type Microsatellites with (GT) <sub>n</sub> Repeat and Flanking Regions	21
12. Comparison of Nucleic Acid sequences for Exons, Linker Introns, and Internal Introns for the <i>Octopus dofleini</i> hemocyanin paralogues	23
13. Percent Identity Among the Exons, Linker Introns, and Internal Introns for the <i>Octopus dofleini</i> Hemocyanin Paralogues	24
14. Hypothetical Formation of the Current Hemocyanin Gene Structure	25
15. Line-up of Seven Molluscan Hemocyanin Subunits	26
16. Estimated Time of Divergence for Several Taxonomic Groups	42
17. Divergence Tree	43



## LIST OF TABLES

<u>Table</u>	<u>Page</u>
1. PCR Thermocycler Programs	10
2. Number and size of PCR fragments resulting from the various primer pairs, if the octopus has both the A- and G-type genes	17
3. <i>Octopus dofleini</i> (Od) Functional Unit Percent Identity: OdA versus OdA and OdG versus OdG	28
4. <i>Haliotis tuberculata</i> (Ht) Functional Unit Percent Identity: Ht1 versus Ht1 and Ht2 versus Ht2	29
5. <i>Octopus dofleini</i> Functional Unit Percent Identity: A-type versus G-type	30
6. <i>Haliotis tuberculata</i> (Ht) Functional Unit Percent Identity: type 1 versus type 2	31
7. Functional Unit Percent Identity: <i>Haliotis tuberculata</i> type 1 (Ht1) versus <i>Octopus dofleini</i> A-type (OdA)	32
8. Functional Unit Percent Identity: <i>Haliotis tuberculata</i> type 1 (Ht1) versus <i>Octopus dofleini</i> G-type (OdG)	33
9. Functional Unit Percent Identity: <i>Haliotis tuberculata</i> type 2 (Ht2) versus <i>Octopus dofleini</i> A-type (OdA)	34
10. Functional Unit Percent Identity: <i>Haliotis tuberculata</i> type 2 (Ht2) versus <i>Octopus dofleini</i> G-type (OdG)	35
11. Functional Unit-c Percent Identity: <i>Haliotis tuberculata</i> type 1 (Ht1), <i>Haliotis tuberculata</i> type 2 (Ht2), <i>Octopus dofleini</i> A-type (OdA), <i>Octopus dofleini</i> G-type (OdG), and <i>Megathura crenulata</i> (Mc)	36
12. Functional Unit-d Percent Identity: <i>Haliotis tuberculata</i> type 1 (Ht1), <i>Haliotis tuberculata</i> type 2 (Ht2), <i>Octopus dofleini</i> A-type (OdA), <i>Octopus dofleini</i> G-type (OdG), and <i>Helix pomatia</i> (Hp)	37

## LIST OF TABLES (cont.)

<u>Table</u>	<u>Page</u>
13. Functional Unit-g Percent Identity: <i>Haliotis tuberculata</i> type 1 (Ht1), <i>Haliotis tuberculata</i> type 2 (Ht2), <i>Octopus dofleini</i> A-type (OdA), <i>Octopus dofleini</i> G-type (OdG), <i>Helix pomatia</i> (Hp), and <i>Sepia officinalis</i> (So, FU-h)	38
14. Functional Unit Percent Identity: <i>Haliotis tuberculata</i> type 1 (Ht1), <i>Haliotis tuberculata</i> type 2 (Ht2), <i>Octopus dofleini</i> A-type (OdA), <i>Megathura crenulata</i> (Mc), <i>Helix pomatia</i> (Hp), and <i>Sepia officinalis</i> (So, FU-h)	39
15. Functional Unit Percent Identity: <i>Haliotis tuberculata</i> type 1 (Ht1), <i>Haliotis tuberculata</i> type 2 (Ht2), <i>Octopus dofleini</i> G-type (OdG), <i>Megathura crenulata</i> (Mc), <i>Helix pomatia</i> (Hp), and <i>Sepia officinalis</i> (So, FU-h)	40

## Comparison of *Octopus dofleini* Hemocyanin Paralogues and Evidence for Recent Divergence

### 1. Introduction

#### 1.1 Origin of Oxygen-Binding Proteins

Between one and two billion years ago, the earth gained an oxygen atmosphere and many organisms began to utilize oxygen for energy. Small organisms, 1-2 mm in size had no need for oxygen transport. They could get all the oxygen needed for aerobic metabolism via diffusion. As the size of organisms increased, diffusion was no longer a viable option. The pressure for a more efficient way of getting oxygen began to arise. The current variety of oxygen carrying proteins found in nature suggests that several oxygen transport proteins evolved to meet this need. The three major classes of oxygen-carrying proteins are; the hemoglobins-found in vertebrates, annelids, and some arthropods, the hemerythrins-found in sipunculids, brachiopods, priapulids, and occasionally other invertebrate phyla, and the hemocyanins-found only in molluscs and arthropods. It appears that ancestors of these three protein types arose around the same time, independently of each other. (1) Two things help reinforce this idea, the fact that these protein types differ so greatly in form, and that their usage shows fairly clear divisions between related phyla. For example, the molluscs and arthropods (which both utilize hemocyanin) are more closely related to each other than either are to the vertebrates (which utilize hemoglobin). (2)

Hemoglobins and hemerythrins are intracellular proteins and both use iron in the oxygen-binding site, but in different ways. Hemoglobins bind one oxygen molecule per heme-iron complex, while hemerythrins need a pair of iron atoms to bind one molecule of oxygen and contrary to their name contain no heme at all. Hemocyanin,

which also does not contain heme, is an extracellular protein and uses a pair of copper atoms to bind oxygen, rather than iron. The presence of heme and the metal ion used is what determines the color of the oxygenated blood. Oxygenated blood containing heme and iron (hemoglobin) is red, oxygenated blood with iron and no heme (hemerythrin) is a reddish-purple, while copper containing blood with no heme (hemocyanin) is blue. (3)

It is clear that hemoglobins, hemerythrins, and hemocyanins have been part of a convergent evolutionary process. Even though they are extremely different in their form, their function is clearly the same. Diversity appears not only between the groups, but within them as well. Among the hemocyanins the form can differ greatly. Molluscan and Arthropod hemocyanin are so different, in fact, that it remains in question whether these two groups were actually the result of convergent, rather than divergent evolution. (4)

## 1.2 Molluscan versus Arthropod Hemocyanins

Molluscan and arthropod hemocyanins are dissolved within the blood serum and are not contained within cells that circulate in the blood, as are most of the hemoglobins and the hemerythrins. (5) Also, both classes of hemocyanin form large molecules composed of many polypeptide subunits. This is where the similarities between the two classes end. The subunits are arranged in vastly different ways. Arthropods have been found to combine several different types of subunits together to create the final protein, whereas molluscan hemocyanins are composed solely of one type of subunit. Arthropod and molluscan hemocyanins have extremely different quaternary structures,

and are currently believed to be very distantly related. Molluscan hemocyanins share one similar copper-binding site with the arthropods, but the other molluscan copper-binding site resembles the copper-binding site found in tyrosinase. At best, the arthropod and molluscan hemocyanins share a very distant common ancestor. (1)

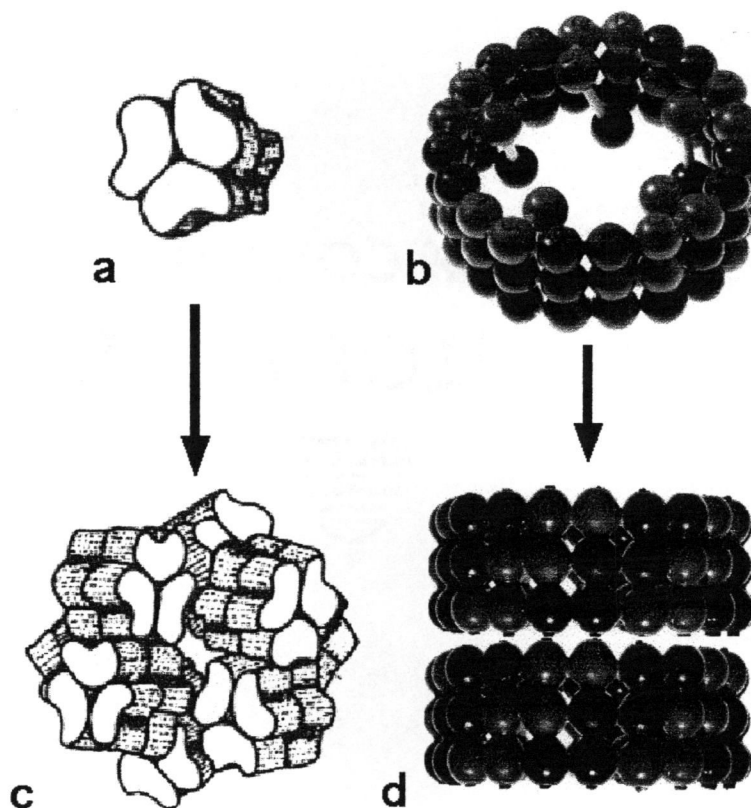


Figure 1. Basic Structure of Arthropod and Molluscan Hemocyanin. Arthropod hemocyanin is found in hexamers. Each subunit of the hexamer weighs  $\sim 75,000$  Da. The whole hemocyanin molecule may be composed of one, two, four, six, or eight hexamers. Molluscan hemocyanin is found in decamers. Each subunit of the decamer weighs 360 kDa (for seven oxygen-binding units) or 450 kDa (for eight oxygen-binding units). The hemocyanin molecule may be composed of one or more decamers. (a) a single hexamer (6-mer, the simplest arthropod hemocyanin found in nature) from spiny lobster, (b) a single decamer, the simplest molluscan hemocyanin found in nature, from *Octopus dofleini*, (c) an eight hexamer hemocyanin (48-mer, the most complex arthropod hemocyanin found in nature) from *Limulus polyphemus*, the Horseshoe Crab, (d) a di-decamer, a more complex molluscan hemocyanin from *Helix pomatia*.

### 1.3 Molluscan Hemocyanins

All molluscan hemocyanins that have been studied to date have been found to contain two distinct forms, or paralogues. After a gene duplication event occurs, there are two copies of the gene instead of one. These copies, which will begin diverging from the time of the duplication event, are known as paralogues of the gene. Only two species, *Octopus dofleini* (the Giant Pacific Octopus) and *Haliotis tuberculata* (the Green Ormer, an abalone), have had their complete hemocyanin proteins and DNA sequenced for both forms. In the case of *Octopus*, the two paralogues have been denoted A-type and G-type (see below). The van Holde lab co-authored a paper comparing the first paralogue of the *O. dofleini* and the first paralogue of the *H. tuberculata* hemocyanin gene and protein.(6) That collaboration resulted in the first presentation of gene sequences for Molluscan hemocyanins. This thesis describes the sequencing and gene structure of the second *O. dofleini* paralogue.

