Abstract

Thompson, Denis Brian. FINDING HOMOLOGOUS GENES WITH PRIMERS DESIGNED USING EVOLUTIONARY MODELS. (Under the direction of Henry Schaffer.)

Genes homologous to a set of known, aligned, genes can be found by screening DNA libraries with PCR. PCR primers for such screens are commonly designed via a method described by Sells and Chernoff (1995). This standard design method does not make use of information about the evolutionary relationship between the known genes. The present study investigated the efficacy of using information about evolutionary relationships (inferred from the sequence data) in the design of PCR primers. This study compares the standard primer design method (represented herein by a modified multinomial distribution) with evolutionary model based primer design methods. The primer design method that, given an alignment of known sequences with one sequence left out, assigned a higher probability, on average, to the left-out sequence, was defined as the better method. By this measure of relative performance, an evolutionary model based primer design method sensitive to states correlated across sites of a sequence, outperformed the standard method, on the alignments studied.

FINDING HOMOLOGOUS GENES WITH PRIMERS DESIGNED USING EVOLUTIONARY MODELS

by
DENIS BRIAN THOMPSON

A dissertation
submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIOMATHEMATICS

Raleigh

2003

APPROVED BY

Henry Schaffer, Chair of Advisory Committee

Michael Purugganan

Charlie Eugene Smith

Jeff Thorne

This dissertation is dedicated to my loving parents.

—D.T. 2003

Personal Biography

Denis moved to North Carolina in 1980.

Acknowledgments

I want to express my appreciation to a few people who helped me bring this dissertation to completion, and who enriched my life during graduate school.

An important part of my life in the last 3 years has been the delightful companionship I've shared with Laurie. I am deeply grateful to Laurie for being on my side, for giving me smart and truly helpful advice, and for giving me caring support. All these gifts have had a positive effect on my grad school experience and on other aspects of my life. But mainly all the time we've spent together means a lot to me.

To my dear family I want to say how much I appreciate your love and support. I appreciate y'all most of all for being there and for being yourselves throughout this grad school process. One relatively small but noteworthy part of all this support was the C++, Igor, and other computer help, including eleventh hour debugging help, I received from my dear brothers.

I thank Jeff Thorne for generously meeting with me some in Summer 2001, and from January 2002 until the completion of this project. The final form of the science of this dissertation is based largely on Jeff's vision. (I did all the research in this dissertation after January 2001.)

I am deeply grateful to Henry Schaffer for noticing in May 2001 (when he attended my seminar) that I was floundering, and for going out of his way to e-mail me, offer help, and then make some phone calls to arrange help for me. Other people must have seen or known I was in trouble, but no one else reached out.

And I give even more thanks to Henry for managing me from April 2002 until the end of the project, and for jumping into the formal role of advisor in September 2002, and for editing this dissertation. I can't thank Henry enough for generously investing time and thought and

energy in my education, and for believing in me.

I also want to thank Henry for some of the most productive and educational discussions about the actual content of Biomathematics I have had in school. These interactions have been what graduate school is supposed to consist of.

This dissertation would not have happened without Jeff and Henry.

A smart person recently told me that the community of fellow students is most educational part of school. For me a big contributor in this respect was Doug Robinson. Doug was a gigantic help during the months I worked on this version of my project. He let me look at his code so I could figure out some algorithms, and let me copy chunks of 3 or 4 of his codon-handling functions. He directed me to the C++ TCL matrix objects. He gave me invaluable help in late September 2003 when I thought I had a bug in my code and needed to figure it out within 24 hours. And he answered e-mail questions from me at other times. Doug also generously explained several issues about evolutionary models, and other bioinformatics concepts. Doug amazingly always responded to my questions and requests in an upbeat and positive way. I can't thank Doug enough for all these particular pieces of help, and in general for being someone I could ask "dumb" questions of.

I want to express my appreciation for some of the precious friends I have had outside of school during graduate school. I treasure my wonderful friendship with Carol. And I am likewise glad for and thankful for many dear friends who have made my life richer and more rewarding in these past few years: Val, Sam, Paul, Leslie, Amy, Marcia, Samantha, Penny (and thanks for the rides into campus in spring 2002), Elise, Harlan, and Corey. Also thanks to Bob, Elisabeth, Jane, Lisa, and Carmen.

I want to thank a few more people who gave me assistance at school: Tim Elston for offering help and talking to me in December 2001 and January 2002, when I approached him in his

role as Director of Graduate Studies; Jim Selgrade for his help in committee reassignments; and Jackie Dietz for suggestions about statistical tests. I thank Terry Byron for being helpful and really caring about stat and biomath students' work environments and computer resources. And thanks to Ann Ethridge for caring about my welfare all along, and for inducing me to take action to improve my situation in December 2001.

A few friends have especially enriched my life around school in these grad school years. These people made my time in the computer lab, in the grad student offices, in class, and also away from school, more fun and connected and real: Russell, Dan, Sarah, Teri, Liz, Jason, Cindy, Brian, Marta, and Virginia. These friends have meant a great deal to me. (Russell showing up out of the blue a few days before my defense is perhaps the happiest surprise I've ever had.)

Finally, I'm glad for Irene's timely questions, for her getting me to think, and for her helping me get work done near the last months of the project. But mostly I appreciate her caring about me.

With much appreciation,

Denis

Table of Contents

		page
List of Tables		x
List of Figures		xii
List of Symbols and Abbreviations		xv
Chapter 1	Introduction	1
	Example of contemporary primer design process	1
	The standard primer design method	4
	Definition of the problem to be addressed	9
	Weaknesses of the standard method	10
	Improving one aspect of primer design —prediction	15
	Algorithms and programs for primer design	17
	Evolutionary model based prediction of new sequences —derivation of equations	24
	Derivation of the performance measure for site-by-site prediction method	30
	A mathematical model of the standard primer design method	38
	Further justification of using a probability distribution to represent the standard primer design method	43
	Comparing prediction methods	44
	Comparing sets of these values with the Wilcoxon-Nemenyi-McDonald-Thompson two-sided all-treatments multiple comparison procedure	47
	Other ways of comparing performance measures	50
	Tests in the amino acid realm are meaningful and interesting	53

Chapter 2 Comparison of evolutionary model based prediction methods and standard method —relative performance at predicting a single related amino acid sequence 57

Descriptions of the three alignments used in this study . . . 57

Breaking the alignments into segments, and results of the statistical test 64

Comparison of the standard primer design method and the single site evolutionary model based method. 68

Summary of comparing differences between performance measures between standard method (multinomial) and single site evolutionary model method 73

Multiple-site information 74

A multisite evolutionary model based method 77

Does the size of the segments affect results? 81

Comparison of the multisite evolutionary model based method and the standard primer design method 83

Summary of comparing differences between performance measures between multisite evolutionary model method and standard method 87

Comparison of the multisite evolutionary model based method and the single site evolutionary model based method 89

Summary of comparing differences between performance measures between multisite evolutionary model method and single site evolutionary model method 93

Correlation of P(LO_seq|LI_set)'s within a segment 95

Chapter 3 Using Information in Clusters 98

Description of the prediction algorithm that makes use of cluster information 99

Results of comparing three prediction methods 100

Comparison of the prediction methods on trees of different average evolutionary distances 104

Creating the Simulated Data 104

	Interpretation of Figures	106
Chapter 4	Comparison of pool construction methods	111
	Detailed description of the all-degenerate-primer pool construction method	113
	Detailed description of the "whole tree" pool construction method	114
	Detailed description of the "one subpool per attachment point" pool construction method	116
	Results	117
	Does pool size affect the "fraction of pool with few mismatches" performance measure?	126
Chapter 5	Conclusions from this research	132
	Directions for future research	136
References	. . . . .	139
Appendix A	Derivation of model to predict related sequences given the species relationships	142

List of Tables

		page
Table 1.1	Bassett et al used sequences from these proteins to design primers	1
Table 1.2	Conserved segments used by Bassett et al	2
Table 1.3	Bassett et al's upstream primer pool	3
Table 1.4	Bassett et al's downstream primer pool	4
Table 1.5	Length of PCR primers used in various research	5
Table 1.6	Segments with low degeneracy are uncommon	11
Table 1.7	Number of citations of primer-design software	23
Table 1.8	Number of citations of primer-design methods papers	23
Table 1.9	Comparison of multinomial and modified multinomial	41
Table 2.1	Results of Wilcoxon-Nemenyi-McDonald-Thompson tests on the G3PD alignment data	65
Table 2.2	Results of Wilcoxon-Nemenyi-McDonald-Thompson tests on the alpha globin and beta globin alignment data	66
Table 2.3	Results of Wilcoxon-Nemenyi-McDonald-Thompson tests on the ribosomal protein L20 alignment data	67
Table 2.4	Summary of comparisons of which method has greater performance measures	74
Table 2.5	Summary of comparisons: Which method has greater performance measures?	88
Table 2.6	Summary of comparisons of which method has greater performance measures	94
Table 3.1	Average Pairwise Percent Identities of sequences within each cluster	98

Table 3.2	Average percent identities of each prediction method	101
Table 3.3	Cluster method comparisons Results of Wilcoxon-Nemenyi-McDonald-Thompson tests on the G3PD alignment data . .	103
Table 3.4	aa APPI for each tree	105

List of Figures

		page
Figure 1.1	Hypothetical Trees	26
Figure 1.2	Attachment Points	27
Figure 1.3	Histogram of Probabilities	47
Figure 1.4	Differences of average $\mathbf{P}(\mathbf{LO\_seq}\|\mathbf{LI\_set})$'s are not normally distributed	48
Figure 1.5	Log scale histogram of a set of differences	51
Figure 2.1	A subset of the Goldman group's G3PD phylogenetic tree	59
Figure 2.2	Phylogenetic tree showing relation of amino acid sequences in concatenated alpha globin and beth globin alignment . . .	62
Figure 2.3	The phylogenetic tree of the 11 sequence ribosomal protein L20 alignment used in the analysis	63
Figure 2.4	Comparison of standard prediction method and evolutionary single-site prediction method for G3PD alignment	70
Figure 2.5	Comparison of performance measures for the two prediction methods, on the alpha globin and beta globin alignments . .	71
Figure 2.6	Log-scale histogram of differences in probabilities	73
Figure 2.7	Comparison of analyses done on different segment lengths	82
Figure 2.8	Log-scale histogram of differences in probabilities —G3PD alignment	83
Figure 2.9	Log-scale histogram of differences in probabilities —globin alignments	85
Figure 2.10	Log-scale histogram of differences in probabilities —ribosomal protein L20 alignment	87
Figure 2.11	Log-scale histogram of differences in probabilities —G3PD alignment	89
Figure 2.12	Log-scale histogram of differences in probabilities —globin alignments	91

Figure 2.13 Log-scale histogram of differences in probabilities —ribosomal protein L20 alignment 93

Figure 2.14 Location, in alpha and beta globin alignments, of sequences for which the single site evolutionary model method predicted significantly better than the multinomial method predicted 96

Figure 2.15 Location, in r. p. L20 alignment, of segments for which the single site performance measure predicted significantly better than the multinomial method 97

Figure 3.1 Percent identities are not normally distributed 102

Figure 3.2 Relative performance of three prediction methods on alignments with different APPI's —segments of length 28 aa's 107

Figure 3.3 Relative performance of three prediction methods on alignments with different APPI's —segments of length 20 aa's 108

Figure 3.4 Relative performance of three prediction methods on alignments with different APPI's —segments of length 14 aa's 109

Figure 3.5 Relative performance of three prediction methods on alignments with different APPI's —segments of length 7 aa's 110

Figure 4.1 alpha and beta globin phylogenetic tree with codon branchlengths 118

Figure 4.2 Comparison of 3 pool construction methods on the alpha globin alignment 118

Figure 4.3 alpha globin pool performance measure vs pool size, for the three pool construction methods 119

Figure 4.4 Comparison of three pool construction methods 121

Figure 4.5 Three histograms for the beta globin data 122

Figure 4.6 Newick representation of the G3PD tree with branch lengths in units of expected nucleotide substitutions per codon . . . 123

Figure 4.7 G3PD pool performance measures vs. APPI 124

Figure 4.8	3 Histograms of “pool fraction” performance measure for G3PD data	125
Figure 4.9	Newick representation of the phylogenetic tree for the ribosomal protein L20 alignment	125
Figure 4.10	Pool fraction vs. Pool size, for alpha globin gene	128
Figure 4.11	Influence of pool size on performance in beta globin . . .	129
Figure 4.12	Pool size and performance, G3PD data	130
Figure 4.13	Pool size and performance, ribosomal protein L50 data . .	131

List of Symbols and Abbreviations

Symbol or Abbreviation	Meaning
AA or aa	Amino Acid
APPI	average pairwise percent identity
cDNA	complementary DNA
Eq.	Equation
G3PD	Glyceraldehyde 3 phosphate dehydrogenase
LI_set	left-in set of sequences
LO_seq	left-out sequence
LHS	Left-hand side of an equation
nt	nucleotide
PCR	Polymerase chain reaction
RHS	Right-hand side of an equation
rp L20	ribosomal protein L20

Chapter 1

Example of contemporary primer design process

Basset et al. (2002) wanted to find the gene for the ethylene receptor, ETR1, in the peach, *Prunus persica*. The ethylene receptor gene had previously been found and sequenced in several species. (See Table 1.1.) Basset et al. predicted that the peach sequence would be similar to the known sequences for this gene in the other species. So to learn the exact sequence in the peach, Basset et al. followed a process familiar to experimental biologists: they designed PCR primers (oligonucleotides) that would bind to DNA sequences coding for parts of the known ethylene receptor amino acid (aa) sequences, used the primers in a PCR reaction with peach genomic DNA, cloned the fragment amplified in the reaction, and screened a peach cDNA library for sequences complementary to the cloned fragment. Using this method they successfully found the sequence of the peach ETR1 gene.

They designed these PCR primers using an alignment of seven known amino acid sequences —sequences of proteins known to bind to ethylene. The proteins in the alignment they used are listed in Table 1.1.

Table 1.1 **Bassett et al. used sequences of these proteins to design primers**
1. ethylene responsive factor from rice 2. putative ethylene receptor from the carnation 3. ethylene response sensor from *Arabidopsis* 4. ethylene receptor from *Arabidopsis* 5. putative ethylene receptor from the tomato 6. ETR1 homolog from the tomato 7. ethylene response sensor from *Rumex palustris* (common name "Marsh Dock")

From the alignment of these aa sequences, they chose two short stretches —one of 6 aa's, one of 7 aa's— of high homology. The two short segments of the alignment they used are shown in Table 1.2

Table 1.2
Conserved segments used by Basset et al.

The parts of the alignment of the known protein sequences that Bassett et al. designed PCR primers to hybridize to the nucleotide sequence for. In this figure, the six aa's in the upstream site are separated from neighboring sequences on each side with a space, as are the seven aa's in the downstream primer site. These spaces do not indicate a gap in the sequences. "ERS" means "ethylene response sensor." ETR1 means "ethylene receptor."

protein	upstream primer site	downstream primer site
rice ERS	LM LVHIIP DLL	DFLA VMNHEMR TPMN
carnation ERS	LW LVYIIP DLL	EFLS VMNHEMR TPIH
Arabidopsis ERS	LM LVHIIP DLL	DFLA VMNHEMR TPMH
Arabidopsis ETR1	LM LVHIIP DLL	DFLA VMNHEMR TPMH
tomato ERS	LM LVHIIP DLL	DFLA VMNHEMR TPMH
tomato ETR1	LM LVHIIP DLL	DFLA VMNHEMR TPMH
Rumex ERS1	LM LVHIIP DLL	DFLA VMNHEMR SAMH

In the downstream segment there is 100% identity between the seven known aa sequences. In the upstream segment the carnation sequence is different from the other aa sequences at one of the six aa sites; otherwise there is 100% identity between the seven sequences here.

Pool of "degenerate primers"

For each of these two short segments of high homology that Basset et al. identified, they designed a set (a "pool") of oligonucleotides that would bind to all or many of the DNA sequences that could code for the conserved aa sequence. That is, for each amino acid occurring at a particular site in the aa sequence, Basset et al. included, at the corresponding site in some of the oligonucleotides in their pool, every codon in the genetic code that codes for that aa.

So, for example, if there were a sequence of 3 aa's, and there were two codons in the genetic code for each of the 3 aa's in this sequence, then the "completely degenerate" primer pool would contain oligonucleotides of 2 x 2 x 2 = 8 different sequences.

The reason to have completely degenerate set of primers in the pool is to make the primers able to amplify the sequence in a related species, even if different species code for the aa sequence using different codons. (The assumption is that the related species will have the same aa sequence, but perhaps not the same nucleotide sequence.) Different species might code for a particular aa at a particular site with different codons because the species have different codon preferences, or just because of chance.

The primers Basset et al. designed, shown in Tables 1.3 and 1.4, do not contain every possible codon that might code for the corresponding amino acids. But in most cases they do. See the row labeled "Fraction of codons coding for this aa represented in primer pool" in Figures 1.3 and 1.4.

Table 1.3
Basset et al.'s upstream primer pool

The molecules present in the primer pool are shown on the sense row. Nucleotides within parentheses are the different degenerate sequences. In the nucleotide sequence, "I" indicates inosine, a synthetic nucleotide that when incorporated into DNA will pair with any of the four natural nucleotides (Watanabe et al. 2001, Bartl 1997). See comments in text on guanine residues at 5' and 3' ends.

sense: 5'-	G	CTI	GTI	CA(C/T)	AT(A/C/T)	AT(A/C/T)	CCT	G	-3'
aa 3-letter symbol		Leu	Val	His	Ile	Ile	Pro		
aa 1-letter symbol:		L	V	H	I	I	P		
fraction of codons for this aa represented in primer pool		4 of 6	4 of 4	2 of 2	3 of 3	3 of 3	1 of 4		

Basset et al. include a guanine at the 5' end of their upstream primers. The aa residues in the

known aa sequences, in the position corresponding to this guanine, are tryptophan (W) and methionine (M). Each of these aa's is coded for by one codon. And both of those codons end in a G. So the guanine would bind to every nucleotide sequence seen in the given alignment.

(It is unclear why Basset et al. include a guanine (G) residue at the 3' end of their upstream primer. The amino acid residues, in the known aa sequences, in the position corresponding the the guanine at the 3' end, are threonine and serine. No codon coding for either of these aa's begins with a G.)

Table 1.4
Basset et al.'s downstream primer pool

The set of molecules in the primer pool (i.e. that are included in the PCR reaction mixture) are shown on the antisense row. The sequence of codons translated into aa's is shown on the sense row. In the nucleotide sequence, "I" indicates inosine. (The rightmost two nucleotides in the primer, "TC" in the antisense strand, will bind with two of the 6 codons for serine, in addition to binding to the two codons for arginine. Only arginine, R, appears in the corresponding site in the known aa sequenes. See Table 1.1.)

sense: 5'-	GTx	ATG	AA(T/C)	CAT	GA(G/A)	ATG	AG	-3'
antisense: 3'-	CAI	TAC	TT(A/G)	GTA	CT(C/T)	TAC	TC	-5'
aa 3-letter symbol	Val	Met	Asn	His	Glu	Met	Arg	
aa 1-letter symbol:	V	M	N	H	E	M	R	
fraction of codons for this aa represented in primer pool	4 of 4	1 of 1	2 of 2	1 of 2	2 of 2	1 of 1	2 of 2	

The standard primer design method

In the research described above, Basset et al. used a primer design method that is promoted by Sells and Chernoff in a methods paper (1995). In that paper Sells and Chernoff sum up the primer design step with this description: "The optimum primers would be based on an invariant stretch six to eight amino acids, [of] low (i.e. <1024-fold) degeneracy, and would

have about 50% GC (guanine:cytosine) content. Rarely can such primers be designed." I will refer to this method as the "standard" primer design method, because it is widely used.

Optimum length of primers

Sells and Chernoff do not cite any science proving that 18 to 24 nucleotides (the length corresponding to six to eight aa's) is the optimum length for a PCR primer. Nor do they even give their reason for advocating this length. But this length is used by many researchers. Table 1.5 shows a small sample of papers that used PCR to screen for genes related to a known alignment of genes, and the lengths of primer used.

Table 1.5
Length of PCR primers used in various research
These PCR primers were all used to screen for related genes.

Paper	Length of primers, in nucleotides	Length of corresponding peptide
Basset et al. 2002	20	7, 8 aa's
Chen et al. 1998	18	6 aa's
Jones et al. 1995	29, 33	10, 11 aa's
Kirimura et al. 1999	23, 24	8 aa's
Oshima et al. 2002	21, 24	7, 8 aa's
Venugopal et al. 2002	25	9 aa's
Wünschiers et al. 2001	20, 21	7 aa's

Although Sells and Chernoff do not list them, there are easily discernible design constraints that make "too short" primers undesirable, and other design constraints that make "too long" primers undesirable.

Any time a primer is made shorter, it becomes more likely to bind to a gene other than the one

the researcher is interested in. Löffert et al. (1997), in a methods paper about PCR primers, discourage the use of primers of length 15 nucleotides or shorter for this reason.

Longer primers are more difficult to design because longer stretches of highly-conserved sequence are more rare than short stretches of highly-conserved sequence. Using a longer primer also makes it more likely that the sequence of interest will have evolved into a different sequence in the species the researcher is attempting to find the gene in.

Longer primers are also undesirable because, if one is using a pool of primers representing all degenerate sequences coding for an aa sequence, then the longer the primer, the smaller the concentration (in the PCR reaction) of the primer with the correct sequence. For example, suppose a researcher wants to probe for an amino acid sequence that is 7 amino acids long. And suppose that 3 of these aa's are each coded for by 4 codons, and the remaining four aa's are each coded for by 2 codons. Then the completely degenerate primer pool would contain representatives of $4^3 \cdot 2^4 = 1024$ different oligonucleotide sequences. Compare that 7 aa long primer pool with the pool for a sequence 11 aa's long, 5 aa's of which are each coded for by 4 codons, and six aa's of which are each coded for by 2 codons. In this second case the completely degenerate primer pool would contain representatives of $4^5 \cdot 2^6 = 65{,}536$ different oligonucleotide sequences.

The researchers listed in Table 1.5, and Sells and Chernoff, have decided that 18 to 24 nucleotides is the approximate optimum range. Bartl (1997) in a methods paper essentially agrees. She advocates using primers of a length 18 to 30 nucleotides. But her methods paper is about using inosines to allow one to construct longer primers that will bind to many degenerate sequences, but do not contain so many different-sequence primers, so do not have to be so dilute.

The advantage of low degeneracy

Sells and Chernoff advocate the use of degenerate primers for the reasons discussed above in the section titled "Pool of degenerate primers."

Sells and Chernoff advocate finding low degeneracy primers so the correct primer will not be too diluted. Bassett et al. designed for low degeneracy when they chose the site for their downstream primer to be an aa sequence containing two methionine residues among the seven aa's. Only one codon codes for methionine. Making a pool of primers "completely degenerate" for a methionine at one site results in multiplying the number of different-sequence primers in the pool by a factor of one.

GC Content

Sells and Chernoff write that 50% GC content is desirable because that content gives the primer-target DNA hybrid a predictable melting temperature. This characteristic is not necessary for success of amplification. Times and temperatures of the PCR reaction can be adjusted so the reaction is successful for different primer-target DNA melting temperatures.

Optimizing all the design desiderata

After describing their ideal primers, Sells and Chernoff note that "rarely can such primers be designed." They make this statement because it can be difficult to simultaneously optimize the design requirements they list, while also optimizing other design requirements of PCR primers. (Other design requirements are discussed below.) For example, the alignment of known genes might not contain a completely invariant stretch of seven amino acids. Or the only invariant stretch might contain a high proportion of amino acids coded for by 4 or 6 codons each, making the pool of degenerate primers not meet their degeneracy requirement. (The aa's Ser, Leu, Pro, Arg, Thr, Val, Ala, and Gly are each coded for by four or six codons.)

Sells and Chernoff point out that the ideal conditions they list are condition to strive for, but

that "there are no infallible rules to guide the final selection."

The number of sequences in the given set

Usually when designing primers to screen for related genes, researchers use an alignment of several sequences to design the primers. One example of this method is the Basset et al. example presented above. Other examples are: Wünschiers et al. (2001) used several prokaryotic sequences to design primers to find Fe-hydrogenase in a green alga; Venugopal et al. (2002) used "all available fish growth hormone protein sequences" to identify a conserved region and design a primer to bind to it, to find the gene in carp; Jones et al. (1995) designed primers based on an alignment of sequences from eight species.

Other researchers have used just two genes to identify conserved regions and design their primers to find related genes. For example Kirimura et al. (1999) used two sequences —one from a lily and one from a yeast— to design a primer to find a gene in the fungus *Aspergillus niger*.

The number of species the "new" sequence is sought in

In the examples given so far, the researchers' goal was to find the gene in one particular species. Sometimes the researcher has the aim of finding a gene in multiple species. For example Gould et al. (1989) used one set of PCR primers to find the succinate dehydrogenase (SDH) gene in humans, mice, rats, *Xenopus*, *Drosophila*, *Arabidopsis*, two species of yeast, and *Dictyostelium*. The primers were designed based on conserved sequences identified by comparing the SDH amino acid sequence in just two species: *E. coli* and bovine.

Gould et al. (1989) were interested in comparing a gene in multiple well-studied species. Today, the genome sequences of many well-studied species are available online, so one would

not have to chemically screen for gene sequences in these species. As time goes on, researchers will have less need to chemically screen for genes in commonly studied species. As more complete genomes become available, researchers will be more able to simply search sequence databases.

But there will still be the need for chemical screening because, of course, some researchers study species for which the genomes are not available online. For example Kevin Moulton screened for genes in multiple non-*Drosophila* fly species. (Kevin Moulton, 2002, personal communication.)

Definition of the Problem to be addressed

The problem stated as given-find

Above are examples and descriptions of the primer design problem. Here is a more precise expression of the primer design problem this dissertation addresses:

Given: a set of n aligned, related, protein-coding sequences (either nucleotide, aa, or codon) of length k.

Find: the one sequence, out of all possible sequences, that is "most likely" to be the sequence of a new member of the set ("new" defined just below).

The Basset et al. example above is an example of an experimentalist solving this problem.

Meaning of terms in given-find statement

By a "new member of the set" I simply mean another sequence related to the given sequences, that is not yet considered to be a member of the set. So the sequence is new to the set in question. "New" does not indicate the sequence is newly evolved. Rather, the sequence is not yet found.

The new sequence is the sequence that is sought experimentally with the PCR reactions. The probability that is used to judge which sequence is "most likely" is not adjusted based on whether the gene is sought with PCR or another laboratory screening procedure.

This dissertation addresses this problem only in protein-coding sequences. Calculations are done in the amino acid realm in Chapters 2 and 3, and in the codon realm in Chapter 4.

I will sometimes refer to the given set of sequences as the "known" sequences.

Further specification of the problem

I think of each sequence in the known set as belonging to a different species, although the methods presented here would probably be valid to apply to duplicated genes within the same species. (This dissertation does not attempt to apply the methods presented here to the problem of screening for related genes within one species.) I also think of the unknown or new sequence as being in yet another species. I will sometimes refer to a sequence as "the sequence of a known species" —the "known species" (singular) being one of the species to which the known sequences belong.

I will address these questions in unrooted trees, so the conclusions will be helpful to experimentalists. Often an experimentalist will have a set of sequences, but not have easy access to a rooted topology for the species. He might not have a rooted topology because he does not have easy access to an outgroup sequence.

Weaknesses of the standard method

Table 1.6
Segments with low degeneracy are uncommon

The degeneracies were calculated based on one most-frequent aa at each site. "aa APPI" is average pairwise percent identity on the amino acid level.

Alignment	7 aa segment starting at aa site	aa APPI	Degeneracy
alpha globin	35	1	2048
alpha globin	84	1	2304
alpha globin	91	1	3072
alpha globin	133	1	9216
alpha globin	98	0.89	10,368
beta globin	21	1	18432
beta globin	91	1	1536
beta globin	28	0.94	4608
beta globin	133	0.90	12228
beta globin	56	0.89	1024
beta globin	77	0.89	3072

Requires segments with low degeneracy and other characteristics

One weakness of the standard primer design method is that it requires the presence, in the alignment of interest, of segments with high identity and low degeneracy. ("Degeneracy" here means the product of the number of codons in the genetic code, for the most-prevalent aa at each site in the segment.) If such segments are not present in the known alignment, or are not present at suitable locations within the alignment, then the researcher would want to have an alternative primer design method. (Currently, a researcher faced with this problem would do the best he could with the given sequences. Sells and Chernoff even write that the ideal part of an alignment is often difficult to find. Sells and Chernoff list a precise cutoff value for degeneracy, but not for percent identity.) This dissertation aims to demonstrate a primer design method suitable for use in segments that might not meet the Sells and Chernoff

criteria.

Sometimes the set of aa's in an aligned segment happen to be of too-high degeneracy to allow the construction of a pool of primers that contains every combination of degenerate codons at each site. For example, Table 1.6 shows the degeneracies and average pairwise percent identities for selected segments of the alpha globin and beta globin alignments. The table shows 11 seven-aa segments (out of the 40 total seven-aa segments of the two genes) with high identities. One can see from Figure 1.6 that only one of these eleven most-conserved segments of the two genes is of low enough degeneracy to fit the Sells and Chernoff criterion for degeneracy: the segment from aa sites 56 to 62 in the beta globin gene.

There are other criteria a researcher must worry about. One must find two appropriate sites an appropriate distance apart, for the upstream and downstream primers of the PCR reaction.

An alternate method for constructing a pool of primers, that would not constrain the researcher's choice of segment within alignment, as much as the standard method, would be useful.

Lack of success with standard method

A primer design method that has a higher success rate than the contemporary primer design method would also be welcomed by researchers. (By "success" I mean the designed primers amplify the gene of interest in the new species.) Primers designed via the Sells and Chernoff method do not always amplify the gene of interest. That fact is not surprising —no experimental technique works all the time.

It would be difficult to determine the success rate, in practice, of the Sells and Chernoff method. One would have to gather data on successful and unsuccessful screening attempts, many of which are not reported in the literature —especially the unsuccessful ones. But

based on anecdotes such as the ones described in the next section, it happens a significant fraction of the time.

Examples of not finding genes

Here I note a couple of examples of the standard primer design method not working, just to give a sense of how the method works in the lab.

The work of Basset et al. (2002) demonstrates that "Sells and Chernoff" designed primers do not always amplify the target gene. The two primer pools shown in Figures 1.3 and 1.4 (of this dissertation) were the primers that did successfully amplify. But Bassett and her coworkers report that they made several (at least 6) upstream and downstream primer pools designed to bind to different short stretches of the sequence. They tried them in PCR reactions in various combinations —thus empirically determining which amplified from the target DNA better, worse, or not at all.

Basset et al.'s work shows one way researchers deal with this lack of guaranteed success: by empirically trying multiple primers designed to bind to different stretches of the aa sequence. Sells and Chernoff (1995) suggest this approach. My goal in this dissertation is to invent an improved primer design method that will decrease the amount of this extra experimental effort that is necessary.

Another example of lack of success in finding related genes with PCR was related to me by Michael Purugganan. He was searching for MADS-box transcription factor genes in Clarkia species and mosses. He designed primers based on an alignment of approximately 20 genes, but was not able to amplify the desired gene from his species of interest. (Michael Purugganan, personal communication, June 2003.)

Kevin Moulton described instances of designing PCR primers, by essentially the Sells and

Chernoff method, in which the primers failed to amplify the desired product from fly species. (J. Kevin Moulton, 2002, personal communication.)

Reasons to think evolutionary model based prediction method might work better

In the standard method, each sequence in the known alignment is given the same weight in "predicting" the new sequence. That is, in most cases researchers do not take into account if one species in the alignment is relatively closely related to the "new" species of interest, (if they are looking in just one species) and give that sequence more weight than the other sequences. Nor do researchers usually consider if the set of known sequences is biased toward part of the evolutionary tree. For example, perhaps the gene of interest has been studied in several species of insects, and just one or two species each of fish, mammals, and shellfish. In this case a researcher might have an alignment of 15 known sequences of this gene, where half of these known sequences are from insects. If this researcher wants to screen for the gene in a species of mammal, it might make sense for him to give special consideration to the 2 known mammal sequences in the known alignment. From my discussions with researchers, this kind of special consideration is not usually done.

There are reasonable reasons why researchers following the standard method don't usually give extra weight to closely related species: following the Sells and Chernoff method, if one finds a region of 100% identity, relative distance does not matter. Also, if the researcher is interested in finding the gene in multiple new species, and wants to screen with one pool of primers, identifying the closely related known species might not make sense.

But a natural way to take into account biased representation of species is to use the phylogenetic tree. That is, if a lot of sequences in known alignment are close together in tree, that is taken into account, in a flexible and precise way, by making calculations based on mathematical models of evolution.

Improving one aspect of primer design —prediction

This dissertation does not attempt to offer improvements to all aspects and stages of primer design, of which there are many. Nor do I present a comprehensive algorithm for primer design. (I refer the reader to Bartl 1997, and Sells and Chernoff 1995 for such an algorithm.) I present algorithms intended to improve on one aspect of primer design: the prediction of the target sequence.

The following aspects of primer design are important to think about when designing a primer:

- The amplified product must be of an appropriate size. Bartl (1997) advises her readers to choose a distance between the upstream and downstream primers of 200 to 2000 bases. Shorter primers can be difficult to see on the post-PCR gel, and difficult to purify from it. Fragments longer that 2000 bp might not amplify, without special considerations about the duration of the extension step in the PCR reaction.

- If a PCR primer that is designed based on related sequences is used to amplify from the genomic DNA of the target species, unknown introns in the target species can cause problems two ways. First, an intron within the short stretch of nucleotides that a primer is designed to bind to would probably impede primer binding to target DNA. Second, an unknown intron between the two PCR primers could either inhibit successful amplification, or would make the amplified fragment to be of an unexpected size, perhaps causing the researcher to overlook it.

- Hairpin structures and primer-primer binding can impede amplification. Many programs will predict if such binding is likely to occur, given the primer sequences. (e.g. “MacDNASIS” from Hitachi, “PRIMER3” by Lincoln et al.) These programs are not useful though, if one is considering using a primer pool

that contains 1000 different varieties (particular sequences) of oligonucleotides. In this case there will likely be many varieties of the primer that form hairpins, and many that do not.

- If, in the target genome, there are many copies of a short DNA sequence that a primer will bind to, the primer might be competitively bound by those other sites, thus interfering with amplification from the site in the gene of interest (Donehower et al., 1990. p. 34). Mitsuhashi (1996a,b) presented a method for attempting to circumvent this problem during primer design. But his method works only for single primers, not pools of primers.

- Many researchers consider the 3' end of the primer to be especially important in priming, and therefore apply different requirements to these nucleotides than to other nucleotides in the primer. Löffert et al. (1997) urge their readers to avoid a run of 3 or more C or G residues, and to use a special buffer in the PCR reaction so as to diminish the problem of mis-pairing at the 3' end. Kevin Moulton tried to make the 3'-most aa in his primer region be a methionine or tryptophan, each of which is coded for in the genetic code by just one codon. (Kevin Moulton, 2002, personal communication.)

- Primers of different lengths and different sequences will have different optimum PCR reaction conditions. So primer design is interrelated with the design of particular PCR reaction conditions, such as: temperatures of the melting, annealing, and extension steps; duration of each step; buffer; salts; and proofreading or non-proofreading polymerase.

Focus of this dissertation

The methods presented in this dissertation are designed with the goal of improving, relative to the standard primer design method, the prediction of sequences present in related species, given a set of homologous sequences.

Methods are relevant to non-PCR screening

One can screen a DNA library for the presence of a particular sequence using either hybridization of a long nucleic acid probe, or PCR. (Ausubel et al. 2002. p. 6-1) Both methods make use of DNA's property of hybridization between complementary DNA strands, to chemically find the sequence of interest. The methods are similar in that one must have a predicted sequence (in the form of a molecule) to find the desired sequence.

The methods presented in this dissertation could, with some modification, be used to design long nucleic acid probes, in addition to PCR primers. But herein I will always describe the methods in terms of PCR primers.

Algorithms and programs for primer design

Over the years, researchers have developed a number of computer programs to assist with some aspect of designing PCR primers to screen for related genes. Although none of these programs directly solve the problem that is the focus of this dissertation —prediction of sequences in related genes— they solve similar or related problems and deserve attention.

Programs that assist in choosing a primer site

An early program is Montpetit et al.'s (1992) OLIGOSCAN. This program helped researchers choose a primer site by searching through a set of DNA sequences, for occurrences of partial or complete identity between the sequences and a proposed oligonucleotide sequence.

Dopazo and Sobrino (1993) published a program to help researchers find regions of identity

in a set of sequences. Their program suggests the best primer for binding to (and therefore detecting the presence of) a sequence common to groups of species. For example, given a set of hantavirus genomes, the program could find sequences common to all of the genomes. Such a sequence could be used as the target of a diagnostic PCR, to determine if a patient is infected with a hantavirus. This program could, alternatively, be used to find short regions of high homology in an alignment.

One set of programs that advanced the automated design of PCR primers able to bind to aligned sequences was Gibbs et al.'s (1998) GPRIME package. (This package of programs could also be used to help design probes in non-PCR hybridization-based screening methods.) The program scans a set of aligned DNA (not aa) sequences, identifying short stretches of low variability between the sequences in the alignment. This program could be used to quickly identify appropriate primer sites in a DNA alignment.

This program is flexible in that it works on both protein coding and non-protein-coding regions of a DNA sequence. The program apparently does not distinguish between these two types of regions though. So it always identifies variability on the DNA level. It can not take into account possible codon structure of a DNA sequence. Therefore a weakness of the program is that it can not identify a part of an alignment that codes for identical aa sequences with different nucleotide sequences.

Mitsuhashi (1996) presented a program that helps a researcher choose the best primer location within a DNA sequence. Here "best" means strongest binding. The program chooses the strongest-binding short oligonucleotide to bind within one given long DNA sequence. Binding strength is calculated based on thermodynamic stability of the hybrid composed of the oligonucleotide and the target DNA. This measure distinguishes Mitsuhashi's program from others that just count matches and mismatches between two sequences.

Mitsuhashi program takes into account that there are two classes of sequences the oligonucleotide might bind to: 1. the intended target sequence that is the length of the oligonucleotide and 2. other DNA sequences present as "background" in the PCR reaction. (The program allows the user to specify a database of background sequences.) The program calculates how strongly the oligo will bind to each of these two DNA sequences. It tells the user which oligonucleotides have a better ratio of high affinity for their target sequence and low affinity for background sequences.

The program also will inform the user of issues that might interfere with amplification in a PCR reaction: hairpin structures; primer-primer interactions; and GC content of sub-regions of an oligonucleotide, such as the 3' end.

Mitsuhashi's program could be used in some creative ways to decrease problems stemming from primer-primer binding within pools of PCR primers. But the program's intended purpose is to help design effective PCR primers given a single target DNA sequence, not an alignment of aa sequences.

A program with an alternative primer design algorithm

Rose et al. (1998) created the CODEHOP program. This program uses a clever strategy to address the same experimental problem this dissertation addresses: given an alignment of related protein sequences (that probably contains some short, highly conserved segments) design a primer, or pool of primers, that will anneal to another species' gene that is homologous (but whose exact sequence is unknown). CODEHOP requires a highly conserved segment 4 aa's in length. But CODEHOP designs a primer pool to anneal to an approximately 30 nucleotide sequence, i.e. a sequence long enough to code for 10 aa's. The reason for this difference is that CODEHOP uses two different schemes to design two different parts of each oligonucleotide in the pool: one scheme for the part corresponding to the highly conserved aa's, a different scheme for the rest of the oligo. The 3' end of each

oligo corresponds to the highly conserved aa's, the 5' end of each oligo is designed by the second scheme.

Consider the twelve 3'-most nucleotides of the oligos that constitute the pool. These sequences of these twelve nucleotides include all combinations of codons coding for the four highly conserved amino acids.

The sequence of the remaining 18 nucleotides, on the 5' ends of these oligos (corresponding to the remaining 6 amino acids) is a single (non-degenerate) sequence consisting of the most probable nucleotide predicted for each position. The CODEHOP algorithm sets these most-probable nucleotides to be simply the most common codon (according to a codon usage table specified by the user) coding for the most common aa at that position in the aa sequence.

(Sometimes Rose et al. would include all degenerate oligos coding for just three amino acids on the end, not four.)

Rose et al. choose this strategy because the 3' end is widely considered to be more important for primer success in PCR than other parts of the primer are (Löffert et al. 1997; Kevin Moulton, 2002, personal communication). A primer pool designed using Rose et al.'s algorithm will be certain to contain oligos that are an exact match for the 3' part of the target species' nucleotide sequence, assuming the amino acid sequence is completely conserved in the target species. And those "3' exact match" oligos will be present in a higher concentration than "3' exact match" oligos in a pool designed by the Sells and Chernoff method, because the 5' end is not degenerate in the Rose method.

Efficacy of the CODEHOP method

In their 1998 paper, Rose et al. did three separate tests in which they designed primers with their algorithm, and attempted to amplify genes from certain target DNA. In all three of these

tests they successfully amplified genes of the gene family of interest. And the CODEHOP method has been successfully used by researchers since then (see Figure 1.?? below).

Efficacy of CODEHOP relative to the standard method

Rose et al. wanted to demonstrate their strategy's efficacy relative to the standard method of designing primers for screening for related genes. To this end, their compare their results, from one of the three tests in their paper, with results reported by other researchers in two previous papers. The two previous papers reported the use of primers designed against the same short segment (within an alignment of reverse transcriptase genes) that Rose et al. design primers to.

Rose's work compares favorably with these previous reports. For example Wichman and Van Der Bussche (1992) use their primers to amplify reverse transcriptase from human genomic DNA, and found two unique sequences related to reverse transcriptase. But Rose et al. found 24 unique sequences related to reverse transcriptase, when they used the CODEHOP primers with human genomic DNA.

It is unfortunate, though, that Rose et al. did not do direct comparisons between the CODEHOP and standard primers. Direct comparisons would have allowed one to draw more definitive conclusions.

A direct comparison would have been to use primers designed by the standard method in PCR reactions that were otherwise identical to the PCR reactions Rose's primers were used in. Instead of that direct comparison, Rose et al. compared their results in their experiments with results reported in the previous papers. Rose et al. designed their primers to the same particular segments within the reverse transcriptase alignment, as the two previous papers. But besides that consistency between Rose et al.'s experiments, and the other two experiments, there were many differences: the given alignment of known sequences that the

primer designs were based on; how the target DNA was purified; how dilute the target DNA was in the reaction; the genome the target DNA was from; reaction conditions; the number of different pools of primers tested in the search for the one that worked best; how clones were selected from the finished reaction; and others. (Rose et al. 1998, Donehower et al. 1990, Wichman et al. 1992.)

Rose et al. claim that their CODEHOP primers were able to amplify related genes that were "too diverged from known sequences to be readily isolated by standard methods." But they never compared their CODEHOP primers to primers designed by the standard method, in experiments where other reaction conditions were held constant.

Another issue, that confuses exactly what is being compared between the Rose paper and the other two, is that in both of the papers that Rose uses to represent the standard method, the primers contained "extra", non-degenerate nucleotides on the 5' end, containing a restriction site for cloning. Those extra nucleotides make the primers that are supposed to represent the standard method similar, in one way, to CODEHOPE-designed primers.

How often researchers use these programs

Out of all these programs (algorithms) to assist in primer design, the most widely used is Rose et al.'s CODEHOP, as judged by the number of times the papers have been cited. Table 1.7 lists the number of times each paper announcing one of the algorithms above has been cited, between its publication date and June, 2003.

Table 1.8 shows the number of times two primer design methods papers were cited. This table shows that many researchers use the primer-design method Sells and Chernoff describe in their 1995 paper, without citing that paper. For example, all the papers listed in Table 1.1, above, use essentially the Sells and Chernoff method, yet none of them cite Sells and Chernoff. There are so few citations of this paper probably because many researchers

Table 1.7
Number of citations of primer-design software

Each paper listed here describes an algorithm, and announces a program implementing that algorithm, for designing PCR primers, for uses similar to or equivalent to the problem this dissertation addresses. The number of citations is for the time between the paper's publication date and June 2003, according to the ISI® "Science Citation Index Expanded."

Paper	Number of Citations
Montpetit et al. (1992)	10
Dopazo and Sobrino (1993)	11
Mitsuhashi (1996a)	5
Mitsuhashi (1996b)	11
Gibbs et al. (1998)	5
Rose et al. (1998)	85

Table 1.8
Number of citations of primer-design methods papers

The number of citations is for the time between the paper's publication date and June 2003, according to the ISI® "Science Citation Index Expanded."(The Bartl reference is a chapter in a book and so is not in the database.)

Paper	Number of Citations
Sells and Chernoff (1995)	1
Bartl (1997)	N.A.

consider the primer design strategy Cells and Chernoff describe to be "common knowledge". Sells and Chernoff did not invent the method; they described an already widely used method. So researchers probably do not feel the need to cite a reference when they use it.

Evolutionary methods

No research in the literature has attempted to use evolutionary models to predict sequences in related species, or to use those predictions to design PCR primers for screening.

Evolutionary model based prediction of new sequences —derivation of equations

The problem of designing primers (or other probes) to screen for related sequences can be stated as:

Given:
a set of n aligned, related, protein-coding sequences (either nucleotide, aa, or codon) of length k.

Find:
the one sequence, out of all possible sequences, that is "most likely", according to stated assumptions and a model, to be the sequence of a generalized new member of the set.

The section "Definition of the problem to be addressed", above, which first stated the problem as given-find, discusses a few aspects of the problem, including the meaning of the phrase "new member of the set".

The word "generalized", and other terminology

This version of the given-find statement is a little different from the statement in the section above, in that this statement uses the concept of a "generalized" new sequence (a "generalized" new member of a set of sequences.)

Relative to a set of known sequences, a "generalized new sequence" is an unknown sequence, whose relationship to the sequences in the given set is also unknown (unspecified). One is not specifying, for example, the evolutionary distance the new sequence is expected to be from any of the known sequences.

I will refer to the species containing the generalized new sequence as the "generalized new species." Considering a set of species, whose evolutionary relationships are known, a "generalized new species" is one whose evolutionary relationship to the given set is unknown (unspecified). One does not know how the species containing the sequence is related to the known species. One does not know which of the known species the new species is most closely related to.

From a given set of sequences one can infer a most likely phylogenetic tree, including branch lengths. I refer to this tree, which includes only the species in the known set, as the "known tree". Sometimes I refer to this tree as "the inferred tree", if there is little chance the reader will confuse the specified tree with a different inferred tree.

Figure 1.1 A. depicts a simple, hypothetical, known tree. If I learn a new sequence and its relation to the known sequences, I would show the relation to the known sequences by placing a new branch, terminating in a tip node, somewhere on the known tree. An example of such an attached branch is depicted in Figure 1.1 B. I refer to the point at which this new branch attaches to the known tree the "attachment point." For a generalized new sequence, the attachment point and length of new branch are unknown.

A generalized new sequence is "general" in that one is not specifying where the branch and tip that would represent the new sequence are located on the known tree.

Equations involving generalized new sequences

Equations in subsequent sections of this dissertation contain terms that represent generalized new sequences. These equations are statements about sequences that are unknown. The sequence might later be discovered, through experiments, but in the context of the equations containing "generalized new sequence" terms, the sequence is unknown.

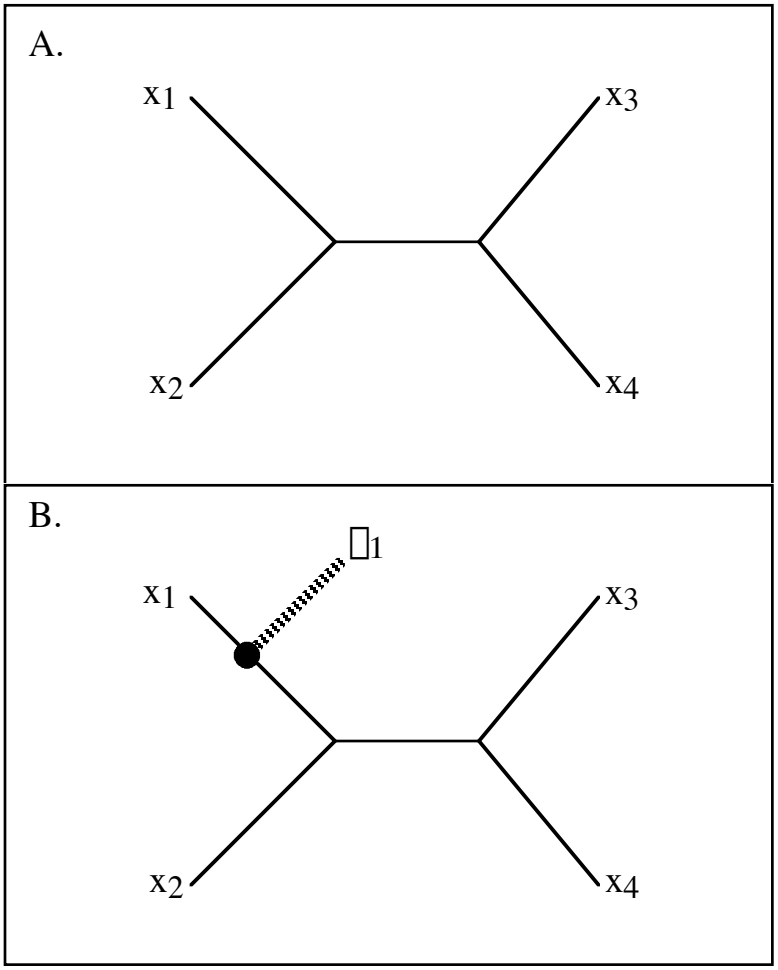

Figure 1.1. **Hypothetical trees.** A) A known phylogenetic tree. B) The tree with "new" sequence □l at the end of a new branch

In the context of these equations, the length and location of the attachment point that would connect the generalized new sequence's branch to the known tree is, as stated above, unknown. It is possible the appropriate attachment point might be any point on any branch of the known tree.

Because the attachment point is not known, it is not straightforward to calculate the

generalized species' probabilities associated with the evolutionary model, as one would calculate these probabilities for known nodes of a tree.

The solution to this difficulty is to model generalized species' probabilities as the average of what the probability would be if the attachment point were located in a finite set of positions on the tree (the finite set representing all possible attachment points on the tree.) An example, hypothetical, finite set of attachment points is depicted in Figure 1.2.

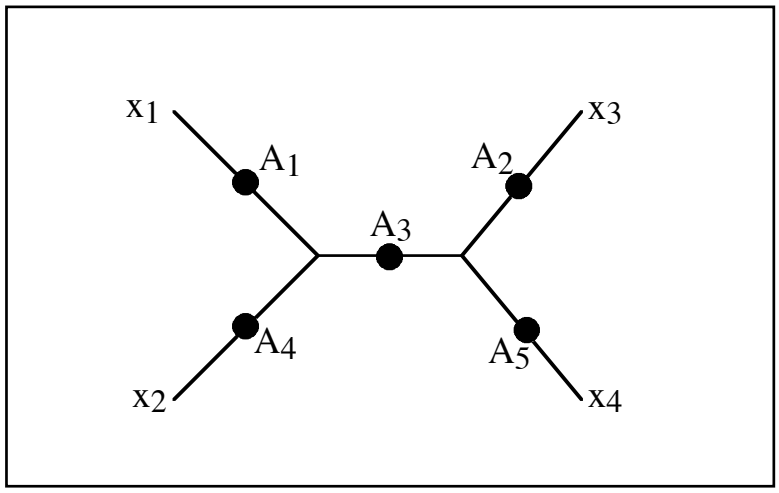

Figure 1.2. **Attachment points.** The finite set of attachment points are designated A_1 through A_5. The known sequences at the tips are indicated by x_1 to x_4.

The goal of the following derivation

To solve the given-find problem in the section above, I define a probability of a particular sequence being the generalized new sequence, given the data and the tree inferred from that data.

$$P(\square_{\text{gen}} | X, \square)$$

Here $\square_{\text{gen}}$ is the generalized new sequence. These equations can model sequences of

nucleotides, amino acids, or codons.

$$\Omega_{gen} = \{\text{all sequences of length k}\}$$
$$X = \text{the set of known sequences } x_1 \text{ to } x_n, \text{ all of length k.}$$
$$\theta = \text{the parameters of the inferred tree}$$

Parameters included in Θ are the topology, branch lengths, and a priori probabilities of states of a root node. Theta does not include information about what internal node of the tree is ancestral to any other internal node. The parameter Θ will be omitted from the equations, for the sake of notational simplicity.

Let the place where a new branch attaches to the known tree be attachment point A.

$$A = \text{the set of attachment points}$$

One could imagine, of course, a new branch attaching at any "point" on a branch of the known tree. But in this dissertation I use the term "attachment point" to refer to a finite set of points. There is one of these "attachment points" at the midpoint of each branch of the tree. Below, I calculate a weighted average over this discrete set of points (a graphical example of which is in Figure 1.2) instead of an average over all possible points on all branches of the tree.

The variable α (alpha) represents the state (the sequence) of the node that is attachment point A.

$$\alpha \in \{\text{sequences of length k}\}$$

The variable B designates a tip node at the end of the attached branch. The state at node B is β (beta), where β is a new sequence.

In the equation $P(\beta_{gen}|X,\theta)$, the purpose of the "gen" subscript is to remind the reader that the attachment point (where the new branch joins the known tree) is not specified. If the attachment point is specified, this term would be:

$$P(\beta_A|X,\theta,A)$$

This probability is: the probability of the particular sequence beta being the new sequence, given the known sequences X, the inferred tree τ, and that the new branch is attached at attachment point A.

The "A" subscript above is to remind the reader that the attachment point is specified.

(Both the "gen" subscript and the "A" subscript are superfluous, since the information they convey to the reader can be discerned from the list of "given" events, by noticing if the attachment point is present or absent in the list. I will sometimes leave off this subscript.)

The probability, $P(\beta_{gen}|X,\theta)$, is the number I want to calculate. To predict a most likely new sequence, one would calculate this probability for all possible sequences β_{gen} and then choose the β_{gen} for which the probability is the maximum. My goal in the derivation below is to arrive at an equation: an expression for this probability as a function of the sequence data and the parameter θ.

Once I have this expression, I will be able to compute this probability, given the sequence data I start with, and use those numbers to design primers and pools of primers.

An aside on notation: events vs. sequences

In equations such as $P(\theta|X,\tau)$, the variables θ, X, and τ are, strictly, "events". That is, they

are sets of outcomes of "experiments", in the probability theory sense. So I am using the notation inconsistently when I define σ as a member of a set of sequences, and at the same time write the probability that the new sequence is σ as $P(\sigma|X,\sigma)$. The two statements are inconsistent because I am on one hand saying

$$\sigma = \text{ sequence } i$$

and on the other hand saying

$$\sigma = \text{ "the outcome of experiment z is: sequence } i\text{"}$$

To be perfectly consistent I would define σ to be a sequence and write that probability as $P(\text{new} = \sigma|X,\sigma)$. But the language is easier in many cases if I am less strict. So I will often refer to σ as a "sequence" while in the same section using expressions such as $P(\sigma|X,\sigma)$ and not $P(\text{new} = \sigma|X,\sigma)$. The reader will always be able to tell what I mean based on context. If the context might not make the meaning completely clear to the reader, I will write expressions such as $P(\text{new} = \sigma|X)$ or $P(\text{root} = \sigma|X)$.

The use of variable σ is just an example of this dissonance between an event —the outcome of an experiment being a particular state— and the state itself. All of the variables in the probability expressions have these dissonant meanings.