Molluscan hemocyanin exists as a decamer (or larger in the case of some gastropods) with molecular weights numbering in the millions of Daltons.(7)

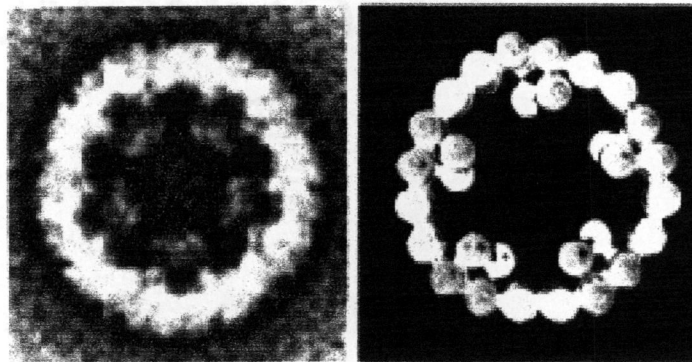


Figure 2. *Octopus dofleini* Hemocyanin Decamer. The left half shows an electron micrograph of the molecule and the right half shows a model of the molecule. Both are shown from a "bird's eye" orientation.

The decamer is composed of 10 identical subunits. A single subunit represents the entire coding region of the hemocyanin gene and for *O. dofleini* weighs 360,000 Da. The subunit itself can be broken down into seven or eight functional units (FU) depending on species. Each FU is capable of binding one molecule of oxygen, so a subunit with seven FUs, such as the one for octopus, can bind seven, enabling the complete decamer to bind 70 oxygen molecules.

The partially unfolded subunit resembles "beads on a string", each FU constituting a "bead" or folded region of protein, and each bead is separated by ~15 amino acids, referred to as a linker, that makes up the "string". See Figure 3.

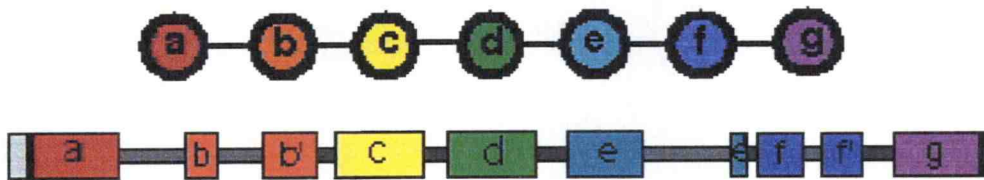


Figure 3. The *Octopus dofleini* Protein Subunit and Gene. The protein subunit is represented by the top model, each circle is an oxygen-binding unit (or Functional Unit, FU) and the lines attaching the FUs are the linker regions, composed of ~15 amino acids each. The gene is shown in similar fashion, the exons are represented by rectangles and are color-coded to match the FU it codes for. The linker regions are elongated by phase one linker introns (gray) and three FUs are split by internal introns (black).

Two subunits come together to form a dimer and five dimers join to create the decamer. Even though two forms of the gene (and therefore protein subunit) exist, the *Megathura crenulata* (Keyhole Limpet) subunits form only homodimers. This is also true for the formation of the decamer from the dimers. So, two distinct decamers exist, one for each paralogue.(8) A complete study has not been undertaken for the *O.*

*dofleini*, however preliminary data exists supporting two distinct homo-decamers for this species.

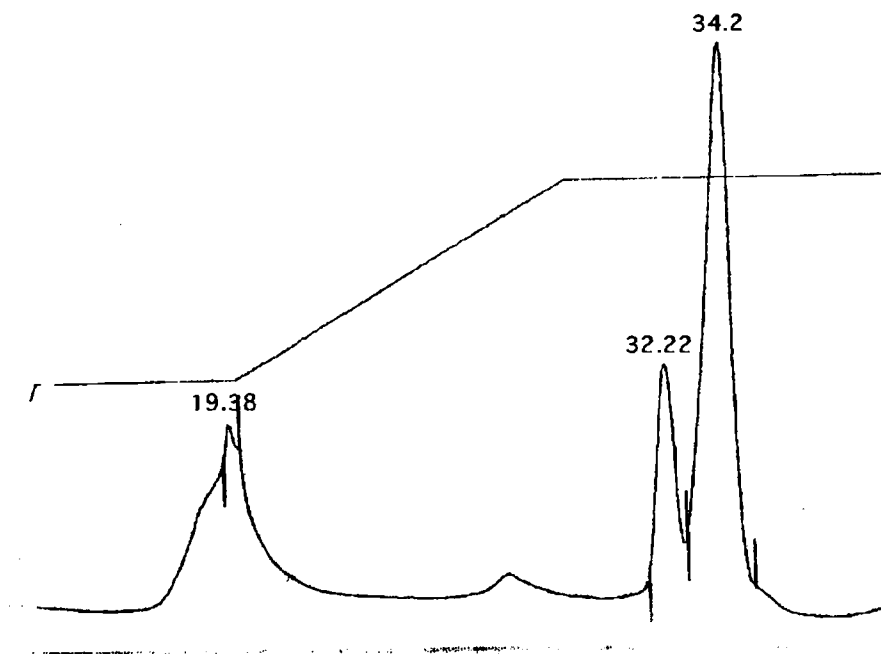


Figure 4. High Pressure Liquid Chromatography (HPLC) results show two distinct peaks for *O. dofleini*. The peak at 19.38 represents the presents of dimers and monomers in the solution. The 32.22 and 34.2 peaks represent the decamers. The peaks are very close together, indicating the two types are extremely similar. The height of the peaks is an indication that one type is more abundant than the other. The presence of two distinct peaks supports the idea that two distinct homo-decamers exist for this species.

The "beads on a string" model is useful when referring to the hemocyanin gene, as well as the protein subunit. See Figure 3. Most FUs are encoded by a single exon; therefore the gene structure retains a bead-like appearance, with exons separated by the string-like intron-containing linker regions. The string has been elongated to include phase one introns that are found regularly, one within each linker. A phase one intron refers to any intron that is inserted between the first and second base of a



nucleic acid triplet that codes for an amino acid. *O. dofleini* and *H. tuberculata* each have three FU exons that are split by internal introns. See Figure 5.

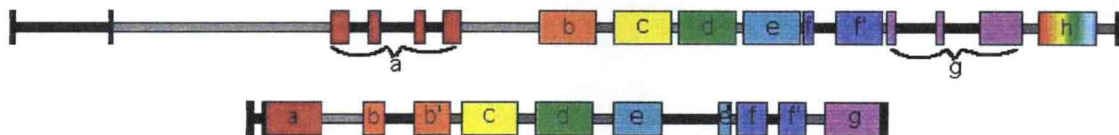


Figure 5. The *Haliotis tuberculata* gene and the *Octopus dofleini* gene. The *H. tuberculata* (Ht) gene (top) has eight FUs, the *O. dofleini* (Od) gene (bottom) has seven FUs. The exons are represented by rectangles and are color-coded to match the corresponding exon in the other gene. Ht-h has no corresponding exon in the *O. dofleini* sequence. The linkers have been expanded by phase 1 introns (thin gray bars), and several FU coding regions in both genes are split into two or more exons for completion of the FU by internal introns (thin black bars).

The *O. dofleini* has a total of ten introns: seven linker introns that are all phase one, and three internal introns of phase 2, 0, & 0, respectively. *H. tuberculata* has a total of fourteen introns: eight linker introns, that are all phase one, and six internal introns. (6) The total intron numbers are not including introns in the upstream/downstream signal regions. The linker introns do not vary in location or phase between the two species. This conservation indicates that these introns were probably in place before the divergence of the cephalopods and gastropods. The internal introns not only range in number, but also in location. *O. dofleini* has internal introns that split FU-b, FU-e, and FU-f each into two exons (one internal intron each). *H. tuberculata* has internal introns that divide FU-a into four exons, FU-f into two exons, and FU-g into three exons. The lack of pattern to the location and phase of the internal introns indicates that these introns probably arose after the cephalopod-gastropod split. For these introns to have been in place before the cephalopod-gastropod split, it would take nine

independent events to get them into the sequence plus six events for *O. dofleini* to remove the ones no longer present, and three events for *H. tuberculata* to remove the ones no longer present. That is a total of eighteen independent events assuming the internal introns preceded the cephalopod-gastropod split. If the internal introns appeared after the split then that requires three independent events for *O. dofleini* and six independent events for *H. tuberculata*. That is a total of nine independent events compared to eighteen for the introns-early idea.

## 2. Materials & Methods

### 2.1 Obtaining *Octopus dofleini* DNA and Sequencing

The *Octopus* used for sequencing was obtained from the Hatfield Marine Science Center, Newport, OR. Blood and tissue were taken and genomic DNA was obtained from the brain, gonad, and branchial glands. Short (15-25 base) Polymerase Chain Reaction (PCR) primers were constructed to "walk" down the sequence. The first primers were created from the G-type cDNA and then later from the genomic sequence itself as pieces were completed. AssemblyLign Version 1.0.9b (Oxford Molecular) was used to create and align contigs from the sequences found by PCR. Some sequences showed consistent ambiguities (later it was known that we were getting both A- and G-type sequences overlapping, See Results & Discussion), we therefore turned to cloning to separate the A- and G-types for sequencing. PCR products were cloned into the vectors pCR4-TOPO or pTrc His2-TOPO (Invitrogen) and then sequenced. The Central Services Lab (CSL) at Oregon State University did the PCR primer synthesis and DNA sequencing. Synthesis of oligonucleotide primers was accomplished on the 394 DNA/RNA Synthesizer, Applied Biosystems. DNA sequencing before October 2001 was processed on the ABI Prism 377 DNA Sequencer, Version 3.3, Applied Biosystems. After October 2001, DNA sequencing was done on the ABI Prism 3100 Genetic Analyzer, Version 3.4.1, Applied Biosystems. Sequencing problems arose when we began sequencing the microsatellite region. Drhodeamine was added to the sequencing mix, which significantly improved the resulting sequence. PCRs were run on the PTC-150 Minicycler, (MJ Research, Cambridge, MA). Two programs were used for PCR, "Alex" for regular PCR and

"Alex-Hot" for Hot Start PCR. The steps for "Alex" and "Alex-Hot" are shown in

Table 1.

	Alex	Alex-Hot	
Step	Temp.	Time	Temp.
1 Pre-heating	N/A	2:30	92°C
2 Denaturation	92°C	:30	92°C
3 Annealing	52°C	1:00	62°C
4 Extension	72°C	2:00	72°C
5 Repeat	N/A	<sup>2-4</sup> 29 times	N/A
6 Final Extension	72°C	7:00	72°C
7 Hold	4°C	∞	4°C

Table 1. PCR Thermocycler Programs. "Alex-Hot" has a pre-heating step that must be done before the Taq polymerase is added to the tubes. After the pre-heating step is complete the tubes are opened and Taq is added. "Alex" does not contain a pre-heating step, the Taq polymerase is added to the tubes at the beginning along with the rest of the components. After the PCR is complete the tubes are held at 4°C until removed from the thermocycler.

The components for the PCR reaction include water, 10x buffer, dNTPs, Mg<sup>++</sup>, Taq polymerase, 2 primers, and template DNA. Every PCR tube contained 10 µl of 10x buffer, 10 µl of dNTPs, 6 µl of Mg<sup>++</sup>, 0.5 µl of Taq polymerase, 2µl each of primer #1 and primer #2. The quantity of the template DNA and the quantity of water were interrelated. Most PCRs had 2 µl of template DNA and 67.5 µl of water, but if there was a need to increase the amount of DNA, then the water level was decreased accordingly to maintain each tube at 100 µl total. The 5U/µl Amplitaq® DNA polymerase, 10x Buffer II, 10mM dNTP mix, and 25mM MgCl<sup>++</sup> were purchased

from Perkin Elmer/Applied Biosystems. PCR results were visualized by running them out on 1% agarose gels with 6x Loading Dye Solution and GeneRuler™ 100 bp DNA ladder Plus, both from MBI Fermentas, and Ethidium Bromide (EtBr) staining. MacVector Version 7.0 was used to translate the DNA sequences into the amino acid sequences. The sequences found by MacVector were then compared to the cDNA translations found by Walter Lang [reference] to verify the exons and find the locations of the introns.

## 2.2 Population Genetics Study

Nine *O. dofleini* were used in the population genetics study to determine if any octopus has only one type of hemocyanin gene (A or G). Octopus tentacle tips were taken from the Hatfield Marine Science Center, the Oregon Coast Aquarium (Newport, OR), and the Seaside Aquarium (Seaside, OR). Genomic DNA was extracted from tentacle tips or brain tissue using the Qiagen DNeasy Kit. The DNA was stored in 10 µl aliquots in the -80°C freezer and used as templates for PCR reactions. The population genetics primers: PGenA, PGenG, PGenN, PGenAR, PGenGR, and PGenNR (See Results for definitions) were synthesized at the CSL at Oregon State University.

### 3. Results and Discussion

#### 3.1 A- and G-types are Paralogous Genes

During sequencing, we would occasionally come across sequence ambiguities. The Staden file would show two distinct nucleotides for the same location, resulting in an "N" being placed in the sequence.

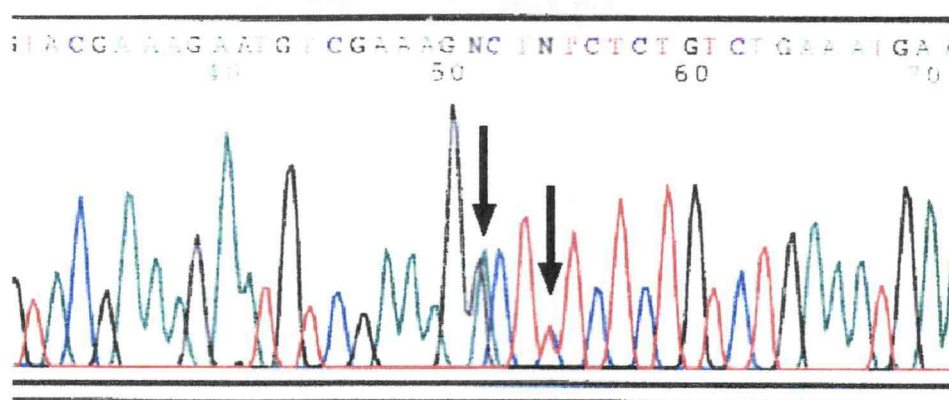


Figure 6. ABI Prism Staden File with overlapping sequence. Each nucleotide is represented by a color, A=green, G=black, T=red, and C=blue. Peaks show where the corresponding nucleotide appears in the sequence. If the signal for two different nucleotides is read in the same location, the sequencer may not be able to distinguish which nucleotide is the correct one and will put an "N" in the sequence. The black arrows show where this has happened, the first arrow shows both A and G giving a signal in that location, the second arrow shows the same for T and C. This is how it was discovered that we were getting two overlapping sequences (the A- & G-type).

In an attempt to clear up these occasional ambiguities, we began cloning our DNA fragments into vectors. We would then select 8-10 colonies so re-sequencing could be done. Generally, about half of the results would come back from sequencing showing the first nucleotide and the other half would come back showing the second nucleotide.

It became clear that the two types were similar enough that primers created from the cDNA were picking up both sequences. That is when it became apparent that we were dealing with two distinct, yet extremely similar, gene sequences. The two types had been detected earlier by Walter Lang and were named according to the first difference observed in their sequences and hence, came to be known as the A-type and the G-type.(5) With two gene types present the question emerged: are we dealing with paralogous genes or alleles of the gene.

For the two types to be alleles, the entire gene, coding and non-coding regions would have to be equal in length. If the two are different sizes than unequal crossing over would occur during meiosis, possibly causing loss of function for this essential gene. Also, if the two types are alleles, then when a population sample is taken, we would expect to find both heterozygous (AG) and homozygous (GG or AA) individuals. By contrast, if the two types are paralogues then there would be no size constraint on them, and we would expect every individual tested to have both types. As sequencing progressed the two types were occasionally found to have different size introns. This is what first led us to believe that A- and G-type were indeed paralogues, and not alleles.

This conclusion was further supported by a limited population study we conducted. *O.dofleini* tentacle tips were collected and genomic DNA was extracted from the tissue. The DNA was used in a PCR reaction to determine if the octopus had one or both types present. If A and G are indeed paralogues than every octopus tested should have both types. It was necessary to devise a simple technique to identify each form.

Six PCR primers were designed that would identify of which hemocyanin gene (A-type or G-type) the octopus carried without having to sequence the DNA.

The three forward primers were named PGenA, PGenG, and PGenN and the three reverse primers were named PGenAR, PGenGR, and PGenNR. PGen stands for Population Genetics, A refers to the A-type, G refers to the G-type, N refers to sequence that is the same in both types, and R refers to reverse.

These primers were created in areas where spots of sequence divergence and spots of total conservation were found in close proximity and covered a region in which the two types differ in length. So, when a sample of DNA was run with PGenA and PGenAR only one band should appear because only the A-type is acting as a template. The same is true for the use of PGenG and PGenGR, only G-type is used as a template so only one band would be visible. When PGenN and PGenNR are used both A- and G-types will be used as templates and therefore both bands should be visible. The forward and reverse primers enclose an intron region that differs in size for the two types. The intron in the G-type is 354 bp long, add to that the length of the primers and their distance to the intron, and G-type yields a band at 457 bp. The intron in the A-type is 223 bp long and similarly yields a band at 326 bp. If both types are present then the gel will show both bands, G-type at 457 bp and A-type at 326 bp, a difference of 131 bp, which is easily detectable on a gel. See Figure 7.

If the sample has only one of the two gene types, then the PGenN and PGenNR reaction should yield only the band corresponding to that type.

DNA from nine animals was sampled, and seven of them showed two bands at the expected sizes when run with PGenN and PGenNR showing they contained both A-



and G-types. The last two showed no bands, and corresponded to the tentacle tips taken from dead animals. For unknown reasons, the primer pair PGenG and PGenGR, never yielded bands. The other two primer pairs worked and still provided the required information, even without PGenG and PGenGR to give G-type only reinforcement data.

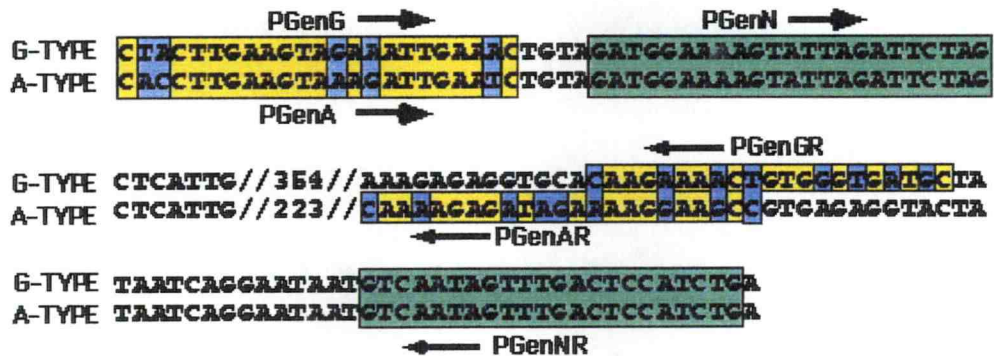


Figure 7. Population Genetics PCR Primer Locations.

The type-specific primers are shown in yellow with the sequence differences between the types highlighted in blue. The PGenN and PGenNR primers, in green, are the same for both sequences. The arrows pointing to the right indicate the forward primers and the left facing arrows indicate reverse primers. The numbers shown (*//354//* and *//223//*) represent the number of base pairs in the intron. This difference is the key to detecting the presence of both types on a gel. For simplicity, only one strand for each type has been shown, therefore to get the actual sequence for the reverse primers, the sequence indicated by the figure would need to be reversed and complimented.

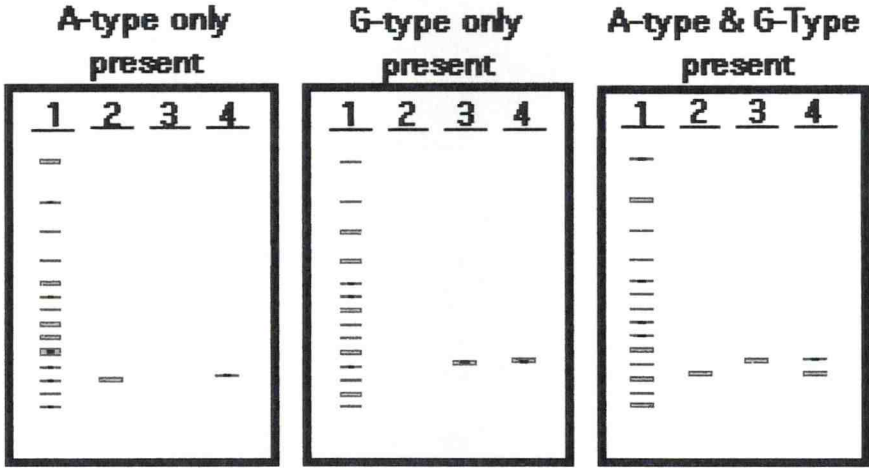


Figure 8. Theoretical Gel Results for the PGen Primer Sets. Lane 1 is a 100 bp DNA Ladder (the fifth band from the bottom is 500 bp). Lane 2 is run with the PGenA/AR primer set. Lane 3 is run with the PGenG/GR primer set. Lane 4 is run with the PGenN/NR primer set. Lanes 2-4 had genomic DNA from one octopus. The gel on the left shows the expected results if the octopus tested had only the A-type gene. The center gel shows the expected results if the octopus tested had only the G-type gene. The gel on the right shows the expected results if the octopus tested had both the A- and G-type genes.