Derivation of the probability of interest for site-by-site prediction method

Derivation of the probability of interest

Here is a statement, simply Bayes' rule, about a generalized new species.

$$P(\theta_{gen}|X,\pi) = \frac{P(\theta_{gen},X|\pi)}{P(X|\pi)}$$

Eq. 1.1

The parameter π does not affect the following derivation, so I will omit it from the equations in this section

$$P(\theta_{gen}|X) = \frac{P(\theta_{gen},X)}{P(X)}$$

Eq. 1.2

The denominator of equation 1.2 is the whole-tree probability. Let τ be the state (sequence) of the root node. Given the structure of the tree:

$$P(X) = \sum_{\text{all possible } \tau} P(X|\tau)P(\tau)$$

Eq 1.3

$\tau \in \{\text{sequences of length k}\}$.

As mentioned above, Γ does not include information about which node is the root node (root meaning ancestral to all other points on the tree). So for these calculations I assume the current attachment point, whichever it may be, is the root. I use this assumption because doing so allows the information from all the known sequences (the tips) to influence the state of the attachment point. This assumption is valid because the whole-tree probability is the same regardless of which node is assumed to be the root when calculating it.

Consider the probability $P(\square_{\text{gen}}, X)$, which is the numerator of the RHS of equation 1.2. This probability equals a weighted average, over all attachment points on the tree

$$P(\square_{\text{gen}}, X) = \sum_{\text{all A}} \left[\mathrm{P}(\square_{\mathrm{A}}, X|A) P(A) \right]$$

Eq. 1.4

where 'A' is an attachment point in the finite set of attachment points. And $\square_A$ is the state of the tip at the end of the attached branch that is attached to the tree at attachment point 'A'. For this version of the calculations I assume the attached branch is a constant length $\square$.

Modeling attachment points

I am modeling the generalized 'new' species (meaning a species that is adding to a set of known species) as the average of the set of hypothetical species at the tips of branches which join the known tree at a finite set of attachment points. Exactly how those attachment points are distributed on the known tree will influence the results I get from this model.

There are different distributions of attachment points one could use, depending on how one wanted to model a new species. For example, the attachment points could be distributed uniformly over the total length of all branches of the known tree. Another distribution would be one (or two, or m) attachment point(s) per branch of the known tree. Or I could distribute the finite set of attachment points over the length of all the branches, but weighted toward a particular tip node in the known tree. The reason for using this last type of distribution would be if, for some reason, I think the new species is closely related to one of the known species. Another possible "distribution" would be one single attachment point for the entire tree. A reason to use just one attachment point would be that I believed that was the one and only place the new branch could join the know tree, that belief based perhaps on a phylogeny

constructed from other genes.

The distribution I use is one attachment point per branch, and weight the probability associated with that attachment point proportionally to the length of the branch. This method models a uniform probability of where the new species appears on the known tree. So the probability $P(A|\square)$ in equation 1.4 is proportional to the length of the branch that attachment point A is on, relative to the branch lengths of all the branches. That is:

$$P(A) = \frac{\text{length of branch containing A}}{\sum_{\text{all branches i}} \text{length of branch i}}$$

Eq. 1.5

So the equation for our probability of interest is now

$$P(\square_{\text{gen}}|X) = \frac{\sum_{\text{all A}} \left[\mathrm{P}(\square_{\mathrm{A}}, X|A)P(A)\right]}{\sum_{\text{all possible }\square} \mathrm{P}(\mathrm{X}|\square)P(\square)}$$

Eq. 1.6

Rewrite $\mathrm{P}(\square_{\mathrm{A}}, X|A)$ as a sum over all the possible states of the attachment point. (I set the root, in the model, to be at the current attachment point. So summing over the states of the attachment point is written, in equation 1.7 and below, as summing over the states of the root. The two statements are equivalent.)

$$P(\square_{\text{gen}}|X) = \frac{\sum_{\text{all A}} \sum_{\text{all }\square} \left[\mathrm{P}(\square_{\mathrm{A}}, \text{root} = \square, X|A)\right]P(A)}{\sum_{\text{all possible }\square} \mathrm{P}(\mathrm{X}|\square)P(\square)}$$

Eq. 1.7

Where $\square \square$ {sequences of length k}

$$P(\square_{\text{gen}}|X) = \frac{\square \square \square \left[\mathrm{P}(\square_{\mathrm{A}} \mid \text{root} = \square, X, A)\mathrm{P}(\text{root} = \square, X|A)\right]P(A)\square}{\square \quad \mathrm{P}(\mathrm{X}|\square)P(\square)}$$

all A all $\square$

all possible $\square$

Eq. 1.8

The probability $\mathrm{P}(\square_{\mathrm{A}}|\text{root} = \square, X, A)$ is just the transition probability from the state // to the state $\square_A$, for a branch of distance $\lceil$. This transition probability is not affected by the structure of the tree, or by two of the given variables: the data X, and the particular attachment point A. So I can use the notation $t_{\square\square\ \square}$ to represent this probability. This transition probability is for the distance $\lceil$, even though $\lceil$ is not written.

Substituting in the transition probability notation, and rewriting the $\mathrm{P}(\text{root} = \square, X|A)$ term as $\mathrm{P}(X|\text{root} = \square, A)P(\text{root} = \square|A)$ yields

$$P(\square_{\text{gen}}|X) = \frac{\square \square \square \left[\mathrm{t}_{\square\square\ \square}\mathrm{P}(X|\square, A)\mathrm{P}(\square|A)\right]P(A)\square}{\square \quad \mathrm{P}(\mathrm{X}|\square)P(\square)}$$

all A all $\square$

all possible $\square$

Eq. 1.9

The term $\mathrm{P}(X|\square, A)$ is just the "likelihood" of the sequence //. In other words, it is the probability of the data, given that the sequence at attachment point (node) A is //, considering the attachment point A to be ancestral to all other points in the tree (and given the topology of the tree, $\lceil$.)

Calculating probabilities using this formula

The goal of the derivation above was to arrive at a formula that, if I have in hand some sequence data X, allows me to calculate a value for $P(\Box_{\text{gen}}|X)$. Equation 1.9 meets that criterion.

In the amino acid realm, I can calculate the RHS of equation 1.9, starting with amino acid alignment X, a phylogenetic tree inferred from the alignment (or known from other research on the represented species), and stationary probabilities of the amino acids known from Jones et al. (1992).

The instantaneous rate matrix for these amino acid models is calculated as follows. One starts with the entries in the PAM 1 matrix (except the matrix is transposed, so the transition is from the amino acid on row i to that on column j.) Each off-diagonal element is multiplied by $\Box_j$, the stationary probability for amino acid j (from Jones et al. 1992) corresponding to the row j the element occupies. The on-diagonal element for each row is now calculated to be the opposite of the sum of the off-diagonal elements for that row. Next one calculates the linear combination that is the sum of all the on-diagonal elements, each multiplied by the $\Box_j$ for the row it occupies. Every on-diagonal and off-diagonal element is divided by this linear combination. The resultant matrix is the instantaneous rate matrix.

If I choose to work in the codon realm, I can calculate the RHS of equation 1.9 using the codon alignment X and a phylogenetic tree inferred from that alignment (or known from other research on the represented species).

For evolutionary models in the codon realm, I follow the methods of Yang (1997). The instantaneous rate matrix Q for these codon models is built using the codon stationary probabilities $\Box$ ($\Box_j$ being the stationary probability for codon j) and Yang's parameters $\Box$

(kappa), the ratio of transition to transversion substitution rates, and ω (omega), the ratio of nonsynonymous to synonymous substitution rates (Yang, 1997). The entry q_{ij} in the instantaneous rate matrix Q equals the substitution rate from codon i to codon j. This instantaneous substitution rate equals: zero, if codons i and j differ at more than one nucleotide position; π_j if the change between i and j is a synonymous transversion; $\kappa\pi_j$ if the change between i and j is a synonymous transition; $\omega\pi_j$ if the change between i and j is a nonsynonymous transversion; and $\omega\kappa\pi_j$ if he change between i and j is a nonsynonymous transition (Yang, 1997).

Notation specifying length of sequences

The variable k, the length of the sequence in aa's, codons, or nucleotides, does not appear in the equations above. The length k influences the size of the sequence space. In the equations, if a summation is over a space of sequences, that space is indicated simply as "over all (sequences) X", or "over all (sequences) Γ". So the equations above could apply to a single site in an alignment, or to a segment, of any length, of an alignment. (Clearly too-long alignments, or segments of alignments, are computationally impractical and not important for primer design.)

It would be useful to have notation that specifies what length sequence an equation is being applied to.

In the equations below I will use write $P\left(\pi_{\text{gen}}\middle|X\right)$ as

$$P\left(\pi_{\text{gen,FL}}\middle|X_{\text{FL}}\right),$$

where the "FL" subscript stands for "full-length" indicating that probability calculated is for a full-length sequence, where "full-length" = k. "Full-length" refers to the full length of

a segment, usually 7 amino acids in length, not the entire full length of an alignment. (The alignments used in this study are above 100 amino acids in length.)

Computing statistics for sequences of length k

A way to make the calculation of $P(\Box_{\text{gen,FL}}|X_{\text{FL}})$, for a sequence of length k, computationally easier is to calculate $P(\Box_{\text{gen}}|X)$ for a single site (i.e. for an aligned "sequence" of length one) and then use the assumption of independent evolution between sites to calculate the full-length probability:

$$P(\Box_{\text{gen,FL}}|X_{\text{FL}}) = \prod_{S=1}^{k} P(\Box_{\text{gen},S}|X_S).$$

Eq. 1.10

In this equation S is the site.

In other words equation 1.9 becomes, for a sequence of length k,

$$P(\Box_{\text{gen,FL}}|X_{\text{FL}}) = \prod_{S=1}^{k} \left[\sum_{\text{all A}} \sum_{\text{all } \Box} \left[t_{\Box\Box} P(X_S|\Box_S, A) P(\Box_S|A) \right] P(A) \right] \Big/ \sum_{\text{all possible } \Box} P(X_S|\Box) P(\Box)$$

Eq. 1.11

Where the $\Box$'s in the denominator are "sequences" of length one.

Each term on the RHS is evaluated at each site in the sequence, the RHS ratio is calculated from those values, and then the RHS product is calculated by stepping through the sites of interest.

Equation 1.11 represents what I refer to throughout this dissertation as the "single site evolutionary model based prediction method", or more simply as the "single site method." In this name, "single site" refers to how the numerator and denominator on the RHS of the equation are both evaluated at a single site, S, and then the ratio and then the full length probability are calculated using those single site values.

A mathematical model of the standard primer design method

To compare evolutionary model based prediction methods against the standard prediction method, I need to express the standard method in a mathematical way. I need a mathematical equivalent of the standard method. This mathematical equivalent will assign a probability distribution to the sequence space. This distribution associates with each sequence a probability of being the particular "new" sequence. These probabilities sum to 1.0 over the entire space.

I will explain this distribution in terms of amino acids, though it could be over codons. So the distribution described here models the standard prediction method just up to the point of assigning a probability to each sequence. (Then choosing the sequence with the highest probability is trivial.) It does not describe the next parts of the standard primer design method: the step of making a pool of primers that includes all degenerate codons at each site.

The distribution I describe here assumes each site is independent. It does not take into account correlations between sites.

The multinomial distribution is a good, simple, place to start, in the construction of a mathematical equivalent of the standard primer design method. Using the multinomial as the prediction distribution would mean each amino acid would be assigned the probability equal

to its relative frequency in the observed alignment (at the site in question.) Consider for example an alignment of 10 aa sequences in which, at site 1, five sequences have a valine (V) and five sequences have a leucine (L). The simple multinomial prediction distribution would assign P(V)=0.5 and P(L)=0.5. And the probability of any of the other 18 aa's being the aa at site 1 in the "new" sequence would be zero.

This prediction distribution would be problematic to use as a representative of the standard prediction method because it assigns a zero probability to some aa's. A zero probability would make some statistical comparisons later in this dissertation unworkable. I want to modify the multinomial distribution so it never assigns a zero probability.

If I use a modified multinomial (described below), that never assigns a zero probability, does it still accurately represent what experimentalists do in the standard primer design method? It does have the same behavior as the standard method in that the amino acid chosen at each site (by the maximum probability criterion) will always be the same as the amino acid chosen by the experimentalist following the standard method. The modified multinomial is different from what experimentalists following the standard primer design method do, in that it simply assigns a specific probability to the different sequences. The standard primer design method does not make statements in terms of probabilities.

Notation for the modified multinomial

To make describing my modified multinomial distribution easier, I will introduce some notation here.

$x \in \{\text{amino acids}\}$

$f_S(x)$ = absolute frequency (the count) of x at site S

n = number of sequences aligned

$P_S(\text{new} = x)$ = the probability the new seq. is x at site S

The simple multinomial would assign probabilities at each site S by this formula

$$P_S(\text{new} = x) = \frac{\text{f}_S(x)}{n}$$

Eq. 1.12

The modified multinomial I will use reserves some of the probability in the prediction distribution for aa's not observed in the known sequences. (More precisely, it reserves some probability to be assigned to all 20 amino acids; but the important aspect of assigning this probability in this way it is that doing so gives some probability to unobserved aa's.)

The amount reserved is equal to the amount a new sequence added to the alignment would garner in the simple multinomial. For example, in an alignment of 10 sequences, 1/11 is set aside. In an alignment of 15 sequences, 1/16 is set aside. This amount of "set aside" probability is divided up uniformly among all amino acids (or codons, if doing all the prediction in the codon realm.)

So in the modified multinomial used in this dissertation, the formula for the probability assigned to an aa that is <u>not</u> observed in the known alignment is

$$P_S(\text{new} = x) = \frac{1}{20}\frac{1}{n+1}$$

Equation 1.13

And in this modified multinomial, the formula for the probability assigned to an aa that <u>is</u> observed in the known alignment is

$$P_S(\text{new} = x) = \frac{\text{f}_S(x)}{n}\frac{n}{n+1} + \frac{1}{20}\frac{1}{n+1}$$

Equation 1.14

The formulas in equations 1.13 and 1.14 are the same. The value of $f_s(x)$, for an aa not observed in the known sequences at site S, is zero.

This modified multinomial fulfills all the following desiderata of a function representing the standard prediction method:

- An aa observed more often in the sample is assigned a higher probability than one observed less often.
- A nonzero probability is assigned to every letter (aa or codon), even if it is not observed in sample.
- Probabilities assigned to aa's unobserved in the data are small, less that 0.01. (Although I make one more refinement of the modified multinomial, below, to make it fulfill this desideratum better.)
- As more sequences are added to an alignment, higher probabilities will be assigned to known aa's, and lower probabilities will be assigned to unobserved aa's.

To return to the example set out above: Consider for example an alignment of 10 aa sequences in which, at site 1, five sequences have a valine (V) and five sequences have a leucine (L).

Table 1.9
Comparison of multinomial and modified multinomial

	Simple multinomial	Modified multinomial
P(V) or P(L)	$\frac{5}{10} = 0.5$	$\frac{5}{10}\frac{10}{11} + \frac{1}{20}\frac{1}{11} = 0.459091$
P(an unobserved aa)	$\frac{0}{10} = 0.0$	$\frac{1}{20}\frac{1}{11} = 0.004545$

One can see in Table 1.9 that the probabilities are not too different between the modified multinomial and the simple multinomial. But importantly, the modified multinomial does not assign a zero probability to any aa.

Another example

For an alignment of 10 sequences, all of which are valine at site 2,

$$P_2(V) = \frac{10}{10}\frac{10}{11} + \frac{1}{20}\frac{1}{11} = 0.913636.$$

And for one of the other 19 (all unobserved) aa's x,

$$P_2(x) = \frac{1}{20}\frac{1}{11} = 0.004545.$$

If at site 3 in the alignment of 10 sequences, a different aa was observed in each of the 10 known sequences, then for y = an observed aa,

$$P_3(y) = \frac{1}{10}\frac{10}{11} + \frac{1}{20}\frac{1}{11} = 0.095455.$$

And the probability assigned to each of the 10 unobserved aa's z at this site would be

$$P_3(z) = \frac{1}{20}\frac{1}{11} = 0.004545.$$

A further refinement

Another desideratum for a distribution that is an analog of the standard primer design method: Keep the part of the probability assigned to unobserved aa's low. If there are not many sequences in the alignment, it tends to get high. For example, by Equation 1.14 above, for an

alignment of 5 sequences, with 18 unobserved aa's, the aggregate probability of the 18 unobserved aa's would be 0.15.

To meet this desideratum, I set an arbitrary limit of 0.10 as the upper limit of aggregate probability of seeing any (not any particular one, but any all together) of the unobserved aa in the new sequence. This limit applies no matter how few sequences are in the alignment. This modified form of the formula has this behavior:

$$P_S(\text{new} = x) = \frac{f_S(x)}{n}(1 - R) + \frac{1}{20}R \qquad \text{Eq. 1.15}$$

$$\text{for } R = \min\left[0.1,\ \frac{1}{n+1}\right]. \qquad \text{Eq. 1.16}$$

This is the modified multinomial function I use to represent the standard prediction method in this dissertation. This version of psuedocounts is similar to that presented in Sjölander et al. 1996.

Further justification of using a probability distribution to represent the standard primer design method

One implicitly assigns probabilities, one explicitly

The previous section proposed a probability distribution analog of the standard primer design method (the "standard method"). One might question if a distribution is a fair representation of the standard method, because the standard method does not formally assign probabilities to anything, while the distribution assigns a probability (the probability that that particular sequence is the new sequence) to every sequence in the sequence space.

I think it is fair to claim that an experimentalist following the standard primer design method is implicitly assigning probabilities to different sequences in the sequence space. I will use the Kirimura et al. (1999) study, cited above, to illustrate this point. In this study, Kirimura et al. used two sequences —one from a lily and one from a yeast— to design primers to find a gene in the fungus *Aspergillus niger*. Their purpose in using the outgroup —the lily— was to identify conserved regions in the alignment. Their logic was: if a sequence is conserved between a lily and a fungus, then it will probably be conserved within fungi. I think it is fair to say that Kirimura et al. are implying:

- The highly conserved sequences in the known alignment have a relatively high probability of being the "new" sequence in the target species, *Aspergillus*.
- Other sequences have a relatively low probability of being the "new" sequence (at those sites in the target species.)
- At sites not highly conserved in the known alignment, all sequences have too low a probability of being the new sequence (at those sites in the target species.)

The modified multinomial described in the section above gives explicit probabilities matching these implied probabilities.

Comparing prediction methods

The most important measure of a primer design method is its success rate —the rate at which primers designed via a particular method successfully amplify the gene of interest in the target species. So the most relevant comparison of different primer design methods would be an empirical test: given the same alignment, primers would be designed by the different methods, synthesized, and tested in the lab to see how well the different primers amplified the gene of interest. Whichever method succeeded more, on average, would be regarded as the "better"

method.

Probability of left-out sequence

That kind of lab experiment is beyond the scope of this dissertation. Instead of comparing success in empirical experiments, I will compare how well different primer design methods (prediction methods), given an alignment with one sequence left out, predict that left-out sequence.

So the general expression for this measure is

$$P(\text{LO\_seq}|\text{LI\_set}), \qquad \text{Eq 1.17}$$

where LO_seq means the left-out sequence, and LI_set means the left-in set of sequences. I will consider a prediction method that returns higher values of these probabilities to be a better prediction method.

The term in Equation 1.17 is exactly analogous to the term in the derivation above:

$$P(\square_{\text{gen}}|X) = P(\text{LO\_seq}|\text{LI\_set}) \qquad \text{Eq. 1.18}$$

Comparison needs to be on many instances

A goal of this dissertation is to be able to recommend, to a researcher wanting to isolate a homologous gene from one or more new species, which prediction method (primer design method) is more likely to result in success.

Obviously, it would not be convincing to claim that prediction method A is better than

prediction method B, based on a comparison of how the two methods performed on just one instance of the problem. A convincing comparison would be based on many different instances of the problem. Such a comparison would determine if one method works significantly better in a significant majority of the instances.

Description of a single instance

The "problem" this dissertation addresses is: given an alignment, find the most likely new sequence. As a foundation for a description (below) of the statistics on a number of these instances, here is a general description of a comparison of how two prediction methods perform on one instance of this problem:

> Given an alignment of 10 sequences of length 7 aa's, leave out one sequence. Calculate, once for method A and once for method B, P(LO_seq|LI_set) for this one left-out sequence, given the 9 left-in sequences. The result is two probabilities: one for method A and one for method B. The method that gives the higher probability is the better method.

Comparison of a number of these instances

Chapter Two describes doing that comparison using the alignments of four genes as data. After choosing the alignments, the process was: break each alignment into segments of 7 aa's, starting at the N-terminal end of the protein. Any excess aa's remaining on the C-terminal end were left out of the analysis. For each segment (7 aa's in length) of the alignment (n sequences in the alignment) step through the n sequences, leaving out each sequence in turn. For each combination of 7aa-segment-of-the-alignment and left-out sequence, calculate P(LO_seq|LI_set), based on the n-1 left-in sequences. Calculate the average of these probabilities over the n sets of n-1 left-in sequences.

For each combination of 7aa-segment-of-the-alignment and left-out sequence, the probability P(LO_seq|LI_set) is calculated for each prediction method. So at this point in the analysis

one has a set of $\left(\text{length of alignment in aa's}\right)/7$ average P(LO_seq|LI_set)'s for each prediction method. Next these sets of performance measures are compared with each other, to see which prediction method is better.

Comparing sets of these values
with the Wilcoxon-Nemenyi-McDonald-Thompson
two-sided all-treatments multiple comparison procedure

Certain statistical tests ruled out

Figure 1.3 shows that the average (over left-out sequence) P(LO_seq|LI_set)'s calculated from the four alignments, via the three prediction methods, are not normally distributed. Nor

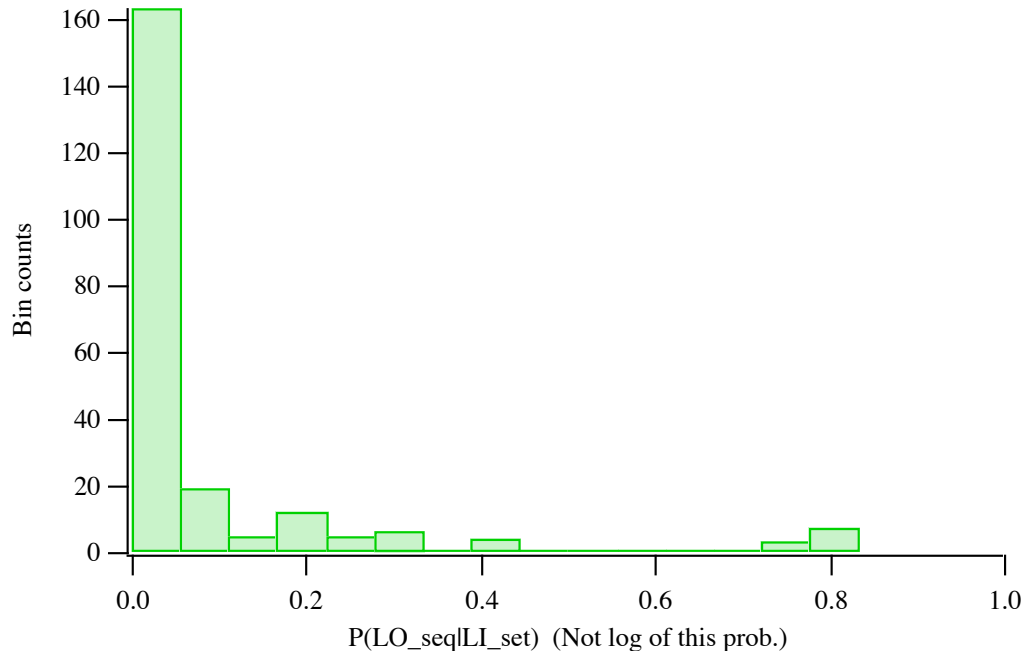

Figure 1.3. Histogram of probabilities. Histogram of the average (over left-out sequences) probabilities, for each 7aa segment, of each of the G3PD, alpha-beta globin, and rpL20 alignments. The probabilities are $P\left(\text{LO\_seq}|\text{LI\_set}\right)$

are the populations of differences between pairs of the average P(LO_seq|LI_set)'s (paired between methods, for the same 7-aa segments) normally distributed, as shown in Figure 1.4. So a t-test or paired t-test would not be valid ways to compare the groups of numbers. (Pollard, 1977. p. 177). Log transforming the probabilities was also considered and decided against, because doing so heavily weighted very small probabilities in an undesirable way.

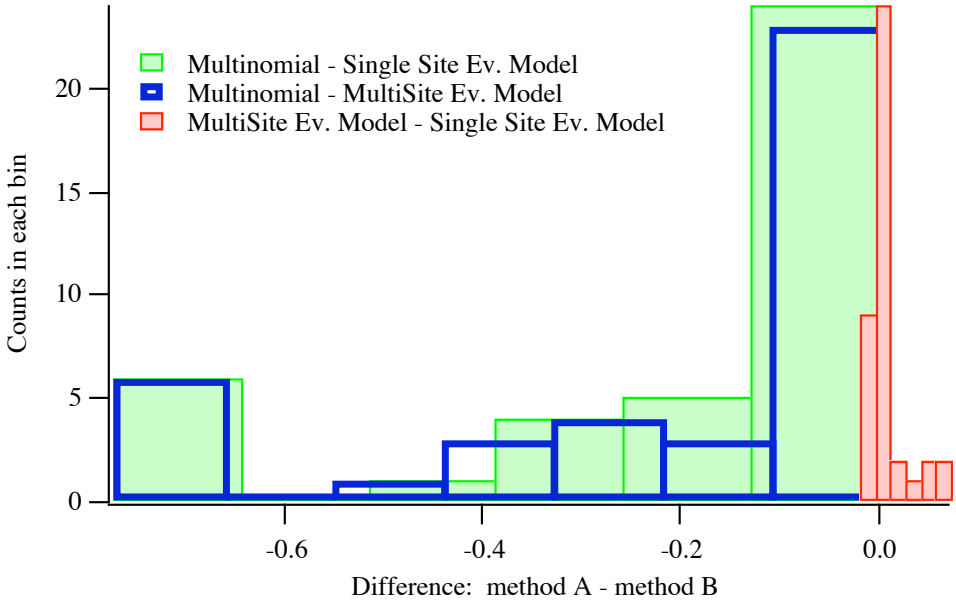

Figure 1.4. Differences of average $\mathbf{P(LO\_seq|LI\_set)}$'s are not normally distributed. The graph shows three distributions: the distributions of differences between the average (over left-in sets) of $P(\text{LO\_seq}|\text{LI\_set})$'s, for each pair of prediction methods. The data is the alpha and beta globin alignments.

The Wilcoxon-Nemenyi-McDonald-Thompson test

I decided to compare the sets of calculated values using the test that Hollander and Wolfe (1999) refer to as the "Wilcoxon-Nemenyi-McDonald-Thompson two-sided all-treatments multiple comparison procedure." This is a non-parametric test that allows one to keep track of the multiple-comparison critical value, for comparisons between each pair of the three

prediction methods studied in Chapter Two. The multiple comparison critical value I use in all of the tests in this dissertation is $\Pi = 0.05$. This Π is the probability of making one or more Type I errors in all the comparisons for an experiment.

The null and alternative hypotheses

The hypothesis tested by the Wilcoxon-Nemenyi-McDonald-Thompson two sided all treatments multiple comparison procedure refers to the treatment effects τ_j in this model of the data:

$$X_{i,j,t} = \mu + \beta_i + \tau_j + e_{i,j,t}$$

Eq. 1.19

In Equation 1.19, X is a data value, μ is the overall effect, β_i is the effect of block *i* ("block" here meaning the 7 aa segment in the alignment), τ_j is the effect of treatment *j*, ("treatment" here meaning the prediction method), and e is the error term.

The null and alternative hypotheses are:

H0: $\tau_1 = \tau_2$

H1: $\tau_1 \neq \tau_2$

Since I consider this performance measure, average $P(\text{LO\_seq}|\text{LI\_set})$, to be a measure of how well a prediction method predicts, the hypotheses being tested are, more relevantly:

H0: The two prediction methods predict "new" sequences equally well.

H1: One predicts better.