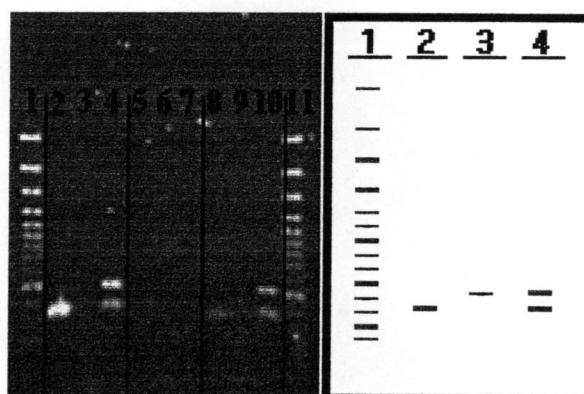


Figure 9. Actual Gel versus Theoretical Gel for A- & G-type. The left shows an actual gel divided into five areas. The two outermost areas (Lanes 1 & 11) contain 100 bp DNA ladder. Lanes 2, 5, & 8 were run with the PGenA/AR primer set. Lanes 3, 6, & 9 were run with the PGenG/GR primer set. Lanes 4, 7, & 10 were run with the PGenN/NR primer set. Lanes 2-4 had genomic DNA from one Octopus, lanes 5-7 had genomic DNA from a second Octopus, and lanes 8-10 had genomic DNA from a third Octopus. Each of the three middle areas (Lanes 2-10) is equivalent to the set up for the theoretical gels. On the right are the theoretical gel results for an Octopus that has both the A- & G-type gene for comparison. With the exception of Lanes 3, 6, & 9 (the runs with the PGenG/GR primer set) the actual gel matches the prediction.

Primer Pair	Number of Fragments	Fragment Length
PGenG & PGenGR	1	450 bp
PGenA & PGenAR	1	303 bp
PGenN & PGenNR	2	457 bp (G) 326 bp (A)

Table 2. Number and size of PCR fragments resulting from the various primer pairs, if the octopus has both the A- and G-type genes. If the octopus had only A type, then no band would appear when run with the PGenG/PGenGR primer pair and only the A band would appear when run with the PGenN/PGenNR primer pair. Similarly, if the octopus is only G-type, then no band would appear when run with the PGenA/PGenAR primer pair and only the G band would appear when run with the PGenN/PGenNR primer pair.

If A- and G-type were alleles, the chance that seven octopus sampled at random would all be heterozygous is very low. Barring heterozygote advantage, a situation in which the fitness of the heterozygous individuals is greater than either of the homozygotes (leading to higher fecundity for the heterozygotes), the greatest number of heterozygous individuals would be expected in a population where the two alleles were found in equal proportion. If the alleles appear in a 1:1 ratio, then the number of heterozygotes could be calculated using the Hardy-Weinberg theorem as a guide. It is very unlikely that the population is actually in Hardy-Weinberg equilibrium, but the theorem will provide a rough estimate. The important part of the equation is the frequency of heterozygotes, which is represented by  $2pq$ , where  $p$  and  $q$  are the allele frequencies.

$$2pq=2(0.5)(0.5)=0.5$$

The likelihood of drawing one heterozygous individual from the population is 50%, the likelihood of drawing seven out of seven heterozygous individuals from the population is the likelihood of drawing a heterozygote each time,  $0.5^7$  or 0.0078 (0.78% chance).

Two distinct genes have been found, not only in *Octopus*, but also in *Haliotis tuberculata*, *Sepia officinalis*,<sup>(9)</sup> and *Megathura crenulata*. So far, every molluscan hemocyanin studied has revealed two types, found in a single individual. This evidence along with the population study and the unequal lengths of the two *O. dofleini* types has led us to conclude that the two types are, in fact, paralogues and not alleles.

It seems likely that both types of the gene can be expressed, although not necessarily in equal amounts. See the HPLC results, Figure 4, shown earlier. Lang (5) observed both A & G type cDNA in a single organism.

### 3.2 Microsatellites in the A- and G-type *Octopus dofleini* hemocyanin

As previously stated, there are three introns that are found within an exon (internal introns), rather than within the linker region (linker introns) for *O. dofleini*. The intron that splits the exon for FU-e contains a microsatellite. A microsatellite is a region containing tandem repeats of short sequence motifs (the motifs range between one and six nucleotides in length). Both types were found to contain a microsatellite in this region, although the two differ in length (1305 bp for A-type and 1244 bp G-type). A large part of the microsatellite is extremely similar between the two types, but regions exist where divergence is evident.

The lengths of the repeats found in this microsatellite are unusually high. A paper by Kruglyak et al. looked at tandem repeat lengths for microsatellites in humans, mice, fruit flies, and yeast. The average lengths of di-, tri-, and tetranucleotide repeats were 13.1, 6.5, and 5.75, respectively.(10) The (GT)<sub>n</sub> repeat in the A-type is 62 repeats (124 bp), and the (ATAC)<sub>n</sub> repeat in the A-type is 21 repeats (84 bp). The comparison tables and graphs in the Kruglyak paper do not even include repeat lengths of this size. The A-type microsatellite has four different tandem repeats (ATAC, AT, GT, and GTAT) present in seven locations. The (AT)<sub>n</sub> repeat is found in four separate areas. The G-type microsatellite has five different tandem repeats (ATAC, AT, GT, GA, and CATA) present in nine locations. The (AT)<sub>n</sub> repeat is found in five separate areas. All

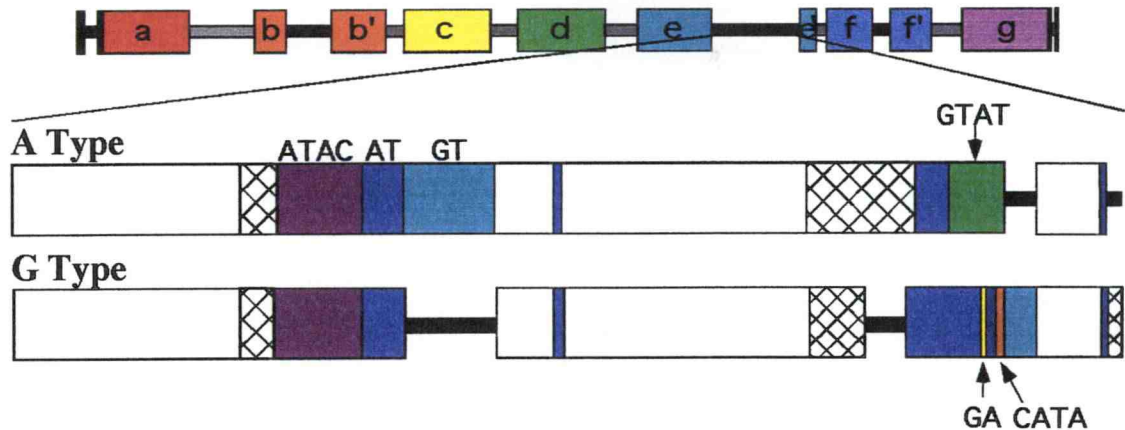


Figure 10. *Octopus dofleini* microsatellites compared. The areas of tandem repeats are represented by colors. The color is labeled with the repeat the first time it is used only. For example, blue always represents  $(AT)_n$  repeats, but is only labeled the first time it is seen in the sequence. Only repeats of 10 base pairs or longer have been labeled. The white regions represent areas of high identity (82-93%) and the hatched regions represent areas of lower identity (41-70%). Thin black bars mark the presence of gaps that have been inserted to facilitate alignment.

the tandem repeats in these microsatellites are either di- or tetra- nucleotide repeats; nowhere throughout the entire length of the microsatellite are tri-nucleotide repeats found. While microsatellite data exists for several other molluscs, the *O. dofleini* microsatellite presented here is the first to be found within a hemocyanin gene. *H. tuberculata*, also fully sequenced, shows no microsatellites for either of its gene types. The microsatellites of the A- and G-types show amazing amounts of identity at various positions throughout their lengths and are found in the same location on the gene. This would support the idea that the microsatellite was present, at least in part, before the gene duplication for *O. dofleini*, but after the cephalopod-gastropod divergence, since no similar structure is found in *H. tuberculata*. Before the gene duplication that created the A- and G-types, there was one microsatellite. So, at the time of the

duplication, the A-type microsatellite was, by necessity, identical to the G-type microsatellite. The organization of the repeats can help us uncover which changes to the microsatellites have contributed to the divergence in sequence.

The A- and G-types each have a di-nucleotide repeat, specifically a  $(GT)_n$  repeat that is surrounded by another repeat sequence. If the  $(GT)_n$  repeat is removed, it is clear that the pattern seen before it is the same as the pattern that follows it. It can be easily imagined that the  $(GT)_n$  repeat was inserted into the pre-existing pattern that now surrounds it. Another clue that these  $(GT)_n$  repeat regions are a recent arrival is the corresponding lack of such a sequence in the other type. In both cases, one type has a gap where the  $(GT)_n$  repeat region is located in the other type.

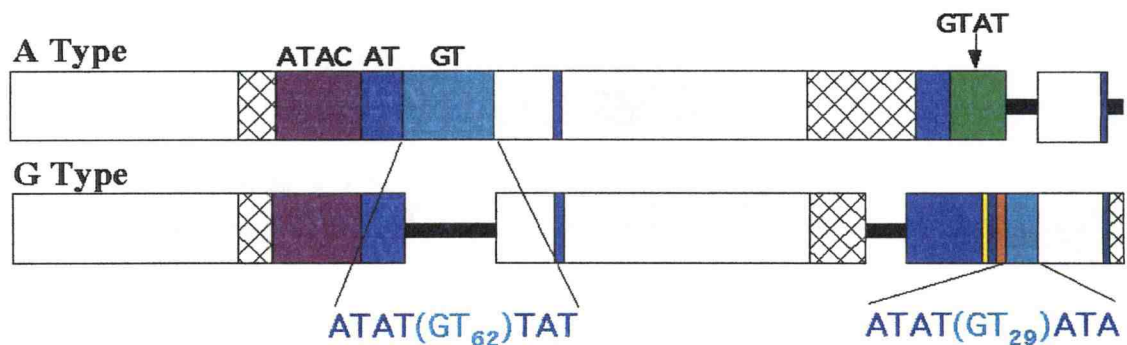


Figure 11. A- and G-type microsatellites with  $(GT)_n$  repeat and flanking regions. The  $(GT)_n$  repeat regions are shown with the flanking base pairs on either side. This suggests that the  $(GT)_n$  repeat regions were inserted recently (after the gene duplication) within an  $(AT)_n$  repeat. Thin black bars represent the gaps in gene sequence. The regions of the  $(GT)_n$  repeat in both types shows corresponding gaps in the other type.