The sample of three alignments used in this dissertation

The three alignments I use in Chapters 2 and 3 were selected as follows. I used the alpha and beta globin sequences because they were included with PAML. I chose the G3PD alignment from the PANDIT database, because it was possible to select a subset of the G3PD alignment in the PANDIT database such that the subset's phylogenetic tree contained "clusters" —subsets of closely related sequences, separated from each other by larger distances. I chose the ribosomal protein L20 alignment from the PANDIT database because this alignment contains approximately 10-12 sequences. (I wanted an alignment containing a number of sequences intermediate between 5 and 22 sequences, the numbers of sequences in the globin and G3PD alignments, respectively.)

Other ways of comparing performance measures

One way I compare sets of performance measures (which are probabilities) from two prediction methods is a log scale histogram of differences. That is, I pair the performance measures from method A and method B, and subtract each method B value from the corresponding method A value. Then I plot these differences in a log scale histogram ranging from -1 to 1. Inspection of the graph reveals if one method performs better.

Interpreting log scale histograms of differences

In a regular histogram each bin covers an equal interval of the number line. In a log scale histogram, each bin covers an order of magnitude. For example one bin will cover the interval from 0.01 to 0.1. I refer to the numbers in this interval as being of order 10^{-2}.

One can think of this graph as showing one log histogram for positive values and one for negative values. The same-magnitude bins are in a mirror-orientation between the two halves

of the histogram. The bottom axis of the graph goes from -1 on the left side of the graph, to zero in the middle, to +1 on the right.

The two halves of the histogram show the distribution of differences between two probabilities. Thus these graphs give an indication of how lopsided, in one direction or the other, a set of differences is. For example a hypothetical set of differences in which 9 out of every 10 differences were positive is shown in Figure 1.5. This kind of result would indicate method B is a better predictor, according to the performance measure in use.

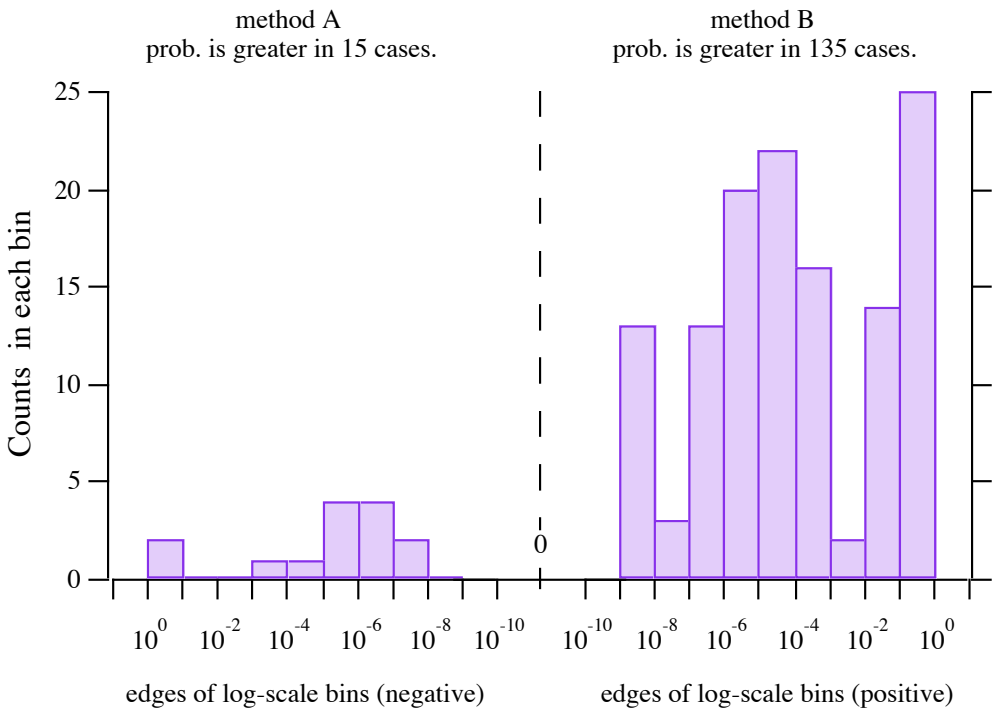

Figure 1.5. Log scale histogram of a set of differences. Made-up data demonstrating lopsided in positive direction.

Sign test

For a sign test to be valid, each individual item in the sample must be independent. (Pollard,

1977). If the performance measures in the sets studied with these histograms were independent of each other, one could perform a sign test on the counts of positive and negative numbers from these histograms, and use the result directly, as an indication of if the graph is truly lopsided. But the performance measures in the two sets (that this graph show the distribution of the differences between) are not independent from other performance measures in their set, because each performance measure is calculated based on a set of left-in sequences (each sequence in the alignment being left-out, in turn) and those left-in sets of sequences have a lot of overlap. So a sign test on the counts of positive and negative differences is biased toward rejecting H_0 that the median is zero.

The maximum possible amount of correlation between the performance measures in one of the sets would be if all n (where n = number of sequences in the alignment) performance measures for a segment were identical. If this were the case, then the appropriate sign test to do would be to divide the counts of positive and negative numbers each by n, and do the sign test using those quotients. But I know, from the Figures presented in the last section of Chapter 2, that there is less than complete correlation between the different performance measures for a segment. So a sign test on these quotients would be biased toward failing to reject H_0 that the median of the set of differences is zero.

Another possible correction of the counts, that would allow for an appropriate sign test, would be to multiply the counts of positive and negative differences each by the value $\frac{1}{\text{correlation coefficient}}$ for an appropriate correlation coefficient. This would probably be the most rigorous option, and future research into primer design methods would probably benefit from study into selecting the correct correlation coefficient for this purpose. But that study is beyond the scope of this dissertation.

The correction that I use, to allow a sign test other than the extremes noted above, is based on the observation that each of the n performance measures is calculated from a set of n-1 left-in

sequences. Analogous to the way one loses one degree of freedom when using a statistic calculated from a set of numbers, the correction is to multiply the counts of positive and negative differences each by the value $\frac{(n-1)}{n}$.

I perform sign tests both on the raw counts, and on the products arrived at my multiplying by the value $\frac{(n-1)}{n}$. With these two sign tests I hope to get a better idea of how much evidence there is, in the distribution of differences, that its median is not zero.

Tests in the amino acid realm are meaningful and interesting

Obviously, a PCR primer, or any other type of nucleic acid probe that uses hybridization to find a gene in a DNA library, needs to be designed in the nucleotide realm. But it makes sense to do some of the comparisons of different prediction methods (that are presented in this dissertation) in the amino acid realm.

First of all, it is valid to do the tests in the aa realm because the evolutionary relationships that the prediction is based on are present among the aa sequences in an alignment, just as they are present in codon or nucleotide sequences.

An advantage of amino acid sequences

One would expect that an evolutionary-model based algorithm for predicting related sequences would predict better on amino acid sequences than on protein-coding nucleotide sequences. Different species might have different codon preferences that would make

homologous nucleotide sequences more variable than the corresponding homologous amino acid sequences. Or homologous nucleotide sequences might have randomly walked (with silent mutations) in different directions while their corresponding amino acid sequences remained unchanged, resulting in greater variability between the homologous nucleotide sequences than between the corresponding amino acid sequences. This extra variability would make prediction more difficult. Jones et al. make a related point when they write that that "The bulk of the selection pressure is on the protein sequence and not on the underlying DNA sequence." (Jones et al. 1992. page 275)

Because a prediction method can be expected to work better in the aa realm, than the codon realm, one might want to test prediction algorithms in the aa realm. If one is trying to invent a method that works (predicts well), one might want to give the method every chance to work well, especially during the early stages of development when we are not even sure if the method works at all. A method that works well in the aa realm can then be studied more extensively in the codon realm.

Practical advantages of amino acid realm

Working in the aa realm also provides an advantage when one wants to use simulated sequence data, such as that used in Chapter 3. (By "simulated sequence data" I mean an aligned set of sequences, created by starting with one ancestral sequence and simulating the evolution of that sequence.) To generate the initial, random, ancestor sequence, one needs a distribution of states, either aa's or codons, from which to choose the sequence. When I generated simulated sequence data, I took this "a priori" distribution to be the stationary probabilities of the states at each site. But in the codon realm, the stationary probabilities are found as part of the process of fitting a most likely model, including most likely values for branchlengths and other parameters of the model. I do these calculations with Yang's PAML program, version 3.13. That is, to find codon stationary probabilities for a set of data, I give PAML the aligned codon sequences, the topology, and suggested values for branch lengths.

Using those inputs, PAML returns maximum likelihood fit of the stationary probabilities of all of the codons, branch lengths, and Yang's tree parameters kappa and omega (Yang 1997).

So the problem with generating simulated codon sequence data is that:
1) To generate aligned sequences via simulated evolution that starts with a random ancestor sequence, I must have stationary probabilities.
2) To get stationary probabilities, I must have aligned sequences from which to estimate them.

Working in the amino acid realm avoids this problem, because I use the stationary probabilities calculated by Jones 1992.

Another, minor, reason to test prediction methods in the amino acid realm is that the development runs run faster when doing the calculations on a 20 letter alphabet than on a 61 letter alphabet. And there are fewer problems with underflow error in the amino acid realm because the probabilities one is calculating are spread over a set of 20 instead of a set of 61.

AA realm is the appropriate place for comparing the predictive part of standard primer design method

In the amino acid realm, one can perform a direct comparison of the prediction part of standard primer design method, and another prediction algorithm.

One can think of the standard primer design method as working in two parts. In the first part, the researcher predicts the amino acid sequence that is likely to exist in the target species. (Or in other words, what amino acid sequence is likely to be the "new" sequence.) The second part of the primer design method is not prediction. In the second part, the researcher creates a primer pool containing all the combinations of codons that could code for the predicted amino acid sequence.

(Another way to describe the first part of the standard primer design method, prediction,

would be: one predicts that the "consensus" amino acid sequence of the given alignment will be present in the "new" sequence. Another way to describe the second, non-prediction part would be: one then creates a pool of all the degenerate primers, so as to cover all possible nucleotide codes for this amino acid sequence.)

In this dissertation I want to compare how well part one of the standard method, the prediction part, performs against other algorithms for predicting "new" sequences. These other prediction algorithms are the evolutionary model based prediction methods introduced in this dissertation. Because the standard method prediction is done exclusively in the amino acid realm, it makes sense to compare it with other prediction algorithms in the amino acid realm. That is, it makes sense to compare how well the different prediction methods predict related amino acid sequences, given an alignment of amino acid sequences, rather than compare how well they predict related nucleotide sequences, given an alignment of nucleotide sequences.

There is a difference between accurately predicting a "new" amino acid (or codon) sequence, and successfully amplifying a "new" gene. Once one has accurately predicted, one has to then design a pool of primers that has a good chance of succeeding in the chemical reaction that is the screening process. This dissertation is concerned with the prediction part of this process, rather than the chemistry part.

Codon realm vs. nucleotide realm

Parts of this study are done in the codon realm instead of the nucleotide realm because I feel the substitution rates between codons capture the evolutionary information in a sequence alignment at a finer "resolution", so to speak, than substitution rates between nucleotides do.

Chapter 2

Comparison of evolutionary model based prediction methods and standard method —relative performance at predicting a single related amino acid sequence

This chapter compares the relative performance at predicting a "new" sequence, in the amino acid realm, of the standard primer design method, represented by the multinomial distribution, and two different evolutionary model based prediction methods.

The prediction methods are compared two at a time. Both prediction methods in a comparison make predictions based on the same input data. Then how well each method predicted is quantified with a performance measure. Finally, the values of the performance measure for the two methods are compared. The method with the better performance measure is considered to be a better predictor.

The performance measure used in this test

This performance measure, like all the performance measures I use in this dissertation, is a statistic from a "leave one out" test. Each sequence in the alignment of n sequences is, in turn, left out, and each prediction method performs its calculations considering the remaining n-1 left-in sequences to be the known sequences. This statistic is in some cases used directly in the comparison, and in other cases this statistic is averaged over the n combinations of left-in and left-out sequences before being compared with the averaged statistics from the other methods.

Descriptions of the three alignments used in this study

I do all the comparisons in this study on three alignments. Here are descriptions of each.

Description of the G3PD alignment

In this study, I used a set of sequences of the C-terminal domain of Glyceraldehyde 3-phosphate dehydrogenase, abbreviated "G3PD." (Pfam accession number PF02800, Bateman et al., 2002.). These aligned sequences are among those made available by the Goldman Group (Whelan et al., 2002.) From the alignment of 76 G3PD nucleotide sequences the Goldman group makes available, I chose a subset of 22 sequences to use in my experiments. I chose these 22 because they group into 3 distinct clusters. By "distinct clusters" I mean the distances between the tips within each of the three clusters are significantly shorter than the distances between the clusters.

In Chapters 2 and 3 of this dissertation, I analyze these sequences as amino acid sequences. In Chapter 4, I analyze them as codon sequences. For analysis in the amino acid realm, I also used the stationary probabilities for each amino acid from Jones 1992.

Here is the Newick representation (Felsenstein, 1986; Olson, 1990) of the tree:

((((O68075:0.368154,(G3PC_ALCEU:0.131982,G3P_PSEAE:0.192865):0.048661):0.084150,(G3P2_RHOSH:0.183687,G3P_XANFL:0.113261):0.118153):0.035770,G3P_ZYMMO:0.385759):0.97648,(G3P_ARCFU:0.313573,((G3P_PYRWO:0.259089,(((G3P_METBR:0.153509,G3P_METTH:0.228401):0.112542,G3P_METFE:0.156903):0.109989,G3P_METJA:0.254988):0.066732):0.119374,G3P_SULSO:0.366910):0.080113):1.221310,((G3P_USTMA:0.116410,(G3P_PHARH:0.129724,(G3P_LYOSH:0.157512,G3P_PHACH:0.097192):0.044796):0.031102):0.026931,(((G3P_COCHE:0.059205,G3P_ERYGR:0.252898):0.041239,G3P1_TRIKO:0.276665):0.036585,(G3P_MONAN:0.132002,G3P2_TRIKO:0.085016):0.007360):0.033849):0.27891);

The phylogenetic tree for the amino acid sequences is shown in Figure 2.1. The three clusters within the tree are: 6 species of Archaea, 7 species of proteobacteria, and 9 species of fungi.

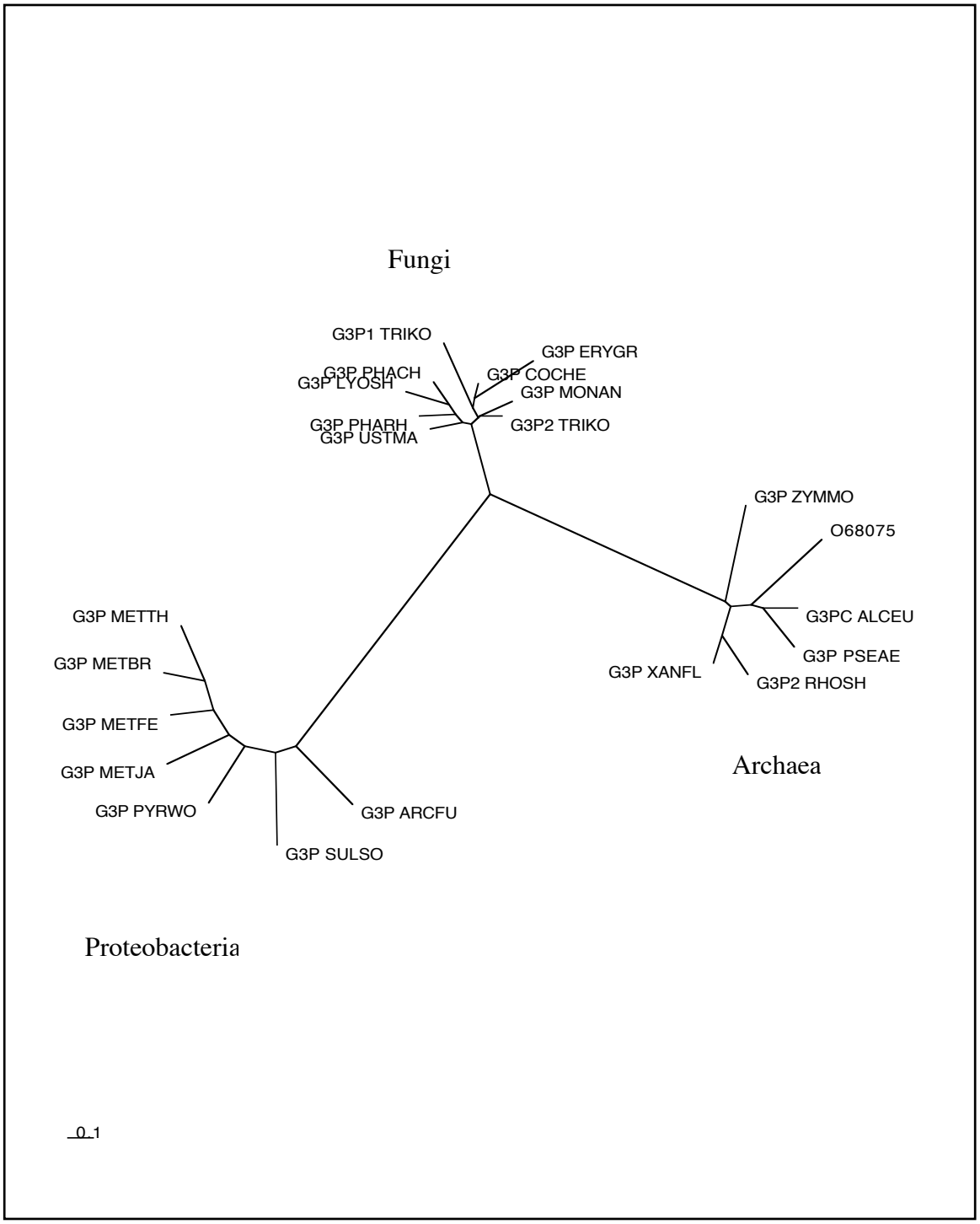

Figure 2.1. A subset of the Goldman group's G3PD phylogenetic tree. The branch lengths correspond to the Newick representation in the text and are in units of expected amino acid substitutions per amino acid site.

The G3PD sequences in the Goldman database are 549 nt (183 codons) long. But these sequences contain gaps. The prediction methods studied in this dissertation do not take into account gaps. (Modeling gaps would complicate the evolutionary model. Many evolutionary models, e.g. Yang et al. 1995, work on sequences with gaps removed.) So I eliminated gaps from the aligned G3PD sequences according to this rule: if any of the 22 sequences had a gap at a particular codon site ("codon site" meaning three nucleotide sites) I would excise that gap and the corresponding part of the sequence in all 21 other sequences. These excisions were done in units of codons, i.e. only multiples of 3 nucleotides were excised. This processing shortened the sequence so that all the sequences used in the experiment were 420 nucleotides, or 140 codons, long.

The branch lengths in the 22-sequence G3PD tree are the same as those in the G3PD tree connecting amino acid sequences given on the Goldman website. (http://www.ebi.ac.uk/goldman/index.html). This Goldman tree was inferred from the entire set of seventy-six 549-nucleotide sequences. The Goldman group derived the topology using neighbor joining, then found branchlengths using maximum likelihood techniques. The tree I use in the experiments presented in this dissertation is a subset of that seventy-six sequence tree —the subset containing only the 22 species, and the branches connecting the nodes of those 22 species.

(A researcher who wanted to use the predictive method I present in this dissertation could obtain the phylogenetic tree necessary to do the prediction in one of two ways. The researcher could infer the phylogenetic tree from the given known sequences. Or the researcher could use a tree reflecting the relationships he believes exist among the species of interest, based on other information.)

The average pairwise percent identity for the full-length 22 sequence alignment is 0.47 (with standard deviation 0.20).

Because of the low average identity between the sequences in this alignment, this alignment is a special challenge to the standard primer design method. The standard method calls for a segment with near 100% identity. There is no such segment in this alignment. If one leaves out, in turn, each sequence from each 7aa segment, and calculates the APPI (on the aa level) of the left-in sequences, the highest APPI over all the segments in the alignment is 0.7. If a researcher were faced with any of the 21-sequence leave-one-out alignments, he would have to work with it as best as possible in order to use the standard method on it.

Description of globin sequence data

I tested the prediction algorithms on segments of alignments of alpha globin and beta globin sequences from 5 species. The alpha globin sequences are 141 amino acids in length. The beta globin sequences are 144 aa's. The alignments, which contains no gaps, are a test alignment included with the PAML program. (Yang 1997.)

In Chapters 2 and 3 of this dissertation, I analyze these sequences as amino acid sequences. In Chapter 4, I analyze them as codon sequences.

The following analysis was done using the topology of the phylogenetic tree inferred by Yang. The branch lengths were inferred from a maximum likelihood analysis (done with PAML) of the concatenated alpha globin and beta globin aa sequences. This inferred tree is shown in Figure 2.2.

Here is the Newick representation of the amino acid tree: (((rabbit:0.07338, rat:0.17121):0.03361, human:0.05841):0.02918, goat_cow:0.09050, marsupial:0.23930).

The full-length average pairwise percent identities between amino acids, are 0.82 for the alpha globin alignment, and 0.81 for the beta globin alignment.

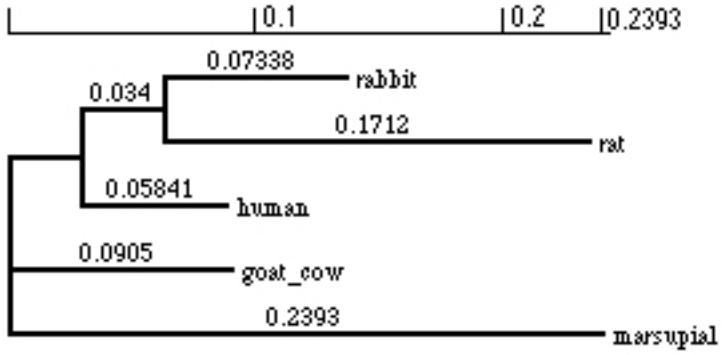

Figure 2.2. Phylogenetic tree showing relation of amino acid sequences in concatenated alpha globin and beth globin alignment. (The length of the alignment is 285 amino acids.) Branch lengths are in expected amino acid substitutions per amino acid site.

Description of the ribosomal protein L20 alignment

Tests were also performed on an alignment of 11 sequences of Ribosomal Protein L20 (abbreviated here "r.p. L20"). This alignment is a slightly modified version of an alignment (accession number PF00453) in the Goldman Groups' PANDIT database. The Goldman group alignment contained 333 nucleotides. I deleted from that alignment any site with a gap in any of the 11 aligned sequences. And when I deleted a gap, I also deleted the corresponding parts of the other 10 sequences in the alignment. These excisions were done in units of codons, i.e. only multiples of 3 nucleotides were excised. These deletions left an alignment of length 324 nucleotides or 108 codons. The Goldman group alignment also contained a twelfth sequence, a gene from *Mycoplasma* named RL20_MYCFE. I deleted that sequence from the alignment because that gene apparently uses a nonstandard genetic code, in which TGA codes for tryptophan instead of being a stop codon.

In Chapters 2 and 3 of this dissertation, I analyze these sequences as amino acid sequences.

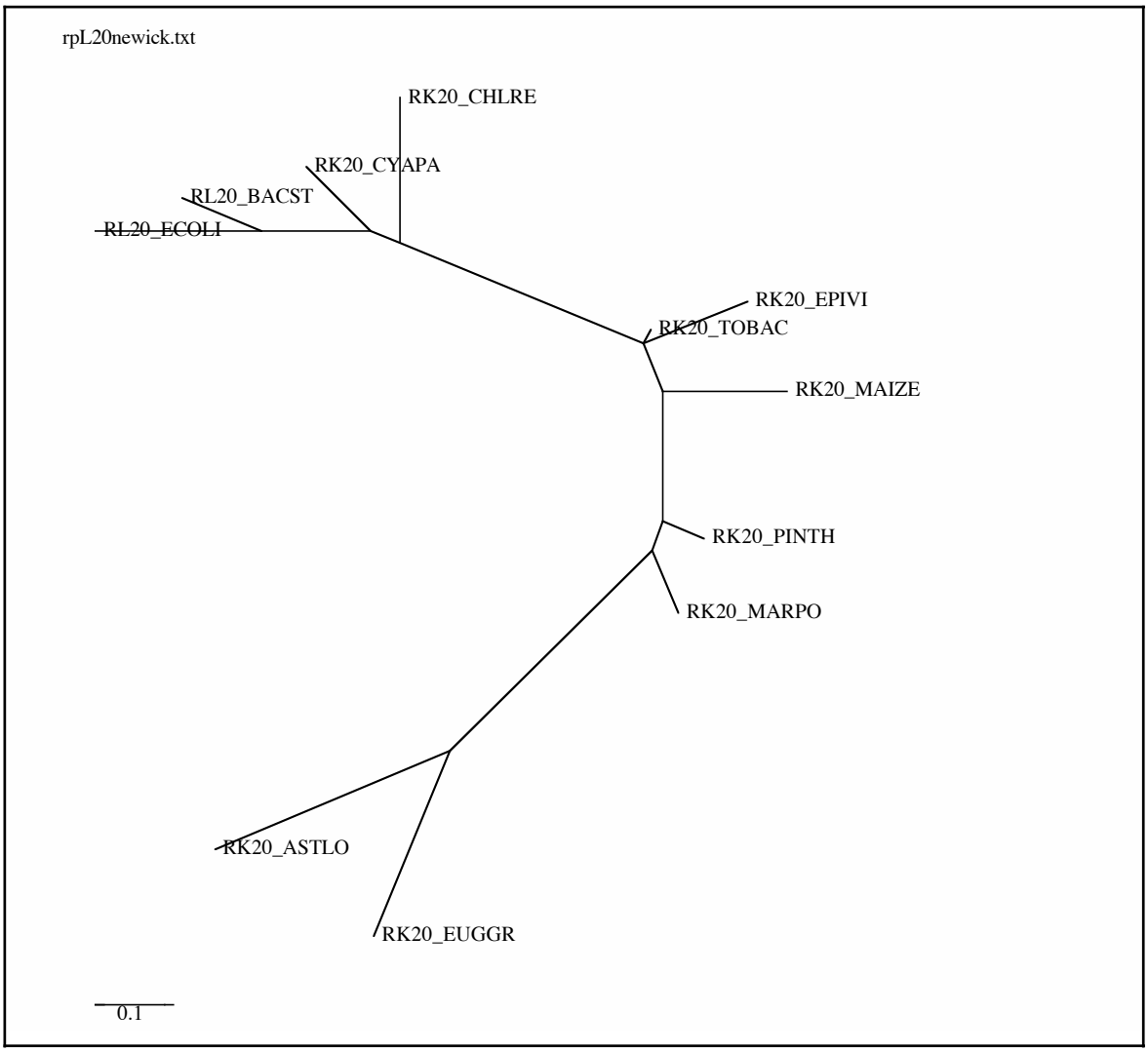

Figure 2.3. The phylogenetic tree of the 11 sequence ribosomal protein L20 alignment used in the analysis.

In Chapter 4, I analyze them as codon sequences.

I used the topology and branch lengths of the phylogenetic tree for amino acid sequences provided by the Goldman group. This tree was inferred from the full 111 aa alignment of the 12 genes, not from the slightly modified alignment I used. Below is the Newick representation of the phylogenetic tree, with the *Mycoplasma* node and its adjoining branch

removed. Branch lengths are in units of expected aa substitutions per aa site. The corresponding tree is shown in Figure 2.3.

(RL20_BACST:0.18603,(RK20_CYAPA:0.20106,(RK20_CHLRE:0.31732,(RK20_TOB
AC:0.03414,(RK20_EPIVI:0.24611,(RK20_MAIZE:0.27080,(RK20_PINTH:0.09825,(R
K20_MARPO:0.14822,(RK20_EUGGR:0.43743,RK20_ASTLO:0.55672):0.62418):0.0
6742):0.28579):0.11169):0.00000):0.57989):0.06601):0.24160,RL20_ECOLI:0.36183);

Where to find detailed descriptions of each of the three prediction methods studied in Chapter 2

In the next section, I refer to "three prediction methods". The first of these is the standard primer design method, represented by the (modified) multinomial distribution described in Chapter 1. The second prediction method is the "single site evolutionary model based prediction method", also described in detail in Chapter 1. The third prediction method is the "multisite evolutionary model based prediction method", which is described in detail in a section of Chapter 2, below. This third method is described in detail before the comparisons between it and the other two methods are discussed.

Breaking the alignments into segments, and results of the statistical test

To use the three alignments (G3PD, globin, and ribosomal protein L20) to compare the three prediction methods, I broke the alignments into segments of 7 amino acids.

The alignment of G3PD genes is 140 aa's long. I broke this length into 20 segments of 7 aa's, and did the analysis described below on each 7 aa segment separately. For each of these segments, in turn, I calculated a probability of the left-out sequence for each of the 22 sets of 21 left-in sequences. I calculated one of these probabilities for each of the three prediction methods. So these calculations yielded a data set of $20 \cdot 22 = 440$ probabilities (values of the

performance measure) for each of the three methods. The 22 probabilities for each segment were averaged, yielding 3 sets of 20 average probabilities.

These 3 sets of 20 average probabilities were compared using the Wilcoxon-Nemenyi-McDonald-Thompson multiple comparison procedure. The results of this comparison procedure for the G3PD alignment are shown in Table 2.1. These results will be discussed in detail below, not in this section.

Table 2.1
Results of Wilcoxon-Nemenyi-McDonald-Thompson tests on the G3PD alignment data

"Rank total" is the sum (over segments) of the within-segment ranks (1, 2, or 3) for that method.. $\square_1$ and $\square_2$ in H0 refer to equation 1.19. Overall $\square$ for the 3 tests = 0.05.

Prediction method	**Rank total**
Multinomial (representing standard method)	48
Single site evolutionary model method	52
Multisite evolutionary model method	20
Methods compared	**Result of test. Reject H0: $\square_1 = \square_2$?**
Multinomial & Single site ev. model method	Fail to reject
Multinomial & Multisite ev. model method	Reject
Single site ev. model method and multisite e model method	Reject

The alpha globin gene in 141 amino acids long. So it was divided into 20 segments of 7 aa's each, leaving the one C-terminal amino acid site out of the analysis. The beta globin sequences are 144 amino acids long. The beta globin gene was also divided into 20 segments of 7 aa's each, therefore excluding the 4 C-terminal amino acid sites from the analysis. So

there are a total of 40 segments of 7 aa's each from the globin alignments. Using these 40 segments, I did the same analysis as described above for the G3PD gene: calculated the P(LO_seq|LI_set) for each combination of segment and left-out sequence, for each of the three prediction methods. These calculations produced 3 sets of $40 \cdot 5 = 200$ probabilities. The 5 probabilities for each segment were then averaged, yielding 3 sets of 40 average probabilities.

These 3 sets of 40 average probabilities were compared using the Wilcoxon-Nemenyi-McDonald-Thompson multiple comparison procedure. The results of this comparison procedure for the alpha globin and beta globin alignment are shown in Table 2.2. These results will be discussed in detail below, not in this section.

Table 2.2
Results of Wilcoxon-Nemenyi-McDonald-Thompson tests on the alpha globin and beta globin alignment data

"Rank total" is the sum (over segments) of the within-segment ranks (1, 2, or 3) for that method.. $\square_1$ and $\square_2$ in H0 refer to equation 1.19. Overall $\square$ for the 3 tests = 0.05.

Prediction method	**Rank total**
Multinomial (representing standard method)	113
Single site evolutionary model method	65
Multisite evolutionary model method	62
Methods compared	**Result of test. Reject H0: $\square_1 = \square_2$?**
Multinomial & Single site ev. model method	Reject
Multinomial & Multisite ev. model method	Reject
Single site ev. model method and multisite e model method	Fail to reject

The ribosomal protein L20 sequences are 108 amino acids long. This alignment was divided

into 15 segments, each of length 7 aa's. The remaining 3 aa sites on the C-terminal end of the alignment were left out of the analysis. I calculated the P(LO_seq|LI_set)'s for each prediction method, leaving out each sequence of each segment in turn, as described above for the the G3PD and globin alignments. These calculations produced 3 sets of $15 \cdot 11 = 165$ probabilities. The 11 probabilities for each segment were averaged, yielding 3 sets of 15 average probabilities.

These 3 sets of 15 probabilities were compared using the Wilcoxon-Nemenyi-McDonald-Thompson multiple comparison procedure. The results of this comparison procedure for the ribosomal protein L20 alignment are shown in Table 2.3. These results will be discussed in detail below, not in this section.

Table 2.3
Results of Wilcoxon-Nemenyi-McDonald-Thompson tests on the ribosomal protein L20 alignment data

"Rank total" is the sum (over segments) of the within-segment ranks (1, 2, or 3) for that method.. $\square_1$ and $\square_2$ in H0 refer to equation 1.19. Overall $\square$ for the 3 tests = 0.05.

Prediction method	**Rank total**
Multinomial (representing standard method)	40
Single site evolutionary model method	29
Multisite evolutionary model method	21

Methods compared	**Result of test. Reject H0: $\square_1 = \square_2$?**
Multinomial & Single site ev. model method	Fail to reject
Multinomial & Multisite ev. model method	Reject
Single site ev. model method and multisite e model method	Fail to reject

Comparison of
the standard primer design method
and the single site evolutionary model based method

This is the initial comparison between the standard primer design method, represented by the multinomial distribution, and an evolutionary model based prediction method.

Quick statement of the single-site evolutionary model prediction method equation and the multinomial equation

From Chapter 1, here is the equation for the probability that a particular sequence $\Box_{gen}$ is the new sequence —that it is the next sequence added to the set of known sequences. The prediction method I call the "single site evolutionary model prediction method" predicts one sequence to be the next related sequence added to the set. The sequence it predicts is found using this equation. At each site (a site is length one aa or one codon) in the sequence, the $P(\Box_{gen}|X,\Box)$ is calculated for each possible "value" or state of $\Box_{gen}$. (The possible states are either the 20 aa's or the 61 codons.) The state with the maximum $P(\Box_{gen}|X,\Box)$ at a site is set as the predicted state at that site. Thus the predicted sequence is constructed.

$$P(\Box_{gen,FL}|X_{FL}) = \prod_{S=1}^{k} \frac{\sum_{all\ A} \sum_{all\ \Box} \left[t_{\Box\Box} P(X_S|\Box_S, A) P(\Box_S|A) \right] P(A)}{\sum_{all\ possible\ \Box} P(X_S|\Box) P(\Box)}$$

Eq. 1.11

The modified multinomial was calculated with equation 1.15,

$$P_S(\text{new} = x) = \frac{f_S(x)}{n}(1 - R) + \frac{1}{20}R$$

Eq. 1.15

for $R = \min\left(0.1, \frac{1}{n+1}\right)$. Since this alignment contains five sequences, R can be calculated and the formula becomes

$$P_S(\text{new} = x) = \frac{f_S(x)}{n}(0.9) + \frac{1}{20}(0.1)$$

Distributions of differences between performance measures of standard method and single site evolutionary model method

A way to compare the two prediction methods is to look at the log histogram of differences between the probabilities returned by the two methods. The log histogram showing these differences for the G3PD alignment is shown in Figure 2.4.

(See Chapter 1 for a discussion of interpretation of the log-scale histograms.)

Figure 2.4 indicates the multinomial prediction method is a little better than the single-site evolutionary model based method, because the performance measure of the multinomial is greater more often. The multinomial is greater 256 of 440 cases. The single site method is greater in 184 of 440 cases.

A two-tailed sign test for the numbers 184 and 256 returns a p value of 0.000713486. To do the adjustment for correlation (as explained in Chapter 1) I multiply the counts by $\frac{n-1}{n} = \frac{21}{22}$, giving 176 and 244 (rounding off). The p value of the sign test for 176 and 244

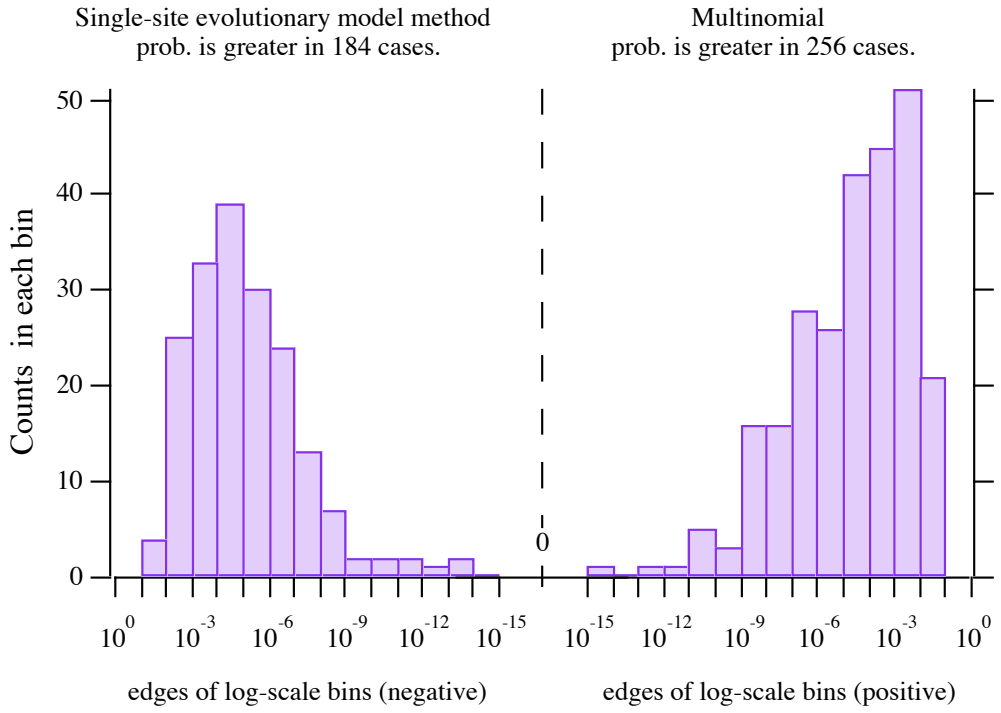

Figure 2.4. Comparison of standard prediction method and evolutionary single-site prediction method for G3PD alignment. Log-scale histogram of differences in probabilities. The differences graphed here are P(LO_seq|LI_set) for the Multinomial prediction method minus P(LO_seq|LI_set) for the Evolutionary model single site method. Notice these are not log(probabilities). The vertical axis is bin counts. The horizontal axis, running from -1 on the left to 1 on the right, indicates the order of magnitude of the edges dividing the bins in the histogram. All 440 differences are sorted into the bins of this histogram —one difference for each of the 22 left-out sequences, for each of the 20 7aa segments of the G3PD alignment.

is 0.00107992.

The distribution of differences for the globin data

Figure 2.5 is a log-scale histogram showing the distribution of differences between paired (paired for each combination of segment and left-out sequence) performance measures for the multinomial and single site evolutionary model methods, for the globin data.

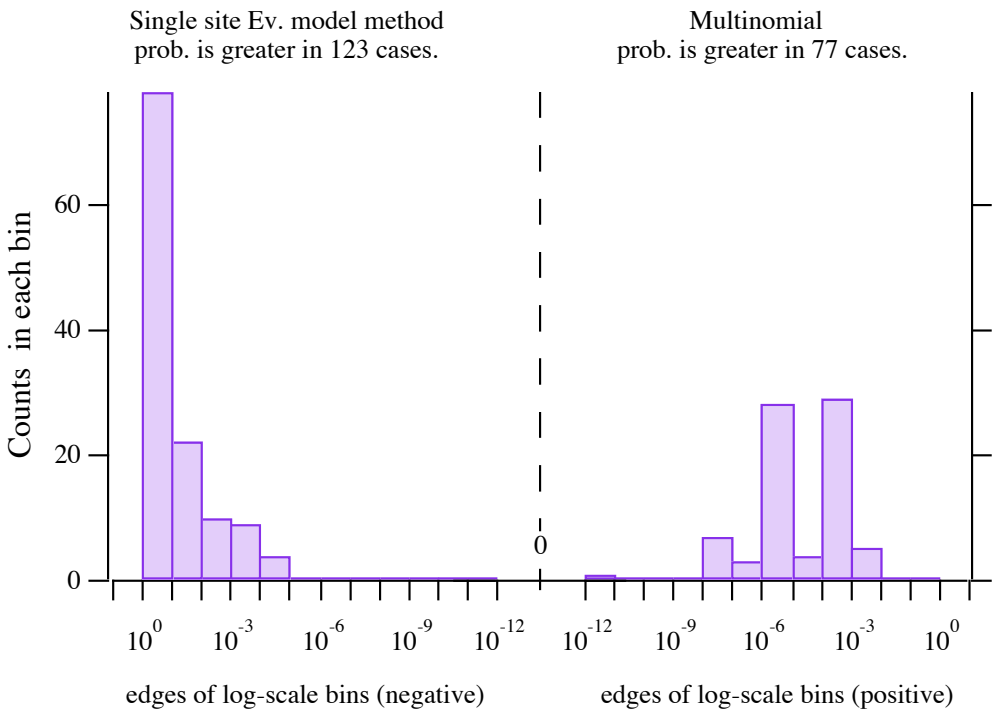

Figure 2.5. Comparison of performance measures for the two prediction methods, on the alpha globin and beta globin alignments. Log-scale histogram of differences in probabilities. The differences graphed here are P(LO_seq|LI_set) for the multinomial prediction method minus P(LO_seq|LI_set) for the single site evolutionary model method. Notice these are not log(probabilities). The horizontal axis indicates the order of magnitude the edges of the bins. All 200 differences are sorted into the bins of these histograms —one difference for each of the 5 left-out sequences, for each of the 40 7 aa segments of the globi alignments.

Figure 2.5 indicates that the single-site prediction method predicts better than the multinomial does on this globin data. Consider the two bars between ⌈10^{0} and ⌈10^{-2} (on the negative, left, side of Figure 2.5). Exactly 100 of the 200 cases are contained in those two bars (77 of magnitude 10^{-1} and 23 of magnitude 10^{-2}). Therefore, in half the cases the evolutionary model method scores not just a higher performance measure, but a significantly higher performance measure. ("Significant" here simply meaning above the threshold of 0.01.) In

zero out of the 200 cases is the multinomial performance measure greater than the evolutionary model performance measure greater by this amount.

The two-tailed sign test for 123 and 77 has a p value of 0.00146486. . To do the adjustment for correlation (as explained in Chapter 1) I multiply the counts by $\frac{n-1}{n} = \frac{4}{5}$. giving 98 and 62 (rounding off). The two-tailed sign test for 98 and 62 is 0.00566406.

Comparison using ribosomal protein L20 data

Figure 2.6 is a log-scale histogram showing the distribution of differences between paired (paired for each combination of segment and left-out sequence) performance measures for the multinomial and single site evolutionary model methods, for the ribosomal protein L20 data.

Figure 2.6 indicates that the evolutionary model method is a better predictor. In Figure 2.6, compare the bins holding positive differences (which are those in which the multinomial performance measure is greater) of order of magnitude 10^{-1} and 10^{-2} with the bins holding negative differences (which are those in which the single site evolutionary model performance measure is greater) of those magnitudes. In 19 of the 165 cases, the single site evolutionary model method performance measure is greater than that of the multinomial method, by a magnitude of 10^{-1} or 10^{-2}. (Here "magnitude 10^{-1}" means a value from 0.1 to 1.0, and "magnitude 10^{-2} means a value from 0.01 to 0.1.) While in zero of 165 cases is the multinomial method performance measure greater than the evolutionary model method performance measure by this magnitude.

The two-sided sign test for outcomes 86 and 79 has a p value of 0.640474. . To do the

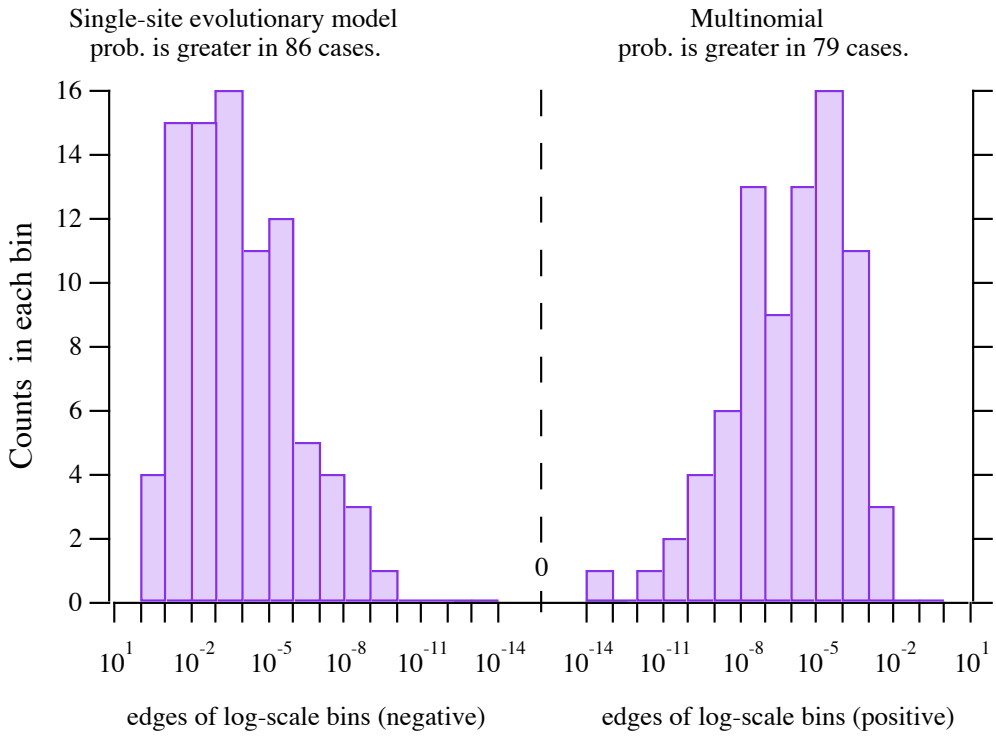

Figure 2.6. Log-scale histogram of differences in probabilities. The differences graphed here are P(LO_seqlLI_set) for the multinomial prediction method minus P(LO_seqlLI_set) for the single site evolutionary model method. Notice these are not log(probabilities). The horizontal axis indicates the order of magnitude that each bin in th histogram covers. All 165 differences are sorted into the bins of this histograms —one difference for each of the 11 left-out sequences, in each of the fifteen 7-aa segments of the ribosomal protein L20 alignment.

adjustment for correlation (as explained in Chapter 1) I multiply the counts by $\frac{n-1}{n} = \frac{10}{11}$

giving 78 and 72. That sign test has a p value of 0.683131.

Summary of comparing differences between performance measures between standard method (multinomial) and single site evolutionary model method

Table 2.4 summarizes the results of this section.

Table 2.4 Summary of comparisons of which method has greater performance measures					
"n.s.d." means "no significant difference", i.e. the sign test failed to reject H0: same number of cases favoring each method. n = num of sequences in alignment.					
2 Methods compared	**Gene**	**Which method has higher performance measures, based on each of these ways to compare differences . . .** count of differences > 0	sign test of count of differences >0	sign test of counts * (n-1)/n	count of differences > 0.01
Standard method (multinomial) vs. Single site evolutionary model method	G3PD	standard method (multinomial)	standard method (multinomial)	standard method (multinomial)	standard method (multinomial)
	globin	single site	single site	single site	single site
	r. p. L20	single site	n.s.d.	n.s.d.	single site

Comparison of average (over left-in sets for a segment) performance measures of the standard method and single site evolutionary model method

The Wilcoxon-Nemenyi-McDonald-Thompson procedure tested H0: no difference between the effect of the two methods. The tests gave these results for the comparisons of the standard method and single site evolutionary model method:

for G3PD, failed to reject H0, (from Table 2.1)

for globin, rejected H0, (from Table 2.2)

for r.p. L20, failed to reject H0 (from Table 2.3)

The results are mixed Not a strong indication of that either method is better.

Multiple-site information

In this section I describe a prediction method that makes use of correlations between the states at multiple sites in a sequence.

Performance Measure

The performance measure for these experiments is the probability —of being the "new", related sequence— that the prediction method assigns to the left-out sequence. The prediction method assigns to any sequence of length k a probability, under the assumptions of the method's model, that that sequence is the "new" related sequence. In a comparison of two prediction methods, the one that assigns the higher probability to the left-out sequence is considered to be the better method.

Here is the probability of interest, that is the performance measure:

$$P\left(\text{LO\_seq}_{\text{FL}} \middle| X_{\text{FL}}\right)$$

In this term X refers to the data, the n aligned left-in sequences. And LO_seq refers to the particular left-out sequence. The "FL" subscripts indicate that the sequences (the left-out sequence and the known sequences) have length k.

The "experiment" (using that word as in statistical theory) relevant to the events X_{FL}, $\theta_{\text{gen,FL}}$ (above), and $\text{LO\_seq}_{\text{FL}}$ is: the determination of the previously unknown sequence of a gene homologous to the known aligned sequences. The set of possible outcomes of this experiment is: the set of all sequences of length k. The left-out sequence $\text{LO\_seq}_{\text{FL}}$ is one member of this sample space. So in the equation above the event "$\text{LO\_seq}_{\text{FL}}$" indicates that the outcome (of the experiment finding the sequence of the new, related gene) is the left-out sequence. I could make the equation clearer, though more busy, by writing "new=$\text{LO\_seq}_{\text{FL}}$" instead of just "$\text{LO\_seq}_{\text{FL}}$".

The variable $\text{LO\_seq}_{\text{FL}}$ refers to the left-out sequence of length k, as opposed to the sequence

at one site. In the equations below I will use LO_seq_S to refer to the S-th amino acid in the left out sequence. The variable LO_seq_S is an outcome of the experiment of determining the sequence at the S-th site of the related gene.

Calculation of performance measure by single-site evolutionary model method

Chapter 1 included a derivation of probabilities of particular sequences being the newly-determined related sequence, arriving at Equation 1.11

$$P(\alpha_{\text{gen,FL}} | X_{\text{FL}}) = \prod_{S=1}^{k} \frac{\sum_{\text{all A}} \sum_{\text{all } \beta} \left[t_{\beta\alpha} P(X_S | \beta_S, A) P(\beta_S | A) \right] P(A)}{\sum_{\text{all possible } \beta} P(X_S | \beta) P(\beta)}$$

Eq. 1.11

In this equation S is the site.

I do this set of calculations for a branch length of zero. So alpha equals beta. So to express Equation 1.11 in terms of alpha

$$P(\text{new} = \alpha_{\text{gen,FL}} | X_{\text{FL}}) = P(\text{root} = \alpha_{\text{gen,FL}} | X_{\text{FL}})$$
$$= \prod_{S=1}^{k} \frac{\sum_{\text{all A}} \sum_{\text{all } \beta} \left[P(X_S | \text{root} = \beta_S, A) P(\text{root} = \beta_S | A) \right] P(A)}{\sum_{\text{all possible } \beta} P(X_S | \beta) P(\beta)}$$

Eq. 2.1

Calculating the performance measure is the simple act of calculating the probability that Equation 2.1 assigns to the left-out sequence being the "new", related, sequence.

$$P\left(\text{root} = \text{LO\_seq}_{\text{gen,FL}} | X_{\text{FL}}\right)$$

$$= \prod_{S=1}^{k} \frac{\sum_{\text{all A}} \sum_{\text{all } \pi} \left[\text{P}\left(X_S | \text{root} = \text{LO\_seq}_S, A\right)\text{P}\left(\text{root} = \text{LO\_seq}_S | A\right)\right]P(A)\pi}{\sum_{\text{all possible } \pi} \text{P}\left(\text{X}_\text{S} | \pi\right)P(\pi)}$$

Eq. 2.2

A multisite evolutionary model based method

This section presents a method that allows correlated states at different sites to be selected together. The method presented above calculated the probability associated with a sequence from independent calculations at each site. In this multisite method, the expressions in the numerator and denominator of the RHS of the expression for $P\left(\pi_{\text{gen}} | X\right)$ (Equation 1.9) are both evaluated for multiple-site sequences, not sequences of length one.

This multisite method calculates the same probability as the single site method, $P\left(\pi_{\text{gen,FL}} | X_{FL}\right)$.

The derivation of this formula can start with Equation 1.9

$$P\left(\pi_{\text{gen}} | X\right) = \frac{\sum_{\text{all A}} \sum_{\text{all } \pi} \left[\text{t}_{\pi\pi'} \text{P}\left(X | \pi, A\right)\text{P}\left(\pi | A\right)\right]P(A)\pi}{\sum_{\text{all possible } \pi} \text{P}\left(\text{X} | \pi\right)P(\pi)}$$

Eq. 1.9

This equation is true for sequences of any length. So rewrite it to specify that it refers to sequences of length k, using the same notation as in Chapter 1.

$$P(\sigma_{\text{gen,FL}}|X_{\text{FL}}) = \frac{\sum_{\text{all A}} \sum_{\text{all } \sigma} \left[t_{\sigma_{FL} \to \sigma_{FL}} P(X_{\text{FL}}|\sigma_{\text{FL}}, A) P(\sigma_{\text{FL}}|A) \right] P(A)}{\sum_{\text{all possible } \sigma \text{ of len k}} P(X|\sigma_{\text{FL}}) P(\sigma_{\text{FL}})}$$

Eq. 2.3

Some terms on the RHS of Equation 2.3 can be replaced.

The transition rate $t_{\sigma_{FL} \to \sigma_{FL}}$, in the RHS numerator, can be rewritten, based on the assumption of independence between sites.

$$t_{\sigma_{\text{FL}} \to \sigma_{\text{FL}}} = \prod_{S=1}^{k} t_{\sigma_S \to \sigma_S}$$

Eq. 2.4

In Equation 2.4, as elsewhere, the subscript S indicates site.

Similarly, the terms $P(X_{\text{FL}}|\sigma_{\text{FL}}, A)$ and $P(\sigma_{\text{FL}}|A)$, also in the RHS numerator, can be replaced based on the assumption of independence between sites.

$$P(X_{\text{FL}}|\sigma_{\text{FL}} A) = \prod_{S=1}^{k} P(X_{\text{S}}|\sigma_{\text{S}} A)$$

Eq. 2.5

$$P(\sigma_{\text{FL}}|A) = \prod_{S=1}^{k} P(\sigma_{\text{S}}|A)$$

Eq. 2.6

The denominator of the RHS can be rewritten, also using the assumption of independence

between sites.

$$\sum_{\text{all } \tau \text{ of len k}} \left\{ P\left(X_{\text{FL}} | \tau_{\text{FL}}\right) P\left(\tau_{\text{FL}}\right) \right\} = \prod_{S=1}^{k} \left[\sum_{\text{all } \tau} \left\{ P\left(X_S | \tau_S\right) P\left(\tau_S\right) \right\} \right]$$

Eq. 2.7

Making these four substitutions into Equation 2.3 yields

$$P\left(\tau_{\text{gen,FL}} | X_{\text{FL}}\right) = \frac{\sum_{\text{all A}} \left[\sum_{\text{all } \tau} \left[\prod_{S=1}^{k} \left(t_{\tau_S \to \tau_S} \text{P}\left(X_{\text{S}} | \tau_{\text{S}} A\right) \text{P}\left(\tau_{\text{S}} | A\right) \right) \right] P(A) \right]}{\prod_{S=1}^{k} \left[\sum_{\text{all } \tau} \left\{ P\left(X_S | \tau_S\right) P\left(\tau_S\right) \right\} \right]}$$

Eq. 2.8

The term $\text{P}\left(\tau | A\right)$ is the *a priori* probability of sequence τ being the state of the root node. I use the stationary probabilities for these. The node A does not affect these values. So $\text{P}\left(\tau_{\text{S}} | A\right)$ can be rewritten as $\text{P}\left(\tau_{\text{S}}\right)$, meaning the stationary probabilities

$$P\left(\tau_{\text{gen,FL}} | X_{\text{FL}}\right) = \frac{\sum_{\text{all A}} \left[\sum_{\text{all } \tau} \left[\prod_{S=1}^{k} \left(t_{\tau_S \to \tau_S} \text{P}\left(X_{\text{S}} | \tau_{\text{S}} A\right) \text{P}\left(\tau_{\text{S}}\right) \right) \right] P(A) \right]}{\prod_{S=1}^{k} \left[\sum_{\text{all } \tau} \left\{ P\left(X_S | \tau_S\right) P\left(\tau_S\right) \right\} \right]}$$

Eq. 2.9

Equation 2.9 is the equation for the multisite evolutionary model based method. (I will sometimes refer to this method as the "multisite" method.)