It is likely that before the gene duplication the first half of the microsatellite more closely resembled the current G-type microsatellite, since insertion of tandem repeats is much more likely than the perfect removal of repeats. The same can be said about

the end of the A-type sequence. It is likely that at the time of the split, the end of the microsatellite probably more closely resembled the A-type microsatellite. For non-repeated sequences, the areas of greatest divergence, represented in Figures 10 and 11 by the hatched regions, directly precede regions that contain tandem repeats. Perhaps these areas are more susceptible to being mutated because of the possible polymerase slippage occurring within the repeated regions. Overall, the A- and G-type microsatellites have 71% identity. With the gaps, and the region corresponding to the gap in the other sequence, removed, the microsatellites have 85% identity. If the tandem repeats are removed altogether, the identity is 92%. Since the mutation rates of microsatellite regions are commonly known to exceed the mutation rates of exon sequences, due to the elongation of repeats rather than an increase in point mutations, the extreme similarity between the two microsatellites supports the idea that the *O. dofleini* paralogues are the result of a very recent gene duplication event.

### 3.3 Evolution

The entire hemocyanin gene sequence is now known for both types. This expands the comparison possibilities. So far, we have been limited to amino acid comparisons, but now the A- and G-type can be compared at the DNA level, as well. This includes not only the exon sequence, but also all of the introns. It would be expected that the intron regions would show a greater percent difference overall, since, unlike the exons, there is no selective pressure restraining them.



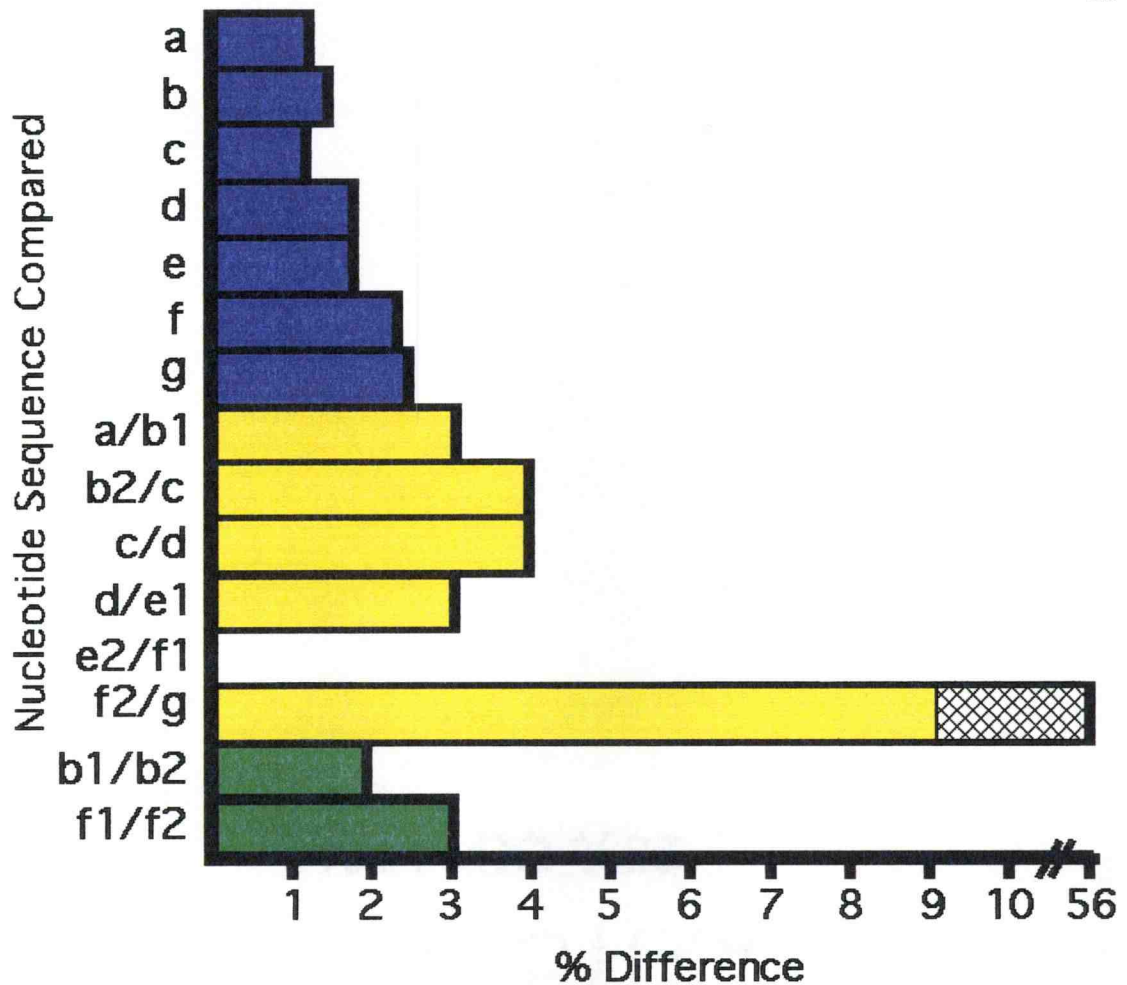


Figure 12. Comparison of Nucleic Acid sequences for Exons, Linker Introns, and Internal Introns for the *Octopus dofleini* hemocyanin paralogues. The exon sequences, shown in purple, range from 1.2-2.5% difference, the Linker Introns (yellow) have been named for which exons they are found between. They range from 0-54% difference. The Internal Introns (green), excluding the microsatellite found between e1/e2 (not shown), show surprisingly little difference, ranging from 2-3%. The percent difference for f2/g was 56%, however if the sequence that existed in only one type was ignored then the percent difference drops to 9%.

The f2/g comparison is remarkably different from the others. The percent difference is very high, even with the removal of sequence only seen in one type. The reason this region shows such high percent difference is very likely the same reason why there are

three (AT)<sub>3</sub> tandem repeats in it. Sequence has been inserted and/or deleted from this region, and the result is a new microsatellite region just forming.

It is surprising to see that the e2/f1 linker intron has no differences at all, but that intron is only 100 bp in length, the shortest region being measured. One thing that is very striking is how the exons show a greater percent difference the further along the molecule one goes. This difference between the units is small, ranging from 0.3-0.7%, but a trend is visible. This trend fits with the idea of tandem duplication and fusion that will be described in Figure 14. Figure 13 shows the percent identity for all for the regions shown in Figure 12.

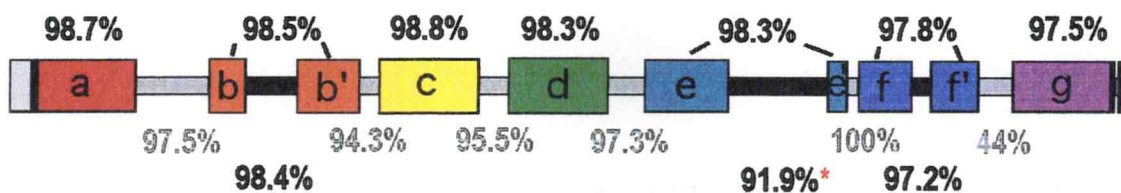


Figure 13. Percent Identity Among the Exons, Linker Introns, and Internal Introns for the *Octopus dofleini* hemocyanin paralogues. The percentages at the top correspond to the exon beneath, the numbers below correspond to the linker introns (gray), and the internal introns (black).

The origins of the single oxygen-binding unit (FU) are still unclear. However it has been noted that the two copper-binding sites found in the current oxygen-binding units appear to be of different origins.(1) One copper-binding site is similar to that found in arthropod hemocyanins, and the other similar to the copper-binding sites found in tyrosinases. This indicates that the original oxygen-binding unit might be the result of a fusion between two ancient genes that each coded for a single copper-binding site. Individually, these ancient proteins might later have given rise to the arthropod hemocyanins and tyrosinases.

Once the first oxygen-binding unit was in existence, tandem duplication events (possibly the result of uneven crossing-over) occurred to give the present seven or eight FU subunit that we have today.

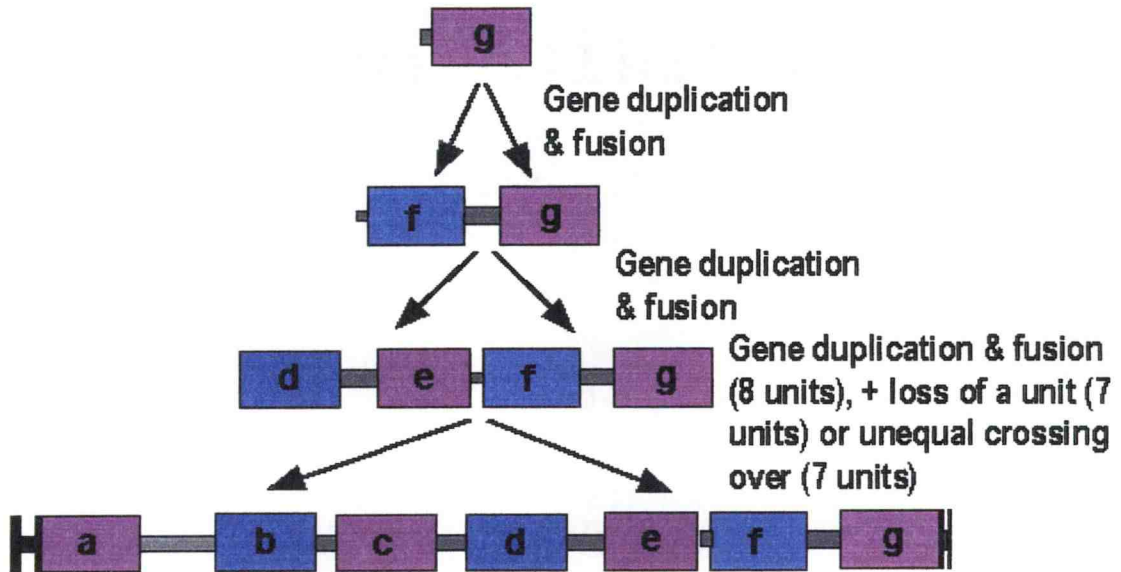


Figure 14. Hypothetical Formation of the Current Hemocyanin Gene Structure. Following the trend of exon differences noted above, the original unit would be the current FU-g (the unit showing the most percent difference). This unit duplicates, giving a two unit structure, another duplication leads to a four unit structure. Units d and e have the same percent identity (98.3%), both greater than the percent identity for the f and g units. A final duplication plus a secondary loss of one unit (or unequal crossing over) would lead to the 7-unit structure.

The degree of identity between functional units can be used as a tool to estimate the time of divergence for the groups of organisms that have this gene. For *O. dofleini*, it can be seen that comparing FU-a of A-type to FU-a of G-type, gives an extremely high percent identity (97% over 1233 bp). The percent identity remains extremely high, 95% and 97%, when comparing the same FU from one paralogue to the other. *H. tuberculata* has lower sequence identity between its paralogues, ranging from 59% to 73%.