In the particular experiments presented below, I consider the new branch from the attachment point node A (with state τ) to the new species tip node B (with state τ) to have length zero. So I can rewrite Equation 2.9, replacing occurrences of τ with τ, and also removing the term

$t_{\square_S \square \square_S}$. At the same time I will explicitly indicate that // is referring to the root state.

$$P\left(\text{root} = \square_{\text{gen,FL}} \middle| X_{\text{FL}}\right) = \frac{\square \square \square \left(\text{P}\left(X_{\text{S}} \middle| \text{root} = \square_{\text{S}} A\right)\text{P}\left(\text{root} = \square_{\text{S}}\right)\right) \square P(A) \square}{\square \square \left\{P\left(X_S \middle| \square_S\right)P\left(\square_S\right)\right\} \square}$$

Eq. 2.10

In these experiments I am interested in calculating a performance measure. Equation 2.10 is the basis for the same performance measure use above, $P\left(\text{LO\_seq}_{\text{FL}} \middle| X_{\text{FL}}\right)$, the probability that a particular method assigns to the left-out sequence. So the express Equations 2.10 in terms of the left-out sequence:

$$P\left(\text{root} = \text{LO\_seq}_{\text{gen,FL}} \middle| X_{\text{FL}}\right) = \frac{\square \square \square \left(\text{P}\left(X_{\text{S}} \middle| \text{root} = \text{LO\_seq}_{\text{S}}, A\right)\text{P}\left(\text{root} = \text{LO\_seq}_{\text{S}}\right)\right) \square P(A) \square}{\square \square \left\{P\left(X_S \middle| \square_S\right)P\left(\square_S\right)\right\} \square}$$

Eq. 2.11

This is the equation I use to calculate the performance measure for the multisite evolutionary model based method in the following experiments.

Numerical difficulties and solution

When calculating the statistics for the multisite cluster prediction method, underflow errors occurred. That is, the numbers in the calculation became smaller than the smallest number that a C++ double variable can represent. The smallest value a double can hold is on the order of $10^{\square 308}$. I worked around this problem by creating a C++ class that can hold larger

and smaller numbers than a C++ double variable can hold. I created this class following the "extended range" solution of Hauser (1996).

Does the size of the segments affect results?

The multisite evolutionary model based method detects correlations between sites in an alignment. So it is reasonable to think that the length of the segments considered in the analysis might affect the results. To investigate this question, I carried out an analysis similar to that described above, but dividing up the alignment into segments of different length each time. I divided the G3PD alignment four different ways: into segments of 7aa, 14 aa, 20 aa, and 28 aa. For each of these ways of dividing the alignment, I calculated the log P(LO_seq|LI_set) for each combination of segment and left-out sequence. So for each way of dividing the alignment, these calculations yielded 3 sets of $m \cdot 22$ log probabilities (where 'm' is the number of segments in the full-length alignment). I then calculated the average of all $m \cdot 22$ of these numbers and plotted these averages in Figure 2.7.

I am not interested in comparing these averages to determine which prediction method is best. Doing so could be misleading for two reasons. First, the $m \cdot 22$ log probabilities are not independent. Second, taking the average of log transformed numbers means giving a lot of weight to small numbers. And I do not want to weight that way for comparison.

The purpose of Figure 2.7 is to investigate if data with different branch lengths results in different relative performances from the three prediction methods. Figure 2.7 indicates that the relative performances of the different prediction methods do not change with segment length. So I will perform all subsequent experiments on alignments broken into 7 amino acid segments. Figure 2.7 indicates I can do experiments on segments of that length, and expect that conclusions I reach will probably apply to segments of other lengths as well.

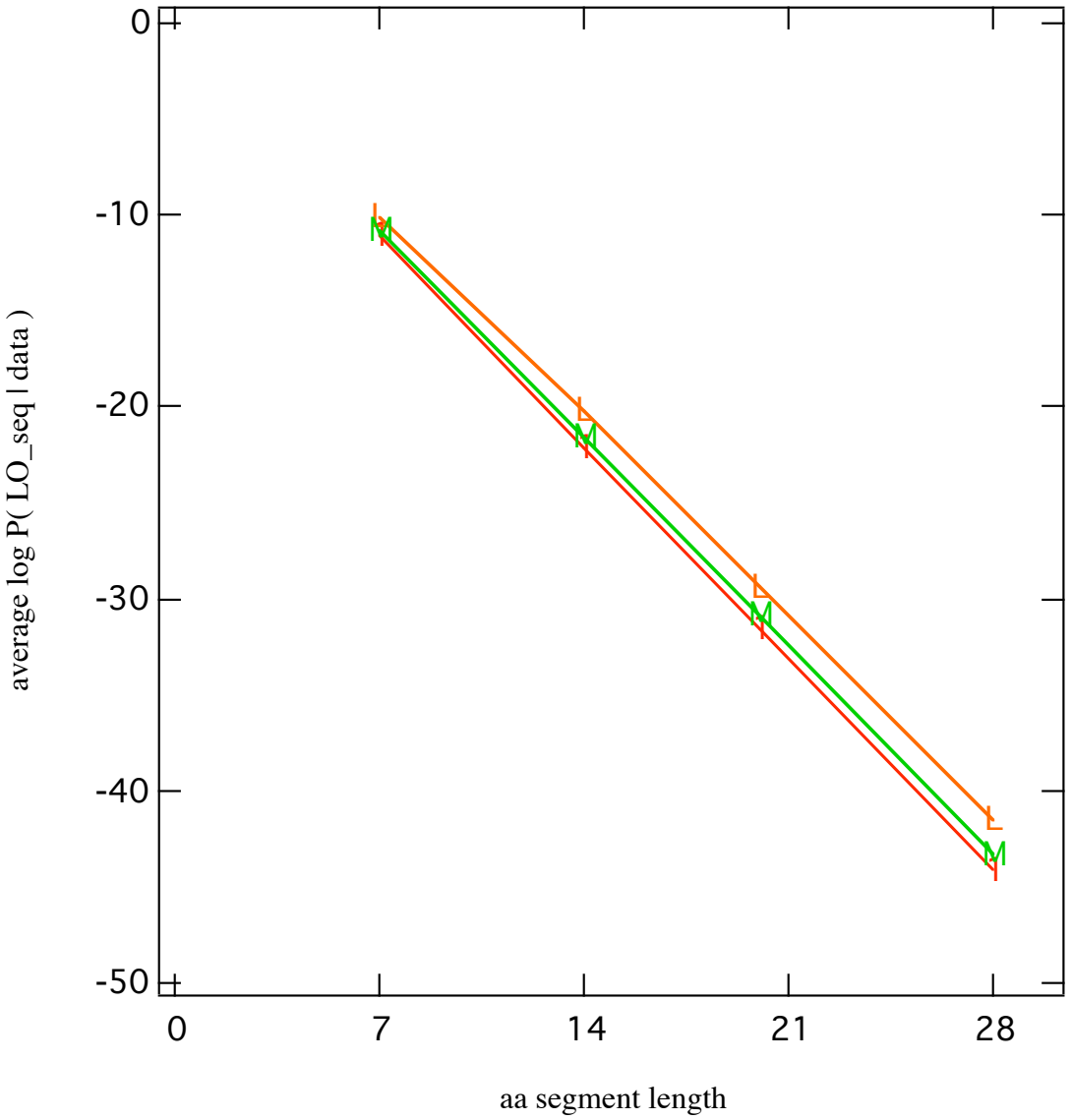

Figure 2.7 Comparison of analyses done on different segment lengths. Analysis of the alignment of 22 G3PD sequences. The 140 aa sequence was broken into shorter segments for analysis. The horizontal axis shows segment length. 'L' (yellow) indicates the multisite cluster evolutionary model-based method. '1' (red) indicates the site-by-site evolutionary model based method. 'M' (green) indicates the modified multinomial prediction

Comparison of
the multisite evolutionary model based method
and the standard primer design method

Using each of these methods, I calculated the "probability of left-out sequence" performance measure using my 22 sequence alignment of Glyceraldehyde 3-phosphate dehydrogenase (G3PD) genes.

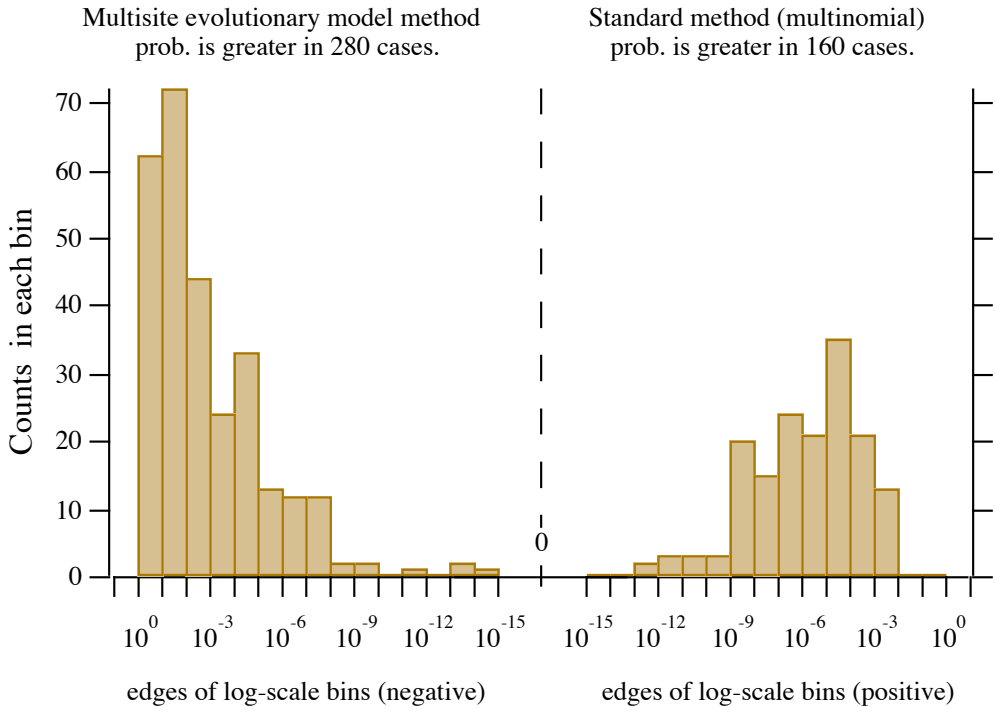

Figure 2.8. Log-scale histogram of differences in probabilities —G3PD alignment. The differences graphed here are P(LO_seq|LI_set) for the multinomial prediction method minus P(LO_seq|LI_set) for the multisite evolutionary model method. Notice these are not log(probabilities). The vertical axis is bin counts. The horizontal axe indicates the order of magnitude of the bins. All 440 differences are sorted into the bins o these histograms —one difference for each of the 22 left-out sequences, in each of the 20 7aa segments of the G3PD alignment.

The distribution of differences for the G3PD data

Figure 2.8 is a log-scale histogram showing the distribution of differences between paired (paired for each combination of segment and left-out sequence) performance measures for the multinomial and multisite evolutionary model methods, for the G3PD data.

Figure 2.8 indicates that the multisite evolutionary model based method is a better predictor that the multinomial distribution. Of the 440 differences, 280 are negative (i.e. the performance measure for the multisite evolutionary model based prediction method was higher than that for the multinomial prediction method) while 160 of the differences were positive. Notice particularly the 72 cases in the left-most bin of the histogram. In these 72 cases the multisite evolutionary model method scored between 0.1 and 1 higher than the standard method. And the second bin from the left of the histogram contains 62 cases in which the multisite evolutionary model method scored between 0.01 and 0.1 higher than the standard method. There are zero cases in which the multinomial method performance measure was that much greater than the multisite evolutionary model method performance measure. In other words: in none of the 440 comparisons did the multinomial method return a performance measure that was more than 0.01 greater than the performance measure returned by the multisite evolutionary model method. While in 134 of the 440 comparisons the reverse happened: the multisite method performance measure was more than 0.01 greater than the multinomial method performance measure. (These performance measures are all probabilities.)

The two-sided sign test for outcomes 280 and 160 returns a p value of 1.40841e-08. So we can reject the null hypothesis that the median of this distribution is zero. Multiply these numbers by $\frac{n-1}{n} = \frac{21}{22}$ yields a test of 267 and 153. The two-tailed sign test for 267 and 153 is 3.52537e-08.

The distribution of differences for the globin data

Figure 2.9 is a log-scale histogram showing the distribution of differences between paired (paired for each combination of segment and left-out sequence) performance measures for the standard primer design method (multinomial) and the multisite evolutionary model based method, for the globin data.

There are 200 such differences to sort into a log-scale histogram, for each pair of prediction

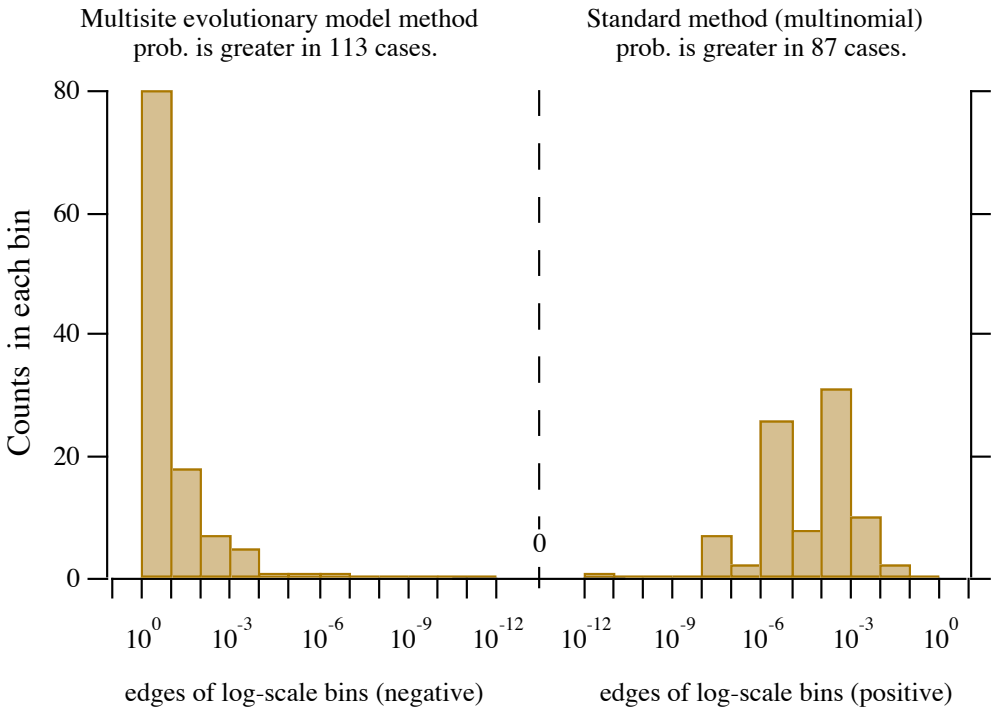

Figure 2.9. Log-scale histogram of differences in probabilities —globin alignments. The differences graphed here are P(LO_seq|LI_set) for the multinomial prediction method minus P(LO_seq|LI_set) for the multisite evolutionary model method. Notice these are not log(probabilities). The vertical axis is bin counts. The horizontal axe each indicate the order of magnitude that each bin in the histogram covers. All 200 differences are sorted into the bins of these histograms —one difference for each of the 5 left-out sequences, in each of the 40 7 aa segments of the alpha and beta globin alignments

methods.

The multisite method looks better. Out of 200 cases, in 98 the multisite evolutionary model performance measure is greater, by a magnitude 10^{-1} or 10^{-2}. (This time 80 of magnitude 10^{-1} and 18 of magnitude 10^{-2}.) While in only 2 cases is the multinomial performance measure greater than the multisite evolutionary model method performance measure, by more than 0.01. And in zero cases is the multinomial performance measure greater by more than 0.1.

The two-sided sign test for outcomes 113 and 87 returns a p value of 0.0771395. To do the adjustment for correlation (as explained in Chapter 1) I multiply the counts by $\frac{n-1}{n} = \frac{10}{11}$ giving 103 and 79 (rounded off). That two-sided sign test has a p value of 0.0882603

Distribution of differences between standard method and multisite evolutionary. model method, in rpL20

Figure 2.10 is a log-scale histogram showing the distribution of differences between paired (paired for each combination of segment and left-out sequence) performance measures for the standard primer design method (multinomial) and the multisite evolutionary model based method, for the ribosomal protein L20 data.

The distribution of differences between the multinomial and multisite evolutionary model based prediction methods indicates that the multisite evolutionary model method is better. Using my standard comparison of the highest order of magnitude differences. Of the 165 differences, 29 are of magnitude 10^{-1} or 10^{-2} where the difference favors the multisite method. But only one of the 165 differences is of this magnitude, where the difference favors

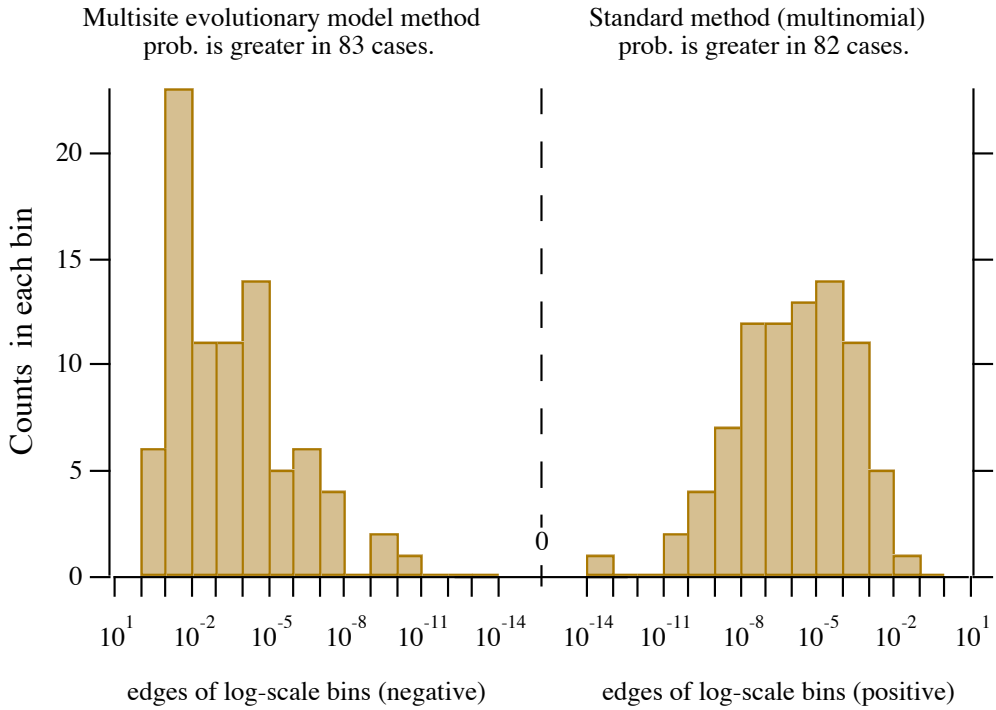

Figure 2.10. Log-scale histogram of differences in probabilities —ribosomal protein L20 alignment. The differences graphed here are P(LO_seqlLI_set) for the multinomial prediction method minus P(LO_seqlLI_set) for the multisite evolutionary model method. Notice these are not log(probabilities). The vertical axis is bin counts. The horizontal axis indicates the order of magnitude that each bin in the histogram covers. All 165 differences are sorted into the bins of these histograms —one difference for each of the 11 left-out sequences, in each of the 15 7 aa segments of the ribosomal protein L20 alignment.

the multinomial performance measure.

I do not need to actually perform the sign test to know that with outcomes 83 and 82, it will fail to reject the hypothesis that the median of the distribution is zero.

Summary of comparing differences between performance measures between multisite evolutionary model method and standard method

Table 2.5 summarizes the results of this section.

Table 2.5
Summary of comparisons: Which method has greater performance measures?

"n.s.d." means "no significant difference", i.e. the sign test failed to reject H0: same number of cases favoring each method. n = num of sequences in alignment.

2 Methods compared	**Gene**	**Which method has higher performance measures, based on each of these ways to compare differences . . .** count of differences > 0	sign test of count of differences >0	sign test of count * (n-1)/n	count of differences > 0.01
Multisite evolutionary model method vs. Standard method (multinomial)	G3PD	multisite	multisite	multisite	multisite
	globin	multisite	n.s.d.	n.s.d.	multisite
	r. p. L20	n.s.d.	n.s.d.	n.s.d.	multisite

Comparison of average (over left-in sets for a segment) performance measures of the multisite evolutionary model based method and the standard primer design method

The Wilcoxon-Nemenyi-McDonald-Thompson procedure tested H0: no difference between the effect of the two methods. The tests gave these results for the comparisons of these two methods:

for G3PD, reject H0 (from Table 2.1)
for globin, reject H0 (from Table 2.2)
for r.p. L20, reject H0 (from Table 2.3)

The results consistently indicate that the two methods are different, that the multisite has better performance measure scores.

Comparison of

the multisite evolutionary model based method

and the single site evolutionary model based method

Distribution of differences in multisite and single site performance measures for G3PD data

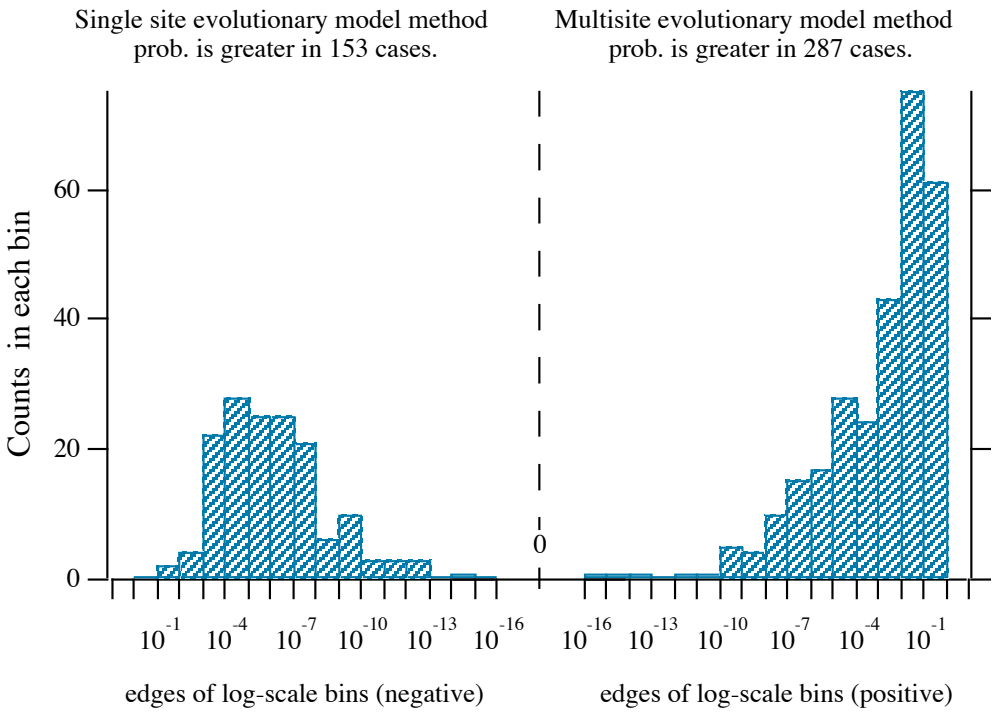

Figure 2.11. Log-scale histogram of differences in probabilities —G3PD alignment. The differences graphed here are P(LO_seq|LI_set) for the multisite evolutionary model method minus P(LO_seq|LI_set) for the single site evolutionary model method. Positive differences (those on the right side of graph) indicate the multisite prediction method gave a greater performance measure. Negative probabilities (those on the left side of graph) indicate the single site method gave a greater probability. Notice these are not log(probabilities). The vertical axis is bin counts. The horizontal axis indicates the order of magnitude that each bin in the histogram covers. All 440 differences are sorted into the bins of these histograms —one difference for each of the 22 left-out sequences, in each of the 20 7aa segments of the G3PD alignment.

Figure 2.11 is a log-scale histogram showing the distribution of differences between paired (paired for each combination of segment and left-out sequence) performance measures for the multisite evolutionary model method and the single site evolutionary model method, for the G3PD data.

Figure 2.11 indicates that the multisite method is a better predictor for this G3PD data. In the majority of the cases (287 to 153) the multisite method had a higher score (performance measure).

And the multisite method also has the large majority of cases in which one method scored significantly higher than the other. The two bins on the far right of the graph contain the cases in which the multisite method scored between 0.1 and 1 higher than the single site evolutionary model method. In 75 of the 440 cases the multisite method outscored the single site method by a value between 0.01 and 0.1. And in 61 of the 440 cases the multisite method outscored the single site method by a value between 0.1 and 1. On the left side of the graph there are many fewer cases in the analogous bins for the single site method. There are only two and zero cases, respectively, in the bins in which the single site method outscored the multisite method by values of those magnitudes.

A two-tailed sign test of 287 and 153 returns a p value of 2.30241e-10. To do the adjustment for correlation (as explained in Chapter 1) I multiply the counts by $\frac{n-1}{n} = \frac{21}{22}$ giving 274 and 146 (rounded off). The p value for that sign test equals 5.78596e-10.

Distribution of differences in multisite and single site performance measures for globin data

Figure 2.12 is a log-scale histogram showing the distribution of differences between paired (paired for each combination of segment and left-out sequence) performance measures for the multisite evolutionary model method and the single site evolutionary model method, for the alpha globin and beta globin alignments.

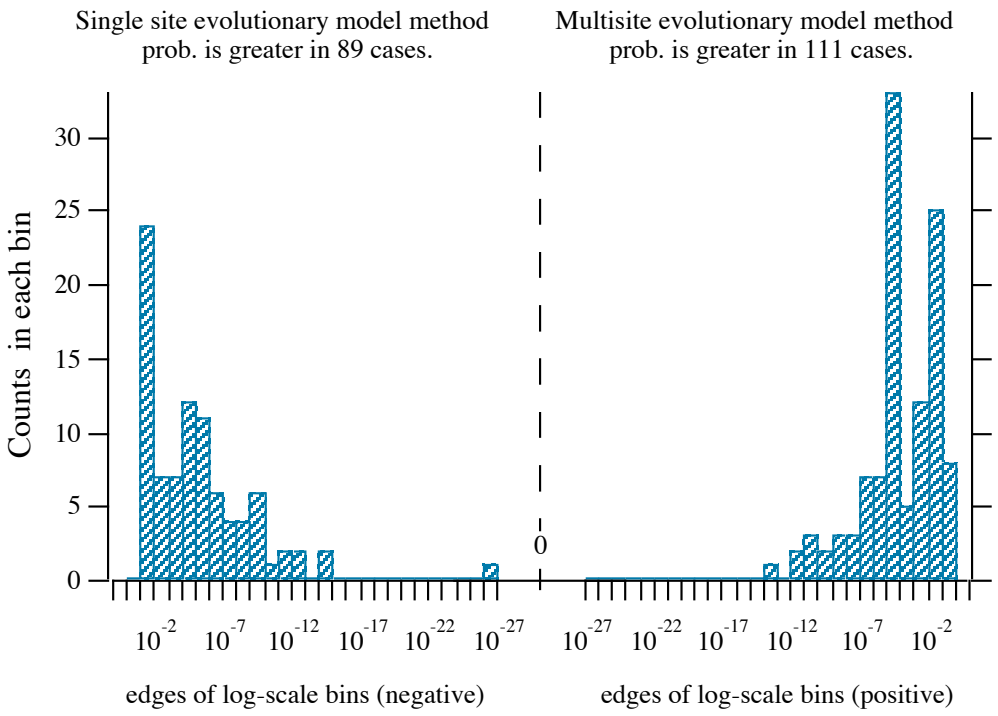

Figure 2.12. **Log-scale histogram of differences in probabilities —globin alignments**. The differences graphed here are P(LO_seq|LI_set) for the multisite evolutionary model prediction method minus P(LO_seq|LI_set) for the single site evolutionary model method. Notice these are not log(probabilities). The vertical axis is b counts. The horizontal axis runs from -1 on the left to 1.0 on the right: the values indicati the order of magnitude that each bin in the histogram covers. All 200 differences are sorte into the bins of these histograms —one difference for each of the 5 left-out sequences, in each of the 40 7 aa segments of the globin alignments.