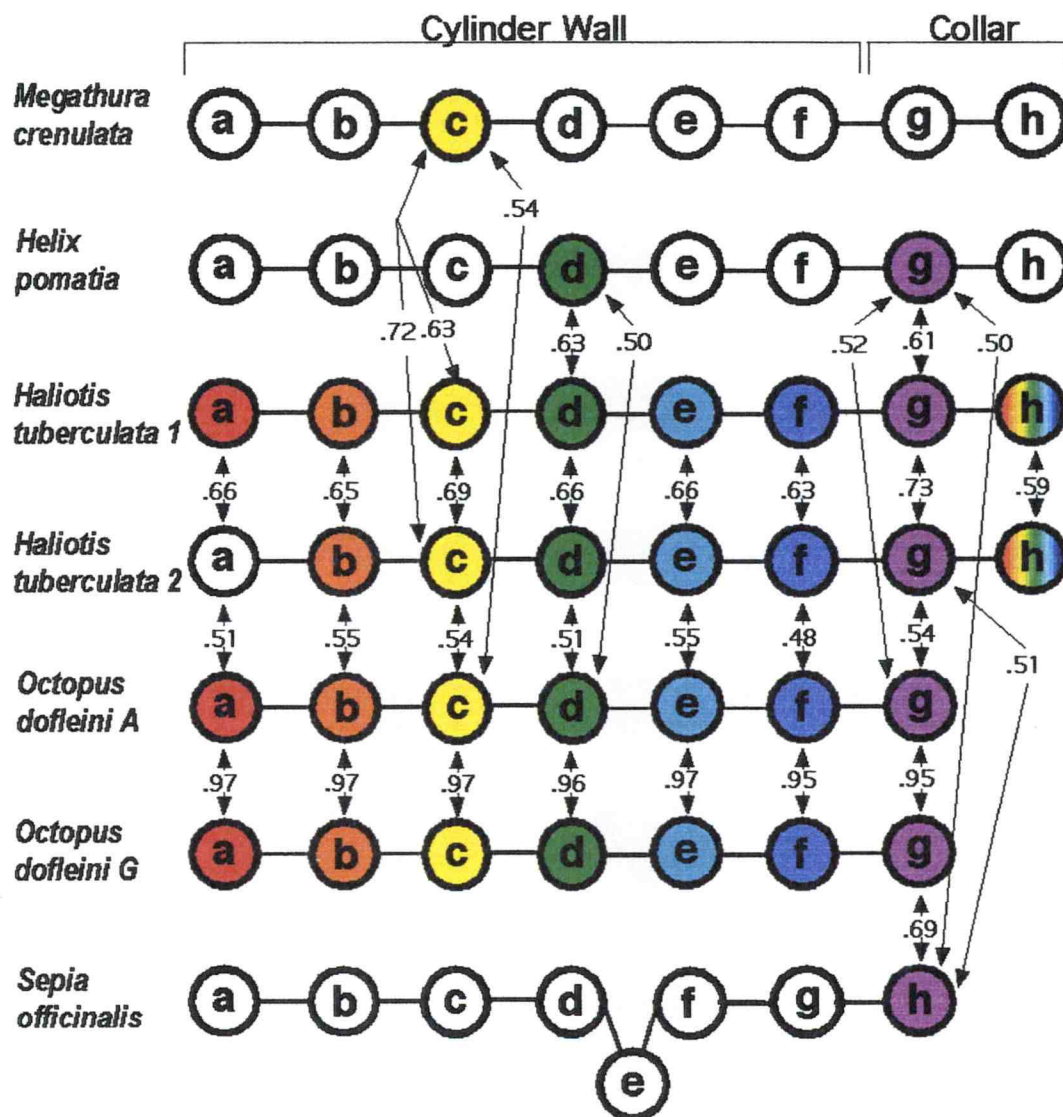


Figure 15. Line-up of Seven Molluscan Hemocyanin Subunits. Two gastropods: *Haliotis tuberculata* (paralogues) and *Megathura crenulata*; one pulmonate: *Helix pomatia*; and two cephalopods: *Octopus dofleini* (paralogues) and *Sepia officinalis*. The units in color represent units with the full amino acid sequence available. Identity across unit G/H (purple) ranges from 53-95%, depending upon the divergence time between the organisms in the comparison. Comparing FUs from the same species reveals a consistent 35-49% identity across all the species.

When the comparisons range across the FUs, comparing FU-a to FU-b through FU-g from A-type etc., the percent identity drops severely, ranging from 38%-45%. When one looks at the *H. tuberculata* in the same fashion, the results are 35% to 47%, strikingly similar to those for *O. dofleini*. The only other pair available for comparison is this way is *Helix pomatia* FU-d vs. FU-g, which shows 42% identity. The identity of the FU-d vs. FU-g pairs for *O. dofleini* and *H. tuberculata* are 40% and 44%, respectively.

Even without further analysis these percentages begin to give a picture of when taxa began to diverge. The lowest percentages 35-47%, correspond to the comparisons between different subunits, and are extremely similar among all the hemocyanins with enough protein sequence available to make the comparison. This common similarity in percent places the formation of the seven or eight FU subunit before the divergence of any of the taxa being looked at. In other words, the seven or eight FU subunit existed before the gastropod/cephalopod divergence, the highest order taxa divergence possible with the given species.

Of all the species in this comparison, the *O. dofleini* paralogues have the greatest identity, 95-97%, showing that they are very likely the most recently diverged. All comparisons that would be possible in Figure 15 have been given in Tables 3-15.

Table 3. *Octopus dofleini* (Od) Functional Unit Percent Identity: OdA versus OdA and OdG versus OdG. The top half shows A-type versus A-type, and the bottom half shows G-type versus G-type. The Each box contains the percent identity and  $\tilde{d}_{ij}$  (calculated from the percent identity using this equation:  $\tilde{d}_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).

*Octopus dofleini* A-type vs. *Octopus dofleini* A-type

	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
<b>a</b>		42% .867	39% .941	42% .867	39% .941	42% .867	42% .867
<b>b</b>	43% .843		44% .820	41% .891	45% .798	42% .867	44% .820
<b>c</b>	39% .941	45% .798		42% .867	41% .891	43% .843	40% .916
<b>d</b>	42% .867	41% .891	42% .867		42% .867	41% .891	40% .916
<b>e</b>	38% .967	44% .820	42% .867	42% .867		40% .916	40% .916
<b>f</b>	42% .867	43% .843	43% .843	41% .891	39% .941		44% .820
<b>g</b>	42% .867	43% .843	41% .891	40% .916	38% .967	44% .820	

*Octopus dofleini* G-type vs. *Octopus dofleini* G-type

Table 4. *Haliotis tuberculata* (Ht) Functional Unit Percent Identity: Ht1 versus Ht1 and Ht2 versus Ht2. The top half shows type1 versus type 1, and the bottom half shows type 2 versus type 2. The Each box contains the percent identity and  $\tilde{d}_{ij}$  (calculated from the percent identity using this equation:  $\tilde{d}_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).

*Haliotis tuberculata* 1 vs. *Haliotis tuberculata* 1

	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>	<b>h</b>
<b>a</b>		40% .916	41% .891	39% .941	38% .967	38% .967	40% .916	35% 1.04
<b>b</b>			43% .843	41% .891	46% .776	41% .891	40% .916	38% .967
<b>c</b>				39% .941	42% .867	41% .891	47% .755	38% .967
<b>d</b>					44% .820	43% .843	44% .820	38% .967
<b>e</b>						44% .820	44% .820	38% .967
<b>f</b>							42% .867	37% .994
<b>g</b>								40% .916
<b>h</b>								

*Haliotis tuberculata* 2 vs. *Haliotis tuberculata* 2

Table 5. *Octopus dofleini* Functional Unit Percent Identity: A-type versus G-type. The numbers in red correspond to the percent identity between like subunits (FU-a vs. FU-a, etc). The Each box contains the percent identity and  $\tilde{d}_{ij}$  (calculated from the percent identity using this equation:  $\tilde{d}_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).

		Octopus dofleini G-type						
		a	b	c	d	e	f	g
Octopus dofleini A-type	a	97% .030	42% .867	39% .941	42% .867	39% .941	42% .867	42% .867
	b	42% .867	97% .030	44% .820	41% .891	45% .798	42% .867	43% .843
	c	39% .941	45% .798	97% .030	42% .867	42% .867	43% .843	41% .891
	d	42% .867	42% .867	42% .867	96% .040	42% .867	41% .891	40% .916
	e	38% .967	45% .798	41% .891	42% .867	97% .030	40% .916	40% .916
	f	42% .867	43% .843	42% .867	41% .891	39% .941	95% .051	44% .820
	g	42% .867	44% .820	40% .916	40% .916	39% .941	44% .820	95% .051



Table 6. *Haliotis tuberculata* (Ht) Functional Unit Percent Identity: type1 versus type 2. The numbers in red correspond to the percent identity between like subunits (FU-a vs. FU-a, etc). The row for the comparisons for Ht2-a is hatched because Ht2-a is not fully sequenced. The Each box contains the percent identity and  $\tilde{d}_{ij}$  (calculated from the percent identity using this equation:  $\tilde{d}_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).

*Haliotis tuberculata* 1

	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>	<b>h</b>
<b>a</b>								
<b>b</b>	38% .967	65% .430	41% .891	42% .867	46% .776	39% .941	42% .867	38% .967
<b>c</b>	41% .891	44% .820	69% .371	43% .843	45% .798	42% .867	47% .755	39% .941
<b>d</b>	39% .941	42% .867	41% .891	66% .415	44% .820	40% .916	42% .867	38% .967
<b>e</b>	39% .941	43% .843	42% .867	41% .891	66% .415	44% .820	45% .798	38% .967
<b>f</b>	40% .916	40% .916	40% .916	42% .867	43% .843	66% .415	45% .798	37% .994
<b>g</b>	43% .843	42% .867	45% .798	43% .843	44% .820	44% .820	73% .314	41% .891
<b>h</b>	35% 1.04	36% 1.02	40% .916	41% .891	40% .916	37% .994	40% .916	59% .527

*Haliotis tuberculata* 2



Table 8. Functional Unit Percent Identity: *Haliotis tuberculata* type 1 (Ht1) versus *Octopus dofleini* G-type (OdG). The numbers in red correspond to the percent identity between like subunits (FU-a vs. FU-a, etc). The yellow boxes show where the percent identity differs from *Haliotis tuberculata* type 1 (Ht1) versus *Octopus dofleini* A-type (OdA). The Each box contains the percent identity and  $\sim d_{ij}$  (calculated from the percent identity using this equation:  $\sim d_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).