Figure 2.12 indicates that the multisite evolutionary model method is a better predictor on the globin data, than the single site evolutionary model method. Though this superiority is not as pronounced as it was for the G3PD data above.

In the globin data, the multisite method is greater in the majority of cases: 111 to 89. And the multisite method has more cases in which it is significantly greater than the single site prediction method. There are 25 cases and 8 cases, respectively, in which the multisite method

performance measure is greater than the single site performance measure by between 0.01 and 0.1, and by between 0.1 and 1.0. There are 24 cases and zero cases, respectively, in which the single site method performance measure beats the multisite by differences of those magnitudes.

The two-tailed sign test for 111 and 89 returns a p value of 0.137616. To do the adjustment for correlation (as explained in Chapter 1) I multiply the counts by $\frac{n-1}{n} = \frac{4}{5}$, giving 89 and 71 (after rounding off) The p value for that sign test is 0.179017.

Distribution of differences between multisite and single site performance measures for ribosomal protein L20 data

Figure 2.13 is a log-scale histogram showing the distribution of differences between paired (paired for each combination of segment and left-out sequence) performance measures for the multisite evolutionary model method and the single site evolutionary model method, for the ribosomal protein L20 data.

Figure 2.13 indicates one thing based on the majority of all differences, and another thing based on the majority of large (between 0.01 and 1.0) differences. The majority of all differences favors the single site method over the multisite method by a ratio of 96 to 69. But the multisite performance measure is greater than the single site performance measure by between 0.01 and 0.10 in 21 cases, and greater by between 0.10 and 1.0 in 5 cases. The single site performance measure is greater than the multisite by those magnitudes in only 3 and zero cases, respectively.

The two-sided sign test for outcomes 69 and 96 returns a p value of 0.0429878. To do the adjustment for correlation (as explained in Chapter 1) I multiply the counts by $\frac{n-1}{n} = \frac{10}{11}$,

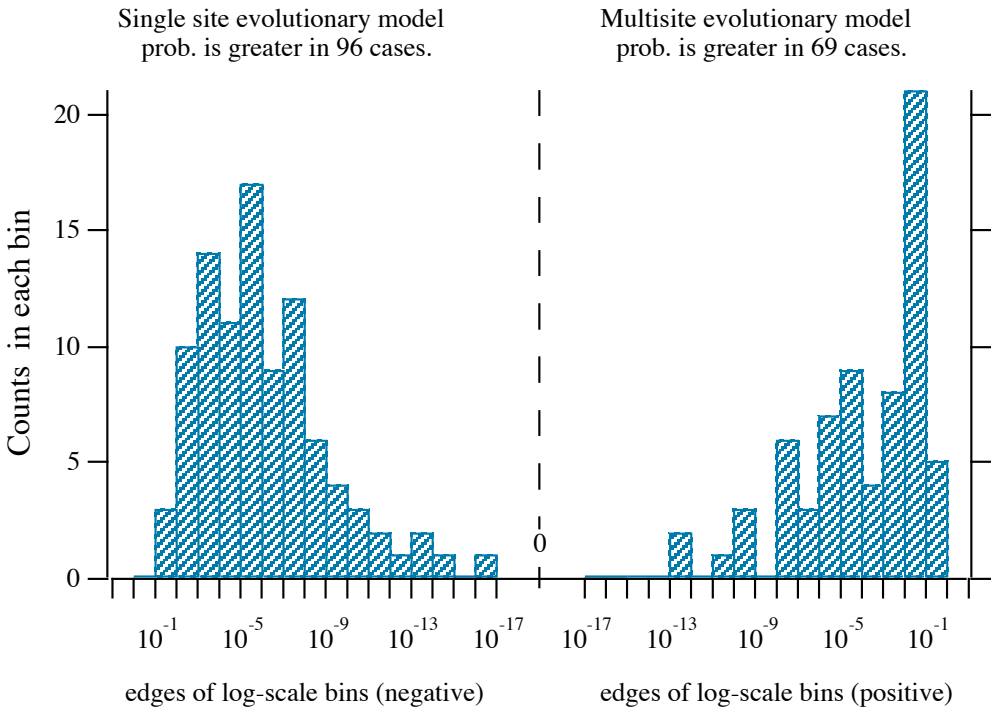

Figure 2.13. Log-scale histogram of differences in probabilities —ribosomal protein L20 alignment. The differences graphed here are P(LO_seqlLI_set) for the multisite evolutionary model prediction method minus P(LO_seqlLI_set) for the single site evolutionary model method. Notice these are not log(probabilities). The vertical axis is b counts. The horizontal axis indicates the order of magnitude that each bin in the histogran covers. The horizontal axis runs from -1 on the left to 1 on the right. All 165 differences are sorted into the bins of these histograms —one difference for each of the 11 left-out sequences, in each of the 15 7 aa segments of the ribosomal protein L20 alignment.

giving 63 and 87 (rounded off). The p value for that sign test equals 0.0604233.

Summary of comparing differences between performance measures between multisite evolutionary model method and single site evolutionary model method

Table 2.6 summarizes the results of this section.

Table 2.6
Summary of comparisons of which method has greater performance measures

"n.s.d." means "no significant difference", i.e. the sign test failed to reject H0: same number of cases favoring each method. n = num of sequences in alignment.

2 Methods compared	**Gene**	**Which method has higher performance measures, based on each of these ways to compare differences . . .** count of differences > 0	sign test of count of differences >0	sign test of count * (n-1)/n	count of differences > 0.01
Multisite evolutionary model method vs. Single site evolutionary model method	G3PD	multisite	multisite	multisite	multisite
	globin	multisite	n.s.d.	n.s.d.	multisite
	r. p. L20	single site	single site	n.s.d.	multisite

Comparison of average (over left-in sets for a segment) performance measures of the multisite evolutionary model based method and the single site evolutionary model based method

The Wilcoxon-Nemenyi-McDonald-Thompson procedure tested H0: no difference between the effect of the two methods. The tests gave these results for the comparisons of these two methods:

for G3PD, reject H0 (from Table 2.1)

for globin, fail to reject H0 (from Table 2.2)

for r.p. L20, fail to reject H0 (from Table 2.3)

The results are not consistent. So there is not strong evidence that the multisite method is a better predictor than the single site evolutionary model based method.

Correlation of P(LO_seq|LI_set)'s within a segment

It is of interest, because of the sign test, and just because of interest in how correlated the sequences are, to know where in the alignment the sequences that score similarly are located. Figure 2.14 shows where in the globin alignments are sequences within segments, whose P(LO_seq|LI_set)'s for the single site method were greater, by at least a difference of 0.01, than the P(LO_seq|LI_set)'s of the multinomial method. One can see that in some cases, all of the sequences within an alignment scored similarly, and in other cases, they scored differently. Figure 2.15 shows similar information for the ribosomal protein L20 alignment. Again one can see that the correlations between performance measures within a segment are not 100%.

Comparisons of other pairs of methods on these alignments (not shown) and on the G3PD alignment have the same mixture of correlations within segment as that shown in Figures 2.14 and 2.15.

A. alpha globin alignment

aa site:	0- 6	7- 13	14- 20	21- 27	28- 34	35- 41	42- 48
aa APPI:	0.83	0.6	0.61	0.8	0.63	1	0.73
human:	VISPANK	TLVKAAW	EKVEAGA	EQYEAQA	IQRMFIS	FPTTKTY	FPGFNIS
goat_cow:	VISAANK	SLVKAAW	EKVEELA	EAYEAQA	IQRMFIS	FPTTKTY	FPGFNIS
rabbit:	VISPANK	TLHKTAW	QKHESGE	EQYEAQA	VQRMFIE	FPTTKTY	FPGFNFT
rat:	VISANNK	TLHKLDW	EKHEEGE	EQYEQQA	ICRMFAA	FPTTKTY	FSGHNVS
marsupial:	VISNANK	TGVKAHW	EKVEEGA	EAYAAQA	IARTFIS	FPTTKTY	FPGFNIS

aa site:	49- 55	56- 62	63- 69	70- 76	77- 83	84- 90	91- 97
aa APPI:	0.71	0.8	0.64	0.69	0.67	1	1
human:	GESACVK	EGEKKVA	NAITLAV	AGVNNMP	LAISAIS	NIGAGKI	RVNPVLF
goat_cow:	GESACVK	EGEQKVA	AAITKAV	EGINNIP	ETISNIS	NIGAGKI	RVNPVLF
rabbit:	GESQCHK	AGEKKVS	QAITKAV	EGINNIP	EAISTIS	NIGAGKI	RVNPVLF
rat:	PESACVK	AGEKKVA	NAIAKAA	NGVQNIP	EAISTIS	NIGAGKI	RVNPVLF
marsupial:	PESACHC	EGEKKVA	NAISCAV	AGINNIP	ETMSKIS	NIGAGKI	RVNPVLF

aa site:	98-104	105-111	112-118	119-125	126-132	133-139
aa APPI:	0.89	0.83	0.57	0.83	0.8	1
human:	KIISGDI	IVTIAAG	IPAQFTP	AVGASIN	KFIASVS	TVITSKY
goat_cow:	KIISGSI	IVTIADG	IPLNFTP	AVGASIN	KFIALVS	TVITSKY
rabbit:	KIISGDI	IVTIALG	GPSQFTP	AVGASIN	KFIALVS	TVITSKY
rat:	KFISGDI	IVTIADG	GPENFTP	AMGASIN	KFIASVS	TVITSKY
marsupial:	KIISGDI	HVTIAAG	ISKNITP	QVGASMN	KFFASVA	TVITSKY

B. beta globin alignment

aa site:	0- 6	7- 13	14- 20	21- 27	28- 34	35- 41	42- 48
aa APPI:	0.64	0.61	0.63	1	0.94	0.79	0.83
human:	ITPQQKS	AVTAIWE	KVLVNQV	EEQAIER	IIVVYPW	TCRFFQS	FENISTP
goat_cow:	ITAQQKA	AVTAFWE	KVKVNQV	EEQAIER	IIVVYPW	TCRFFQS	FENISTA
rabbit:	ISSQQKS	AVTAIWE	KVLVQQV	EEQAIER	IIVVYPW	TCRFFQS	FENISSA
rat:	ITNAQKA	AVLAIWE	KVLPNNV	EEQAIER	IIVVYPW	TCRYFNS	FENISSA
marsupial:	ITSQQKL	DHTTHWS	KVCVNCT	EEQAIER	MIVVYPW	TTRFFES	FENISSP

aa site:	49- 55	56- 62	63- 69	70- 76	77- 83	84- 90	91- 97
aa APPI:	0.64	0.89	0.61	0.63	0.89	0.87	1
human:	NAVMELP	KVKAGEK	KVIEAFS	NEIAGIN	LIKETFA	TISQIGD	NKIGVNP
goat_cow:	NAVMLLP	KVKAGEK	KVINSFS	LEMKGIN	NIKETFA	AISQIGD	NKIGVNP
rabbit:	LAVMLLP	KVKAGEK	KVIAAFS	QEISGIN	LIKETFA	KISQIGD	NKIGVNP
rat:	SAHMELP	KVKAGEK	KVHLAFL	NEIKGIN	LIKETFA	GISQIGD	NKIGVNP
marsupial:	EAVMSLS	KVCAGEA	KVITSFE	QAVKGIN	LIKETYA	KISQIGD	NKIGVNP

aa site:	98-104	105-111	112-118	119-125	126-132	133-139
aa APPI:	0.86	0.66	0.63	0.67	0.77	0.9
human:	QLFRIIE	LVIVDVI	AGGFEKQ	FTPPVCA	AYCKVVA	EVALAIA
goat_cow:	QLFKIIE	LVIVVVI	ARLFEKQ	FTPVICA	NFCKVVA	EVALAIA
rabbit:	QLFRIIE	LVIVHVI	SGGFEKQ	FTPCVCA	AYCKVVA	EVALAIA
rat:	QLFRIIE	LMHVHVI	EGGIEKQ	FTPDACA	AFCKVVA	EVASAIA
marsupial:	QLFKMIE	LHHVHDI	AQGFEKN	FTPQDCV	AWCKIVA	EVAGAIA

Figure 2.14. Location, in alpha and beta globin alignments, of sequences which the single site evolutionary model method predicted significantly better than the multinomial method predicted. Red indicates such sequences for which the performance measure of the evolutionary model method was 0.01 or more greater than the corresponding performance measure for the multinomial method.

```
aa site:        0-  6     7- 13    14- 20    21- 27    28- 34    35- 41    42- 48
aa APPI:          0.6       0.51      0.37       0.5      0.46      0.54       0.5
RL20_ECOLI: ARVKREV   HARARGK   KHIKCAK   EYYEARS   RVYRVAF   CAVHKAE   CYAYRNR
RL20_BACST: PRVKEEP   VTRRRRK   KVIKIAK   EYFEAKG   AIYRVAL   CCVMKSI   MYAYRNR
RK20_CYAPA: TRVKREL   VARKRRK   KHIKIAS   EFREAGS   RIFRVAL   CCVMKAI   RYAYLNR
RK20_CHLRE: TRVKREL   VSRKRGK   KHILMSK   EFREAAS   TIFRTAL   CCLMKAI   RYSYRLR
RK20_TOBAC: TRHKREY   HARRRRT   KHRIFAS   SFREAGS   RITRTHT   CCKHRAI   VSAGRNR
RK20_EPIVI: TRHKREY   HARRRHK   KFRIFAS   SFILAGS   RITRTHT   CCKHRAI   VSSNRNR
RK20_MAIZE: TRVPREY   HARRRRT   KMRSFAS   LFREAGI   RILRVHT   CCVRRAF   VSSGRNR
RK20_PINTH: TRVKREY   HARKRRK   KHIAFVS   ESREAGS   KIFRTAL   CRKARAI   VSAGRNR
RK20_MARPO: TRVKREY   VARKRRK   LHITITS   EFCETGS   KIFRTAL   CCEMRAI   ASSGRNR
RK20_EUGGR: TRHKLLE   HSKKKRK   RKHSKMK   EWVEEGS   KIFRTEL   CCIMKAR   GYAFYNR
RK20_ASTLO: HRHKNLF   KLYKKGK   KHIKISK   EFHESLS   KHFKHSL   CKHMKAI   YFSFSNR

aa site:       49- 55    56- 62    63- 69    70- 76    77- 83    84- 90    91- 97
aa APPI:         0.47       0.7      0.49      0.37      0.35      0.34      0.55
RL20_ECOLI: RCRKRCF   RCIWHAR   HLAAARC   LEHSYSK   FHLEIKK   ASVQHNR   KHIANHA
RL20_BACST: RCRKRNF   RKIWHVR   HLAAARC   LEISYSR   IMGEIKI   AEVQVLR   KHVANIA
RK20_CYAPA: LKRKRNF   RAIWHAR   HLASARI   QEMTYSK   IMESIKK   ILHHILR   KSISCIA
RK20_CHLRE: RCKKRNF   RRMWHTR   VLSAVRR   YEILYSQ   FMLYIKT   GKHCILR   KVHACIS
RK20_TOBAC: NRKKRNF   RRIWHTR   HLAVHRE   VSYSYSR   IHGNIYK   RCIIILR   KHIACHA
RK20_EPIVI: LLKKRKF   RSIWHTR   HLAVHRE   VSYSYKL   FHYACYK   HCIIHLR   KHIACHA
RK20_MAIZE: ERCKRNF   RRIWHTR   HLAATRL   VFLSYSK   IHGLISK   KQIHILR   KMIACVA
RK20_PINTH: EKRKRNI   RRIWHTR   HLAAARA   LEVSYLR   FHCYIYK   RCIIPLR   KTIACHA
RK20_MARPO: EKRKRLI   RRIWHTR   VLAAARN   LEHSYLK   IHQYIYK   KKHIILR   KHIACHA
RK20_EUGGR: KKKKLIL   KTIWHTR   HLEEIKT   HLQKYLH   FVSFIRK   TKTYVLK   KIISQHL
RK20_ASTLO: KKKKRSL   KKHWHLR   HLYFIKF   LLFLYLK   HHLKHKI   LKHDHLK   KHISQHH

aa site:       98-104
aa APPI:         0.29
RL20_ECOLI: VFNKVAF
RL20_BACST: VLNCAAF
RK20_CYAPA: HYNKNAF
RK20_CHLRE: HDNPQAF
RK20_TOBAC: HSLRLDI
RK20_EPIVI: HILRLFF
RK20_MAIZE: VSLPLLI
RK20_PINTH: VINSLDF
RK20_MARPO: HINKFDF
RK20_EUGGR: VRNSKSF
RK20_ASTLO: HLNYKTH
```

Figure 2.15. Location, in r. p. L20 alignment, of segments for which the single site performance measure predicted significantly better than the multinomial method. Red indicates location, in r.p. L20 alignment, of sequences whose performance measures for the single site method were greater, by at least 0.01, than the performance measure for the multinomial prediction method. (There were no sequences in this alignment for which multinomial was greater than single site by this magnitude difference.

Chapter 3

Using Information in Clusters

The following experiment shows that using the information inherent in clusters of tip nodes within a tree allows one to make better predictions of "new", related sequences than simply using the information of the whole undifferentiated tree allows.

By a "cluster" I mean a set of tip sequences that are more closely related to each other than to the rest of the tree. I do not define a quantitative cutoff for "more closely related." One thinks of a cluster as a part of a tree that to the eye stand out as a tight branch on the tree. But statements about clusters can apply, generally, to any contiguous set of tips.

About the sequence data

The G3PD alignment (described above) was chosen because of its cluster nature. The three clusters within the tree are: 6 species of Archaea, 7 species of proteobacteria, and 9 species of fungi. The average pairwise percent identities of the sequences within each of the full clusters is shown in Table 3.1

Table 3.1
Average Pairwise Percent Identities
of sequences within each cluster

Number of species in cluster	Clade	Average Pairwise Percent AA Identity	std. deviation of average pairwise percent identity
6	Archaea	0.599	0.080
7	Proteobacteria	0.540	0.103
9	Fungi	0.752	0.054
22 (entire tree)	entire tree	0.470	0.200

Performance Measure

For this experiment the measure of how well, relatively, two prediction algorithms work is the percent identity between the predicted sequence and the "true" or left-out sequence. A better prediction method will have a higher average percent identity (between the 2 sequences) over all the left-in sets of sequences.

Description of the prediction algorithm that makes use of cluster information

This algorithm predicts a left-out sequence, based on a left-in set of sequences, using the information that is in clusters, in addition to the alignment and tree information used by the previously described methods.

A researcher following this algorithm would do the following steps using all his known sequences. (For the 22 sequence alignment tested, the set of "known" sequences contains 21 sequences.) The first step is the inspect the phylogenetic tree, just by eye, and assign each tip node to one "cluster", based on which sets of tips are relatively close to each other. In Figure 2.1 the clusters are easily seen.

Next, just the sequences in each of the three clusters were used, in turn, as the "known set", each generating one candidate predicted sequence. (For each left-in set, one of the three clusters will have one fewer species in it than that cluster had in the full 22-species set, the missing species being the left-out species. So when I used just the sequences in a cluster to predict, I never included the left-out sequence as a member of the cluster set attempting to predict that sequence.)

The method of generating each candidate predicted sequence was exactly as done as described above in this dissertation. That is, at each amino acid site, I used equation 1.9 to calculate a $P(\square_{\text{gen}}|X)$ for each $\square_{\text{gen}}$ (each of the 20 amino acids).

$$P(\square_{\text{gen}}|X) = \frac{\sum_{\text{all A}} \sum_{\text{all } \square} \left[t_{\square\square\ \square} P(X|\square, A) P(\square|A) \right] P(A)}{\sum_{\text{all possible } \square} P(X|\square) P(\square)}$$

Eq. 1.9

(Each variable in this equation refers to a sequence of length one aa or one codon.) But the "known set" was considered to be only the cluster, not the entire 21 species. That is, the weighted average in the numerator of equation 1.9,

$$\sum_{\text{all A}} \sum_{\text{all } \square} \left[t_{\square\square\ \square} P(X|\square, A) P(\square|A) \right] P(A)$$

was calculated over all the attachment points in a cluster (not all the attachment points in the entire tree of 22 sequences).

The state with the highest probability was chosen as the predicted state at each site, for the current cluster.

At this point in the algorithm, each cluster has generated a full-length candidate sequence. Now one final sequence (the "predicted sequence") is chosen from these candidates. The method of choosing is to simply choose the sequence, of the three, that has the best value of its full-length $P(\square_{\text{gen}}|X)$performance measure. So at this point in the algorithm, the full-length percent identity is calculated for each of the candidates, and the candidate with the highest percent identity is chosen as the one predicted sequence.

Results of comparing three prediction methods

The prediction method using cluster information, described just above, was compared with the

multinomial prediction method (representing the standard primer design method) and the single-site evolutionary model based prediction method. (Both of these methods are described in detail in Chapter 1.) I compared the three methods' accuracies at predicting sequences left out from the G3PD alignment. Averages of the results are shown in Table 3.2.

The performance measure used for this comparison (and shown in Table 3.2) is "average percent identity between the predicted sequence and the left-out sequence." A value of this measure is calculated for each of the 3 prediction methods. This measure is an average percent identity. A percent identity is calculated for each combination of site and left-in set of sequences at that site. A percent identity for one of these combinations has a value of either zero or one: zero if the predicted amino acid sequence and the left-out amino acid sequence do not match, one if they do match. So the average percent identity is an average of zeros and ones. The averages shown in Table 3.2 are averages over all combinations of site and left-in set of sequences at that site. (Notice that this average percent identity is different than the average pairwise percent identity of the alignments themselves.)

The values in Table 3.2 indicate that using the information in the clusters results in better prediction: the cluster method gives 69% identity, whereas the other two give about 50% identity.

Table 3.2
Average percent identities of each prediction method

"% identity" is between predicted aa sequence and left-out aa sequence. It is an average over all combinatio of site and left-in set of sequences at that site.

Prediction Method	mean % identity	std. dev. of % identity
Cluster-information-using	0.69	0.23
Multinomial	0.52	0.22
single site evolutionary model method	0.50	0.23

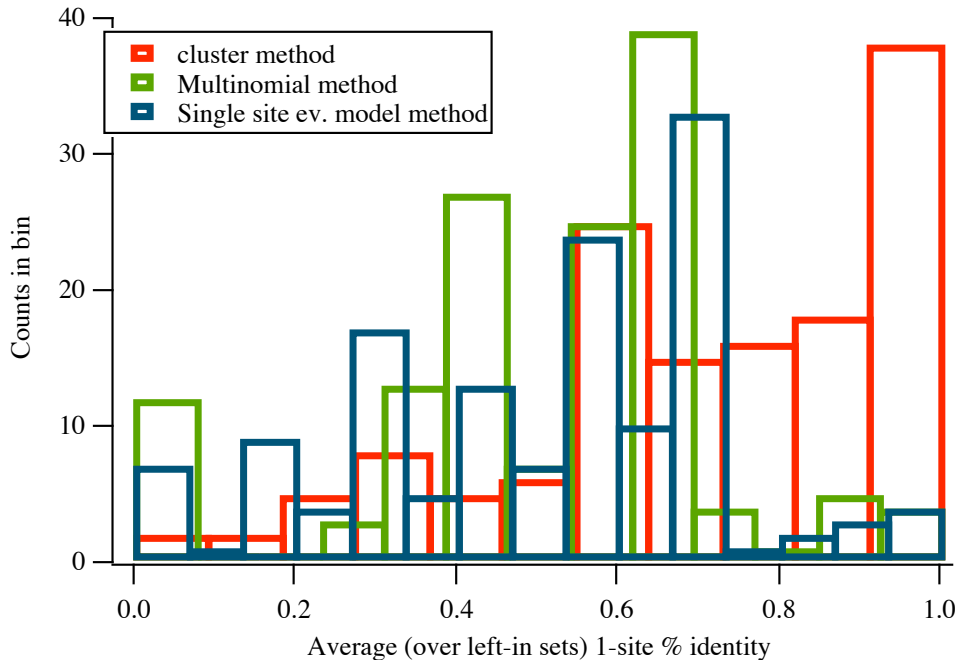

Figure 3.1. Percent identities are not normally distributed. The graph shows three histograms. Each histogram shows the distribution of average (over left-in sets for a site) percent identities calculated using one of the three prediction methods.

The average (over left in sets for a site) percent identities are not normally distributed, as shown in Figure 3.1. So it was decided to again use the Wilcoxon-Nemenyi-McDonald-Thompson multiple comparison procedure.

Table 3.3 shows that the cluster method has the best (ranked closer to "1st" more often) rank sum of the three methods. These results indicate that using the information in clusters lets one predict related sequences better than methods that do not use such information.

It is interesting to note that the multinomial method does a little better, in ranking, than the single site evolutionary model method, on the G3PD data. This result is consistent with the results of the comparisons done on the P(LO_seq|LI_set) performance measure in Chapter 2. Those Chapter 2 data, as the data here do, lean toward the standard method being better than the single site evolutionary model based method, but the Wilcoxon-Nemenyi-McDonald-

Table 3.3
Cluster method comparisons
Results of Wilcoxon-Nemenyi-McDonald-Thompson tests
on the G3PD alignment data

"Rank total" is the sum (over segments) of the within-site ranks (1, 2, or 3) for that method.. θ_1 and θ_2 in H0 refer to equation 1.19. Overall α for the 3 tests = 0.05.

Prediction method	**Rank total**
Cluster method	205.5
Multinomial (representing standard method)	311.5
Single site evolutionary model method	323

Methods compared	**Result of test. Reject H0: $\theta_1 = \theta_2$?**
Cluster & Multinomial (representing standard method)	Reject
Cluster & Single site ev. model method	Reject
Multinomial & Single site ev. model method	Fail to reject

Thompson procedure (in both Chapters 2 and 3) does not reject the null hypothesis that there is no difference between the methods.

Comparison of the prediction methods on trees of different average evolutionary distances

Results in Chapter 2 indicate that the multisite evolutionary model method predicts better than the multinomial distribution method. It would be interesting to see if this relative performance holds when the prediction is done for trees (alignments) with different average evolutionary distances between the tip species. Of course the three trees used in Chapter 2 have different average distances between the sequences. But it would be interesting to compare the methods on a set of tree (alignments) in which the only difference is the average distance between the sequences, and other parameters such as the topology (meaning just the pattern of connections between the nodes and edges) and the relative lengths of all the branches in the tree were held constant. To create a set of trees and associated amino acid sequence alignments that have these desired characteristics, I started with topologies and the various branch lengths, and using these generated artificial alignments via simulated evolution. Then, using these artificial alignments, I carried out a set of analyses comparing the three different prediction methods.

Creating the Simulated Data

The topology of the tree I used in these experiments is that of the tree associated with the Goldman group alignment of amino acid sequences that are the 22 C-terminal domains of glyceraldehyde 3-phosphate dehydrogenase (G3PD).

But the aa sequences I used in the current set of experiments were not the actual G3PD sequences. Instead they were sequences generated randomly using the following process. First a random ancestor sequence of length 140 aa's was created. The aa at each site in this ancestor sequence was drawn from the stationery distribution of aa's reported in Jones

(1992). This sequence was associated with the node designated to be ancestral to all other nodes in the tree. The Goldman tree is unrooted, so I simply chose one node to be the ancestral one. A different choice of root node would not affect the results of these tests.

Then this root sequence was subjected to simulated evolution, following each daughter branch from the ancestor node. The probabilities of amino acid substitutions in this simulated evolution was calculated using the same kind of models of amino acid evolution used throughout this dissertation. This simulation continued until a sequence was generated for every tip node.