		<i>Haliotis tuberculata</i> 1							
		a	b	c	d	e	f	g	h
<i>Octopus dofleini</i> G-type	a	50% .693	40% .916	41% .891	38% .967	40% .916	40% .916	42% .867	35% 1.04
	b	41% .891	57% .562	43% .843	44% .820	45% .798	42% .867	43% .843	36% 1.02
	c	39% .941	44% .820	54% .616	40% .916	44% .820	39% .941	44% .820	36% 1.02
	d	40% .916	39% .941	42% .867	52% .653	41% .891	44% .820	44% .820	37% .994
	e	39% .941	45% .798	41% .891	44% .820	51% .673	42% .867	40% .916	36% 1.02
	f	40% .916	42% .867	41% .891	42% .867	46% .776	50% .693	42% .867	38% .967
	g	41% .891	42% .867	44% .820	42% .867	44% .820	39% .941	55% .597	38% .967

Table 9. Functional Unit Percent Identity: *Haliotis tuberculata* type 2 (Ht2) versus *Octopus dofleini* A-type (OdA). The numbers in red correspond to the percent identity between like subunits (FU-b vs. FU-b, etc). The yellow boxes show where the percent identity differs from *Haliotis tuberculata* type 2 (Ht2) versus *Octopus dofleini* G-type (OdG). The column for the comparisons for Ht2-a are hatched because Ht2-a is not fully sequenced. The Each box contains the percent identity and  $\sim d_{ij}$  (calculated from the percent identity using this equation:  $\sim d_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).

		<i>Haliotis tuberculata</i> 2							
		a	b	c	d	e	f	g	h
Octopus dofleini A-type	a	/	39% .941	42% .867	39% .941	39% .941	43% .843	41% .891	35% 1.04
	b	/	55% .597	45% .798	40% .916	47% .755	42% .867	42% .867	45% .798
	c	/	42% .867	54% .616	40% .916	40% .916	40% .916	43% .843	38% .967
	d	/	40% .916	43% .843	51% .673	39% .941	45% .798	44% .820	37% .994
	e	/	44% .820	44% .820	43% .843	55% .597	41% .891	41% .891	37% .994
	f	/	39% .941	42% .867	42% .867	44% .820	48% .733	39% .941	38% .967
	g	/	43% .843	45% .798	43% .843	44% .820	43% .843	54% .616	39% .941
	h	/							



Table 11. Functional Unit-c Percent Identity: *Haliotis tuberculata* type 1 (Ht1), *Haliotis tuberculata* type 2 (Ht2), *Octopus dofleini* A-type (OdA), *Octopus dofleini* G-type (OdG), and *Megathura crenulata* (Mc). Each box contains the percent identity and  $\sim d_{ij}$  (calculated from the percent identity using this equation:  $\sim d_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).











		OdA	OdG	Ht1	Ht2	Mc
						
OdA			97% .030	53% .634	54% .616	54% .616
OdG				53% .634	55% .597	54% .616
Ht1					69% .371	63% .462
Ht2						72% .328
Mc						

Table 12. Functional Unit-d Percent Identity: *Haliotis tuberculata* type 1 (Ht1), *Haliotis tuberculata* type 2 (Ht2), *Octopus dofleini* A-type (OdA), *Octopus dofleini* G-type (OdG), and *Helix pomatia* (Hp). Each box contains the percent identity and  $\sim d_{ij}$  (calculated from the percent identity using this equation:  $\sim d_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).











	OdA	OdG	Ht1	Ht2	Hp
					
OdA 		96% .040	52% .653	50% .693	50% .693
OdG 			52% .653	50% .693	51% .673
Ht1 				66% .415	63% .462
Ht2 					63% .462
Hp 					

Table 13. Functional Unit-g Percent Identity: *Haliotis tuberculata* type 1 (Ht1), *Haliotis tuberculata* type 2 (Ht2), *Octopus dofleini* A-type (OdA), *Octopus dofleini* G-type (OdG), *Helix pomatia* (Hp), and *Sepia officinalis* (So, FU-h). Each box contains the percent identity and  $\sim d_{ij}$  (calculated from the percent identity using this equation:  $\sim d_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).













		OdA	OdG	Ht1	Ht2	Hp	So
							
OdA			95% .051	56% .579	54% .616	52% .653	70% .356
OdG				55% .597	54% .616	51% .673	69% .371
Ht1					73% .314	61% .494	52% .653
Ht2						64% .446	51% .673
Hp							50% .693
So							



Table 14. Functional Unit Percent Identity: *Haliotis tuberculata* type 1 (Ht1), *Haliotis tuberculata* type 2 (Ht2), *Octopus dofleini* A-type (OdA), *Megathura crenulata* (Mc), *Helix pomatia* (Hp), and *Sepia officinalis* (So, FU-h). The numbers in red correspond to the percent identity between like subunits (FU-a vs. FU-a, etc). Each box contains the percent identity and  $\sim d_{ij}$  (calculated from the percent identity using this equation:  $\sim d_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).

	OdA	OdA	OdA	OdA	OdA	OdA	OdA	Mc	Hp	Hp	So
	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>	<b>c</b>	<b>d</b>	<b>g</b>	<b>h</b>
Mc <b>c</b>	41% .891	43% .843	54% .616	40% .916	44% .820	40% .916	46% .776		44% .820	45% .798	43% .843
Hp <b>d</b>	39% .941	40% .916	41% .891	50% .693	41% .891	41% .891	42% .867			43% .843	40% .916
Hp <b>g</b>	42% .867	42% .867	44% .820	42% .867	44% .820	42% .867	52% .653				50% .693
So <b>h</b>	40% .916	39% .941	38% .967	37% .994	40% .916	41% .891	70% .356				

Table 15. Functional Unit Percent Identity: *Haliotis tuberculata* type 1 (Ht1), *Haliotis tuberculata* type 2 (Ht2), *Octopus dofleini* G-type (OdG), *Megathura crenulata* (Mc), *Helix pomatia* (Hp), and *Sepia officinalis* (So, FU-h). The numbers in red correspond to the percent identity between like subunits (FU-a vs. FU-a, etc). Each box contains the percent identity and  $\sim d_{ij}$  (calculated from the percent identity using this equation:  $\sim d_{ij} = -\log(1 - d_{ij})$ , where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table).

	OdG <b>a</b>	OdG <b>b</b>	OdG <b>c</b>	OdG <b>d</b>	OdG <b>e</b>	OdG <b>f</b>	OdG <b>g</b>	Mc <b>c</b>	Hp <b>d</b>	Hp <b>g</b>	So <b>h</b>
Mc <b>c</b>	41% .891	43% .843	54% .616	40% .916	43% .843	41% .891	46% .776		44% .820	45% .798	43% .843
Hp <b>d</b>	39% .941	41% .891	40% .916	51% .673	41% .891	41% .891	42% .867			43% .843	40% .916
Hp <b>g</b>	42% .867	43% .843	43% .843	42% .867	42% .867	41% .891	51% .673				50% .693
So <b>h</b>	40% .916	40% .916	38% .967	37% .994	39% .941	41% .891	69% .371				

Each of the comparison tables shows the percent identity between the functional units and below that the  $d_{ij}$  calculated from the percent identity using this equation:

$$\tilde{d}_{ij} = -\log (1 - d_{ij})$$

where  $d_{ij}$  is equal to the percent of amino acid differences/100 (or the fractional difference), calculated from  $(1 - f_{ij})$ ,  $f_{ij}$  being the percent identity /100 (or the fractional identity), which is the top number in the table. The use of  $\tilde{d}_{ij}$  approximately accounts for multiple substitution and back mutation.(11) These numbers can be plotted against a fossil record timeline, the width of the line corresponding to presumed divergence time as given by the fossil record.

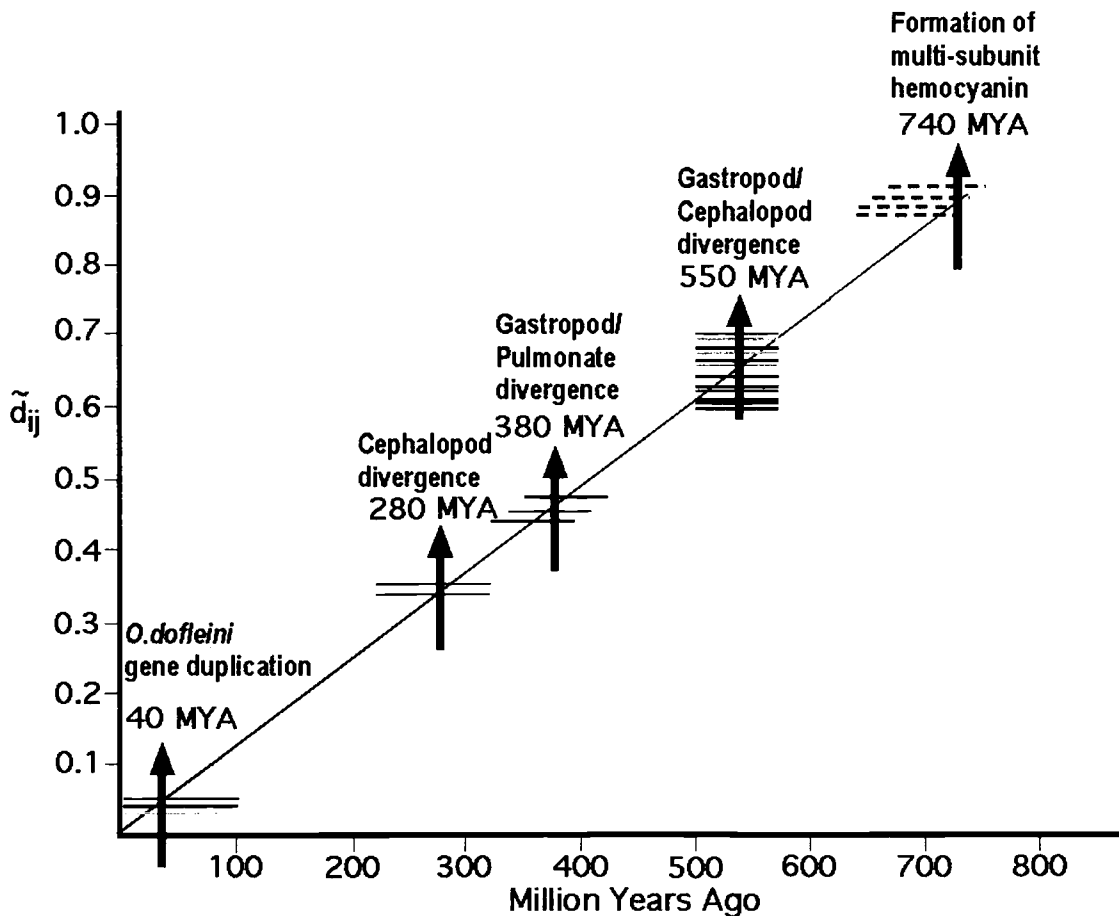


Figure 16. Estimated Time of Divergence for Several Taxonomic Groups. The purple bars are comparisons made using FU-g, the green was used for FU-d, and the yellow for FU-c. The dashed black bars represent comparisons made across all of the possible subunits from the species used in Tables 3-15. The width of the lines for the Cephalopod divergence and the Cephalopod/Gastropod divergence represents the estimated time for that divergence given by the fossil record. The slope line was drawn using those two points and the origin, giving a rough measure of when several other gene divergence events took place.

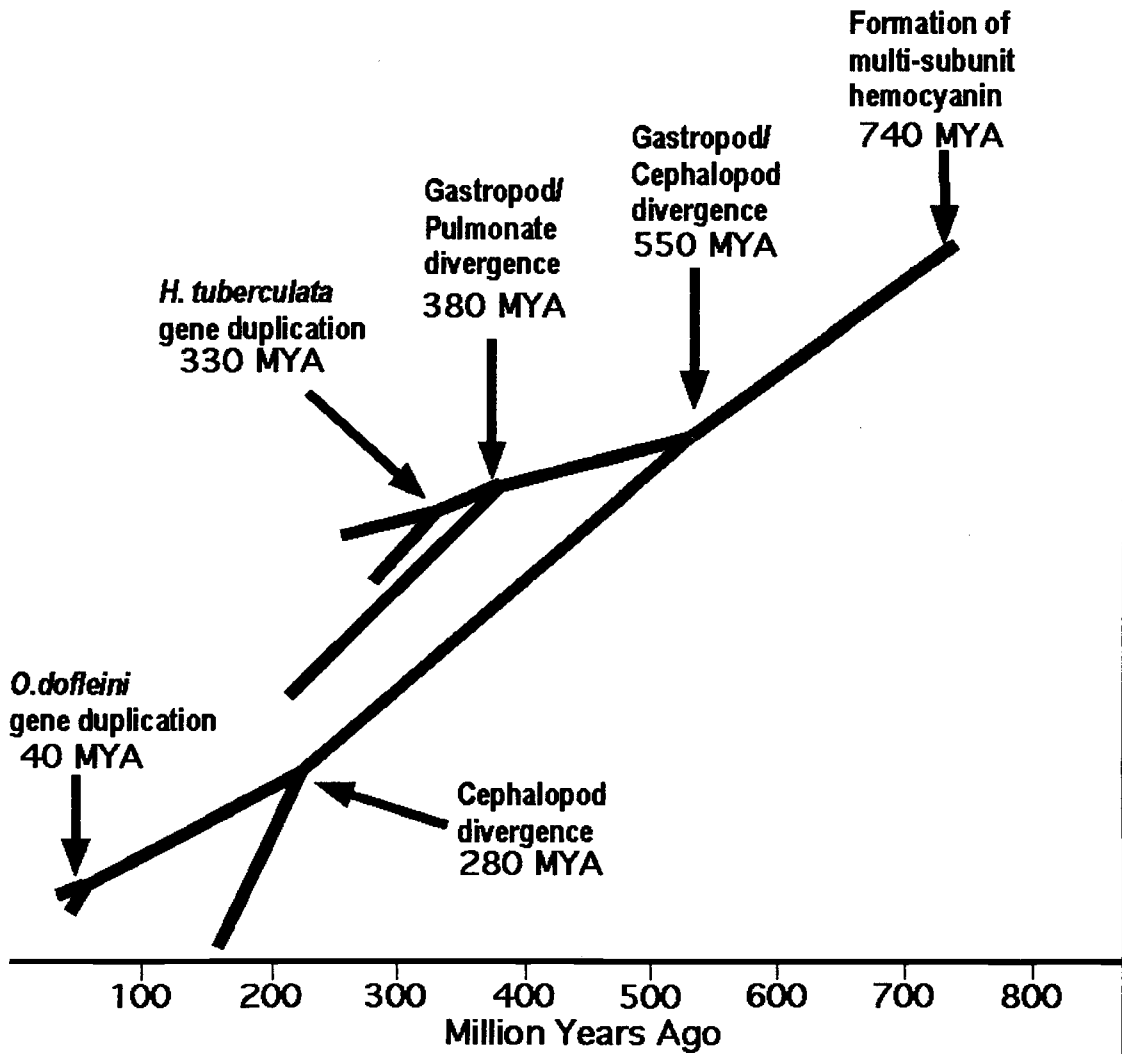


Figure 17. Divergence Tree. This tree gives a visual way to think about the estimated time of divergence for these taxonomic groups. The x-scale is the same as that for Figure 16.

## 4. Conclusion

### 4.1 Paralogous Genes

It is quite clear from all of the evidence presented that the A- and G-type genes for *O. dofleini* hemocyanin are recently diverged paralogues, and not alleles. Two gene forms are also known to exist for *Haliotis tuberculata*, *Megathura crenulata*, and *Sepia officinalis*. At first glance, one would expect the duplication to have occurred back before the divergence of the taxa represented by these organisms (Class Gastropoda and Class Cephalopoda). However, the percent identity between the paralogues of each organism yields different results. The *O. dofleini* paralogues have 95-97% identity and the *H. tuberculata* paralogues have 59-73% identity, with 66% average identity. This surprising result eliminates the possibility that the gene duplication occurred before the Cephalopod-Gastropod split. In fact, we estimate the duplication of the *O. dofleini* hemocyanin gene at roughly 40 MYA. See Figure 16.

### 4.2 A- & G-type Microsatellites in *Octopus dofleini* hemocyanin

So far, these two microsatellites are the only known microsatellites within a hemocyanin gene. The insertion of at least part of the microsatellite must have occurred before the gene duplication for *O. dofleini*. This is demonstrated by the fact that each of the paralogues has a microsatellite and because the two have maintained such a high degree of identity in nucleotide sequence (71% overall, 85% if the gaps and their corresponding regions are removed, 92% if all of the tandem repeats are ignored). It is expected that the microsatellite region would mutate at a higher rate

than the exons, because they are not coding for anything, and at a higher rate than the introns because of the slippage due to the tandem repeats, which are absent in the regular introns.

### 4.3 Comparisons Among Corresponding Functional Units

Functional Units (FUs) for two cephalopods, two gastropods, and one pulmonate are currently available on GenBank. Comparing the FUs across species, shows us how long ago they diverged from one another when plotted against the information gained from the fossil record. Tables 11-13 in Chapter 3 represent the percent identity for every FU-c, FU-d or FU-g (or FU-h in the case of *S. officinalis*) for which complete protein sequences exists. Excluding the percent identity found between paralogues, the two cephalopods *O. dofleini* and *S. officinalis* have the highest percent identity (69-70%), the two gastropods, *H. tuberculata* and *Megathura crenulata* are next (63-72% identity), gastropod-pulmonate (*H. tuberculata* and *Helix pomatia*) are third (61-64%), cephalopod-gastropod are fourth (50-54%), and cephalopod-pulmonate are last (50-52%).

This order is the exact divergence order shown on Figure 16, except for the cephalopod-pulmonate comparison. The pulmonates diverged from the gastropods after the gastropods and cephalopods had diverged. So, the last common ancestor for the pulmonates and cephalopods was at the cephalopod-gastropod split. See Figure 17. It therefore makes sense that the pulmonates would be more diverged from the cephalopods, than the gastropods, the group it split away from more recently. This accounts for the low percent identity seen for the cephalopod-pulmonate comparison.

Tables 3-15 give the percent identity comparisons for every FU for which full sequence is available. Any comparison made within a subunit, therefore between FUs of differing letters, always results in 35-48% identity regardless of which species the subunit is from. This implies that the multi-FU subunit is on the order of 740 million years old.



## Bibliography

1. van Holde, K.E., and Miller, K. I. (1995) *Hemocyanins*. *Advances in Protein Chemistry* **47**, 1-81.
2. Willmer, Pat. (1990) *Invertebrate Relationships: Patterns in Animal Evolution*. Cambridge University Press. pgs. 263-267.
3. van Holde, K.E., Miller, K. I., and Decker, H. (2001) *Hemocyanins and Invertebrate Evolution*. *Journal of Biological Chemistry* **276**, 15563-15566.
4. Burmester, Thorsten. (2001) *Molecular Evolution of the Arthropod Hemocyanin Superfamily*. *Mol. Biol. Evol.* **18**, 184-195.
5. Lang, Walter H., (1990) *cDNA Cloning and Sequencing of Octopus dofleini Hemocyanin*. Doctoral Thesis, Oregon State University.
6. Lieb, B., Altenhein, B., Markl, J., Vincent, A., van Olden, E., van Holde, K. E., and Miller, K. I. (2001) *Structure of two molluscan hemocyanin genes: Significance for gene evolution*. *PNAS* **98**, 4546-4551.
7. van Holde, K.E., Miller, K. I., and Lang, W. H. (1992) *Molluscan Hemocyanins: Structure and Function*. *Advances in Comparative and Environment Physiology* **13**, 257-300.
8. Swerdlow, R. D., Ebert, R. F., Lee, P., Bonaventura, C., and Miller, K. I. (1996) *Keyhole Limpet Hemocyanin: Structural and Functional Characterization of Two Different Subunits and Multimers*. *Comp. Biochem. Physiol.* **113B**, 537-548.
9. Personal correspondance from Gisèle Préaux and Constant Gielens to Karen I. Miller.
10. Kruglyak, S., Durrett, R. T., Schug, M. D., and Aquadro, C. F. (1998) *Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations*. *PNAS* **95**, 10774-10778.
11. Kimura, Motoo. (1987) *Molecular Evolutionary Clock and the Neutral Theory*. *J. Mol. Evol.* **26**, 24-33.
12. Lamy, J., Gielens, C., Lambert, O., Taveau, J. C., Motta, G., Loncke, P., De Geest, N., Préaux, G., and Lamy, J. (1993) *Further Approaches to the Quaternary Structure of Octopus Hemocyanin: A Model Based on Immunoelectron Microscopy and Image Processing*. *Archives of Biochemistry and Biophysics* **305**, 17-29.

13. Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J., and Deka, R. (1997) *Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci*. PNAS **94**, 1041-1046.
14. Doyle, P. (1994) *Phylogeny and Systematics of the Coleoidea*. The University of Kansas Paleontological Contributions **5**, 1-15.
15. Clarkson, ENK. (1998) *Invertebrate Paleontology and Evolution*, 4<sup>th</sup> Edition Blackwell Science Ltd. pg. 234.
16. Miller, K. I., Cuff, M. E., Lang, W. H., Varga-Weisz, P., Field, K. G., and van Holde, K. E. (1998) *Sequence of the Octopus dofleini Hemocyanin Subunit: Structural and Evolutionary Implications*. J. Mol. Biol. **278**, 827-842.
17. Cuff, M. E., Miller, K. I., van Holde, K. E., and Hendrickson, W. A. *Crystal Structure of a Functional Unit from Octopus Hemocyanin*. J. Mol. Biol. **278**, 855-870.