I created five different alignments via this simulated evolution. The difference between the process used to generate these five alignments was the branch length factor. I used a different set of branch lengths each time: the branch length reported by Goldman, and those lengths times 0.25, 0.5, 2.0, and 4.0. The purpose of this distribution of lengths was to see if the three different prediction methods might have different relative strengths and weaknesses, depending on how similar to each other the sequences in an alignment are.

I calculated the average pairwise percent identity (between amino acids) for each alignment. These statistics are shown in Table 3.4

Table 3.4
aa APPI for each tree

branch length factor	average pairwise % identity
0.25	0.766
0.5	0.502
1.0	0.322
2.0	0.234
4.0	0.125

For this comparison, I broke the G3PD aa alignment into segments four different ways: 5 segments of 28 aa's, 7 segments of 20 aa's, 10 segments of 14 aa's and 20 segments of 7 aa's.

The P(LO_seq|LI_set)'s for the single site evolutionary model method and the multinomial method were calculated as described above in Chapter 1. The P(LO_seq|LI_set)'s for the multisite evolutionary model method were calculated as described in Chapter 2. The P(LO_seq|LI_set)'s were log transformed before being plotted. There was one of these probabilities (expressed as its log) for each combination of segment (of the 140 aa alignment) and left-in set of sequences. So for q = the number of segments in the alignment, the data consisted of q x 22 data points for each of the three methods, for each of the five simulated-evolution alignments.

The means of these data sets are graphed in Figures 3.2, 3.3, 3.4, and 3.5.

Interpretation of Figures

Figures 3.2 through 3.5 indicate that the three prediction methods maintain similar relative performances as the average distance between species in a tree is varied. The aspect that stays the same is that the multisite method consistently performs better than the other two prediction methods (in a rough inspection of the graphs, with no formal test of significance.)

One aspect that changes is that the multinomial method performs better than the single site evolutionary model method in the alignments with higher (above 0.4) APPI's, but worse than the single site method in trees with low (less than 0.4) APPI's.

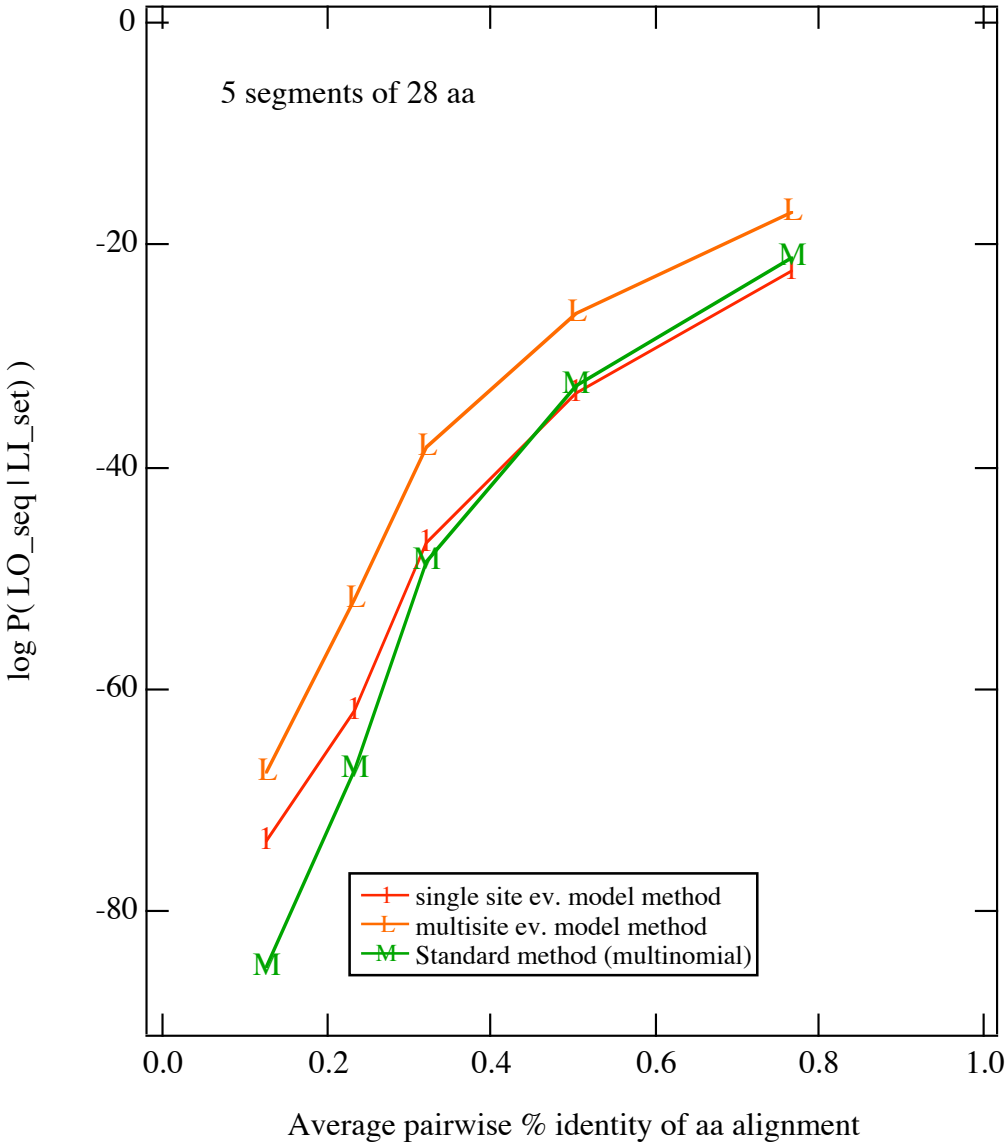

Figure 3.2. Relative performance of three prediction methods on alignments with different APPI's —segments of length 28 aa's. Data is simulated. (See text.) L=multisite cluster prediction method. 1=single site evolutionary model prediction method, M=multinomial prediction method.

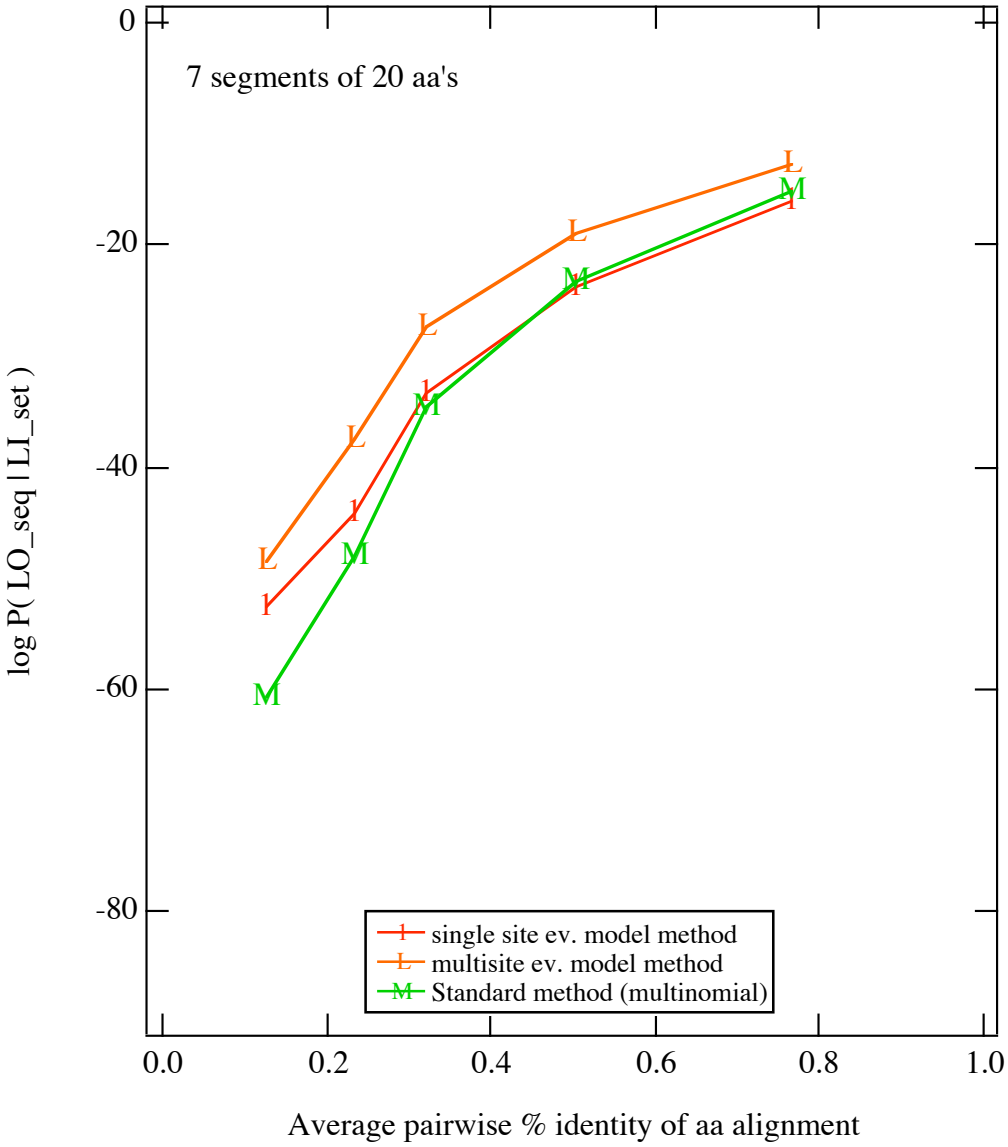

Figure 3.3. Relative performance of three prediction methods on alignments with different APPI's —segments of length 20 aa's. Data is simulated. (See text.) L=multisite cluster prediction method. 1=single site evolutionary model prediction method, M=multinomial prediction method.

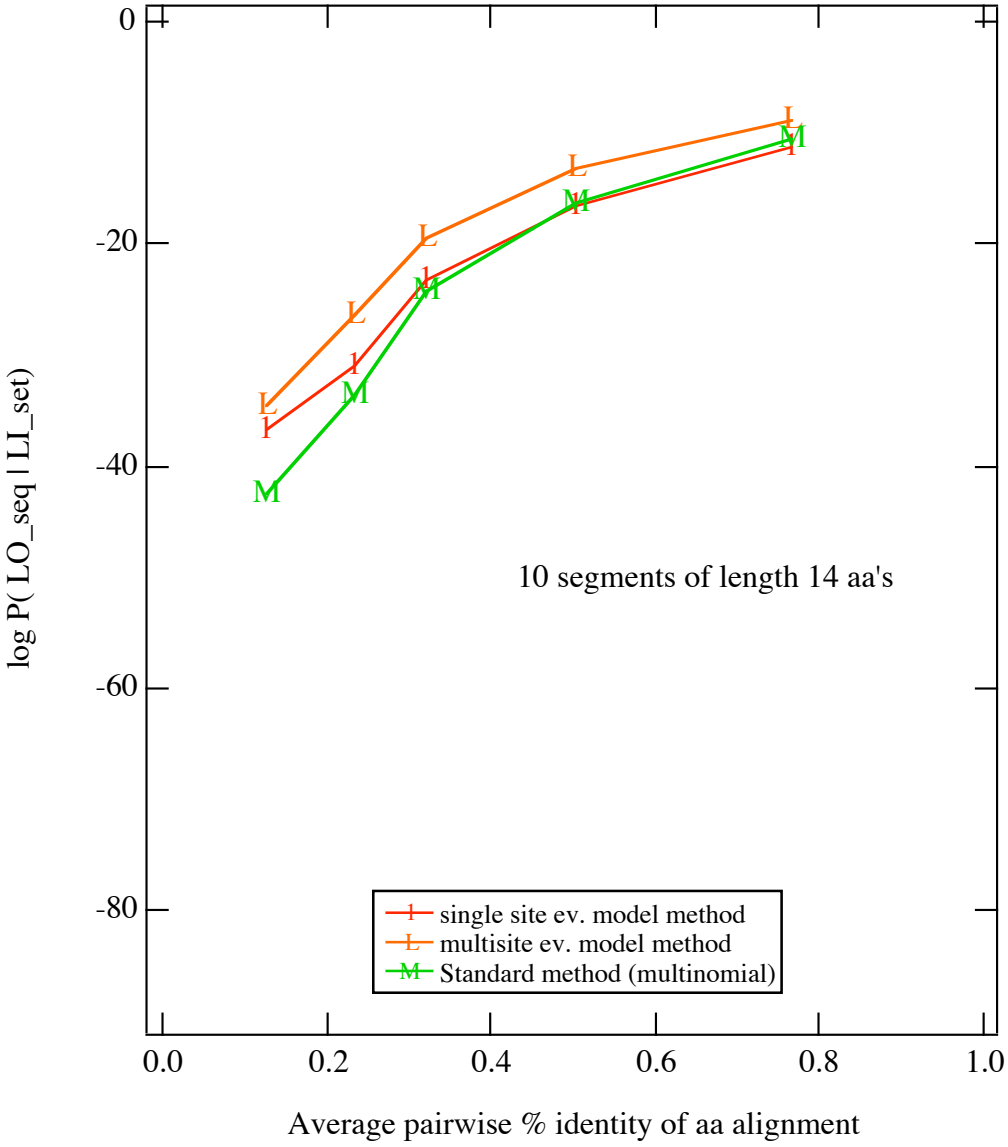

Figure 3.4. Relative performance of three prediction methods on alignments with different APPI's —segments of length 14 aa's. Data is simulated. (See text.) L=multisite cluster prediction method. 1=single site evolutionary model prediction method, M=multinomial prediction method.

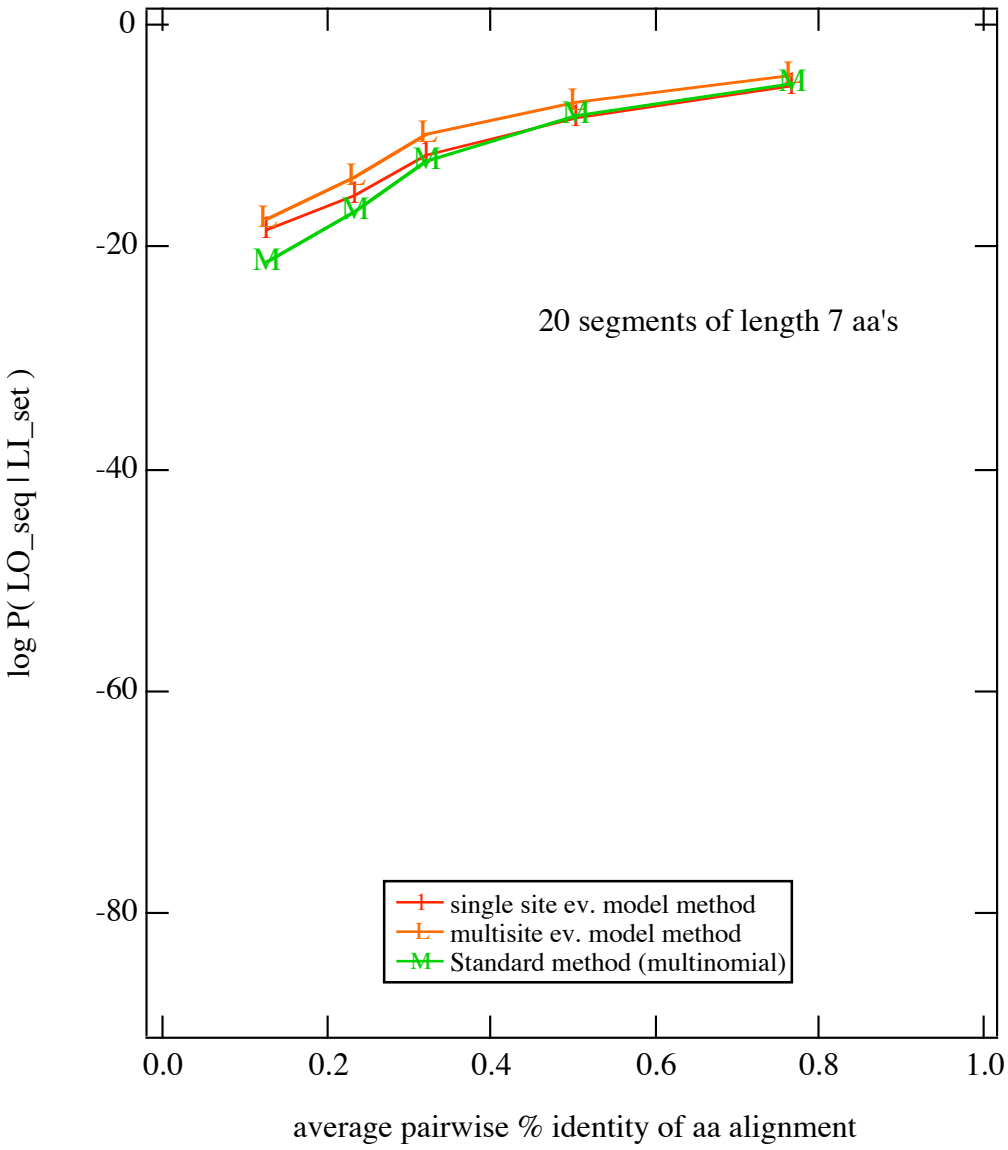

Figure 3.5. Relative performance of three prediction methods on alignments with different APPI's —segments of length 7 aa's. Data is simulated. (See text.) L=multisite cluster prediction method. 1=single site evolutionary model prediction method, M=multinomial prediction method.

Chapter 4

Comparison of pool construction methods

Chapters 2 and 3 compared methods of predicting a single sequence in a new, related species. But when researchers search for a gene in a new species, they make more than one prediction as to what that sequence is. Concretely, these predictions constitute a pool (mixture) of primers. The experimentalist puts a pool of primers into the PCR reaction, knowing that not all of these will bind, but hoping that one or some will bind well enough for the amplification reaction to work.

The research presented in Chapter 4 compares methods of constructing pools of primers. Another difference between this chapter, and Chapters 2 and 3, is that all predictions in this chapter are done in the codon realm, rather than the amino acid realm.

Performance measure for pool construction methods

As in the previous chapters, I will use a leave one out tests to compare the performance of the different methods. A leave one out test for a pool involved comparing the constructed pool with the one left-out sequence, and via that comparison calculate a performance measure (score) for the pool.

There is not a simple, obvious, best measure of performance of a pool. One possible measure would be to simply check for the presence or absence of the left-out sequence, and to give a pool a score of zero or one, based on the presence of that one sequence. But a pool might be missing a perfect match to a left-out sequence, but contain a number of oligos that mismatch the left-out sequence by just one nucleotide. Such a pool would probably work well in a PCR reaction. Similarly a pool could contain no oligos with zero mismatches, and no oligos with

one mismatch, but 50% of the oligos in this hypothetical pool could still have only 2 nucleotide mismatches. Again, this pool would probably amplify well, so it would not make sense to score it badly.

The performance measure that will be used in this chapter, to compare pools, is "fraction of pool that has zero, one, two, or three mismatches compared to the left-out sequence."

The "fraction of pool" performance measure is relative

This performance measure is intended to be a relative measure, not an absolute one. It is difficult to make an absolute mapping between the fraction of a pool with a few mismatches, and the success that pool will have in a PCR reaction. That is a chemistry question, and is complicated.

One unknown that makes such a mapping complicated is the question of what fraction of a pool must have few mismatches, for a PCR reaction to successfully amplify a gene. If 10% of a pool has zero, one, two, or three nucleotide mismatches, will amplification happen? Or is a 1 in 100 fraction, or a 1 in 1024 fraction, enough to allow amplification? I expect that in some cases, in some PCR reactions, pools with small fractions successfully amplify, and in other cases pools with small fractions do not amplify well. The outcomes of these chemical reactions depend on inputs other than sequence and fraction of pool with few mismatches. Parameters such as reaction conditions and amount of primer-primer binding will affect the chemical reaction.

Another reason this performance measure must be seen as relative is that the cutoff I use —3 or fewer mismatches— is ultimately arbitrary. There is no perfect or obvious line between "a few" nucleotide mismatches and "too many" nucleotide mismatches. If I used two or four mismatches as my cutoff, the values of the calculated performance measures would be different. But I think it reasonable to expect that in that situation (a different cutoff point) the

relative performance of different pool construction methods would be similar to the relative performances shown by measures using three mismatches as the cutoff.

This performance measure is a rough model of PCR chemistry

One consideration that this performance measure does not take into account is where in the primer the mismatches occurs. As mentioned above in Chapter 1, many researchers consider the 3' end of an oligo to be more important in determining the oligo's ability to prime DNA replication than nucleotides farther from the 3' end.

Another chemical complication, that this performance measure does not take into account, is primer-primer binding, or primer self-annealing. The former occurs when there are complementary regions in two primers in the pool. In this case the two "species" of primers would bind to each other, and not be available for binding to the target DNA in as high a concentration as one would expect. It would take a lot of computational effort to consider all of the possible combinations of primers within a pool, predict the chemistry of binding for each combination, decide how much primer-primer binding is "too much", and not include in the pool primers that would bind too much to other primers. Such calculations are beyond the scope of this dissertation. (This might be an area for future research).

Hairpin structures and other binding of nucleotides within a single primer can also make a primer unavailable for priming in the amplification reaction.

The conclusions I reach based on analysis using this performance measures will have to be confirmed with lab experiments.

Detailed description of the all-degenerate-primer pool construction method

I will describe this pool construction method, using the globin alignment as an example. For the 5-sequence alpha globin and beta globin alignment, there are 5 "leave-one-out" sets, each containing 4 sequences. The following process was done for each of the leave-one-out sets of sequences. So below, "alignment" means an alignment of n-1 sequences.

The alignment was broken into segments of 7 aa's. At each site, the most prevalent amino acid in the aligned sequences was chosen. If two aa's were equally prevalent, one was chosen at random to be the "most prevalent". All the codons in the genetic code that might code for this most prevalent aa were included at this site in the oligonucleotides in the primer pool. The pool consisted of every combination of codons that could code for this amino acid sequence. For example, if the 7 most prevalent amino acids in the 7 amino acid long segment each had 2 codons for it in the genetic code, then the pool would consist of $2^7 = 128$ oligonucleotides.

Detailed description of the "whole tree" pool construction method

This method was performed on the same 7 amino acid long segments (of the alignments) used throughout this dissertation.

This method used the multisite evolutionary model based method described above in Chapter 2.

Scoring

A number of sequences from the sequence space were scored for probability of being the new, related sequence, based on Equation 2.11

$$P\left(\text{root} = \text{LO\_seq}_{\text{gen,FL}} \middle| X_{\text{FL}}\right) = \frac{\prod_{\text{all } A} \prod_{\text{all } \Pi} \prod_{S=1}^{k} \left(\text{P}\left(X_S \middle| \text{root} = \text{LO\_seq}_S, A\right)\text{P}\left(\text{root} = \text{LO\_seq}_S\right)\right) P(A)}{\prod_{S=1}^{k} \prod_{\text{all } \Pi} \left\{P\left(X_S \middle| \Pi_S\right) P\left(\Pi_S\right)\right\}}$$

Eq. 2.11

The sequences with the highest scores were included in the pool.

Space searched

The space of 7-codon long sequences, that was identical to one of the the sequences in the alignment (of n-1 sequences, i.e. alignments with one left out sequence) except for zero, one, two, or three mismatches, on the codon level, were scored. (For G3PD just 0,1, or 2 mismatch space was searched.) That is $61^3 \binom{7}{3} = 226{,}981 \cdot 35 = 7{,}944{,}335$ sequences out of the entire

$61^7 = 3.14 \times 10^{12}$ sequence space were searched. This many sequences were scored for each of the n leave-one-out sets of sequences (for an alignment of n sequences) for each 7 aa long segment in each alignment. ("Scored" here means the performance measure in Equation 2.11 was calculated.) This space was chosen as the space most likely to contain the new, related sequence. And also the highest scoring sequences would be in this space. It would not make sense to score sequences with, for example, 6 or 7 mismatches from any of the sequences in the known alignment. Those sequences would obviously score less well, by the scoring methods used, and so would not be included in the pool.

Pool size

The pool size —the number of oligonucleotide sequences in the pool— was based on the all degenerate primer pool size, so the two methods could be better compared. If the all degenerate primer pool size was less than 1024, that number was also set as the size of the

pool constructed by the "whole tree" score method. If the all degenerate pool was greater than 1024, the pool size of the "whole-tree" method pool was set to equal 1024.

Detailed description of the "one subpool per attachment point" pool construction method

(I will refer to this pool construction method as the "subpool" method, for short.)

In the "whole-tree" pool construction method, there is one pool of primers. In the "subpool" pool construction method, there are considered to be q smaller pools of primers, where q equals the number of attachment points in the relevant tree. One subpool corresponds to each attachment point on the tree. The maximum size of each subpool is set equal to

$$\frac{\#\text{ of primers in corresponding "whole-tree" pool}}{q}.$$

In the "subpool" pool construction method, the same sequence space was searched for primer sequences, as was searched for the "whole-tree" pool construction method. But the criteria for inclusion in the pool were different. In the "whole-tree" pool construction method, primers with, simply, the highest scores, are chosen from the entire space searched.

As described in the derivation in Chapter 2, the multisite evolutionary model based prediction method calculates the probability $P(\square_A, X|A)$. (Expressed that way in the numerator of the RHS of Equation 1.6, and expanded into

$$\prod_{\text{all } \square} \left[\prod_{S=1}^{k} \left(\mathrm{P}\left(X_S | \text{root} = \text{LO\_seq}_S, A \right) \mathrm{P}\left(\text{root} = \text{LO\_seq}_S \right) \right) \right]$$

in equation 2.11.) In the multisite method, this probability is calculated once for each attachment point: under the assumption that that attachment point is the true attachment point. These probabilities are averaged over all the attachment points.

In the "subpool" pool construction method, $\mathrm{P}(\square_A, X|A)$ for a particular sequence $\square_A$ is calculated for each attachment point A. Then these probabilities are compared. The sequence $\square_A$ is declared to be a candidate to be in the "subpool" of the attachment point with the maximum probability $\mathrm{P}(\square_A, X|A)$ for that sequence $\square_A$. If $\square_A$'s value of $\mathrm{P}(\square_A, X|A)$ is greater than the $\mathrm{P}(\square_A, X|A)$ of the primer with the smallest $\mathrm{P}(\square_A, X|A)$ in that subpool, then $\square_A$ is added to the subpool. (Adding a sequence to the subpool will cause another sequence to be bumped out of the subpool, if the subpool is at its maximum capacity.) If $\square_A$'s value of $\mathrm{P}(\square_A, X|A)$ is less than that cutoff point, then it is not included in the subpool.

Thus the sequence space is searched and the subpools are constructed. The resultant subpools are then combined. This combined pool is then scored for the "fraction with zero, 1, 2, or 3 nucleotide mismatches" performance measure, just as the whole-tree pool or the all degenerate primer pool are scored.

Results

The alpha and beta globin codon tree is shown in Figure 4.1.

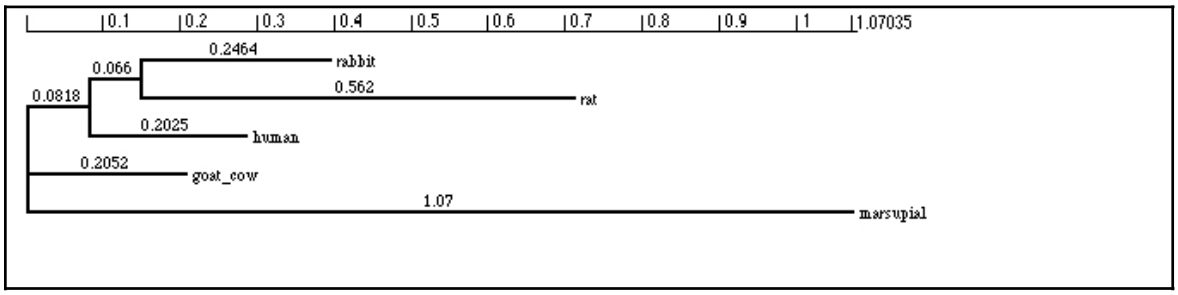

Figure 4.1. alpha and beta globin phylogenetic tree with codon branchlengths.

Branch lengths are in expected number of nucleotide replacements per codon. Fit by PAML.

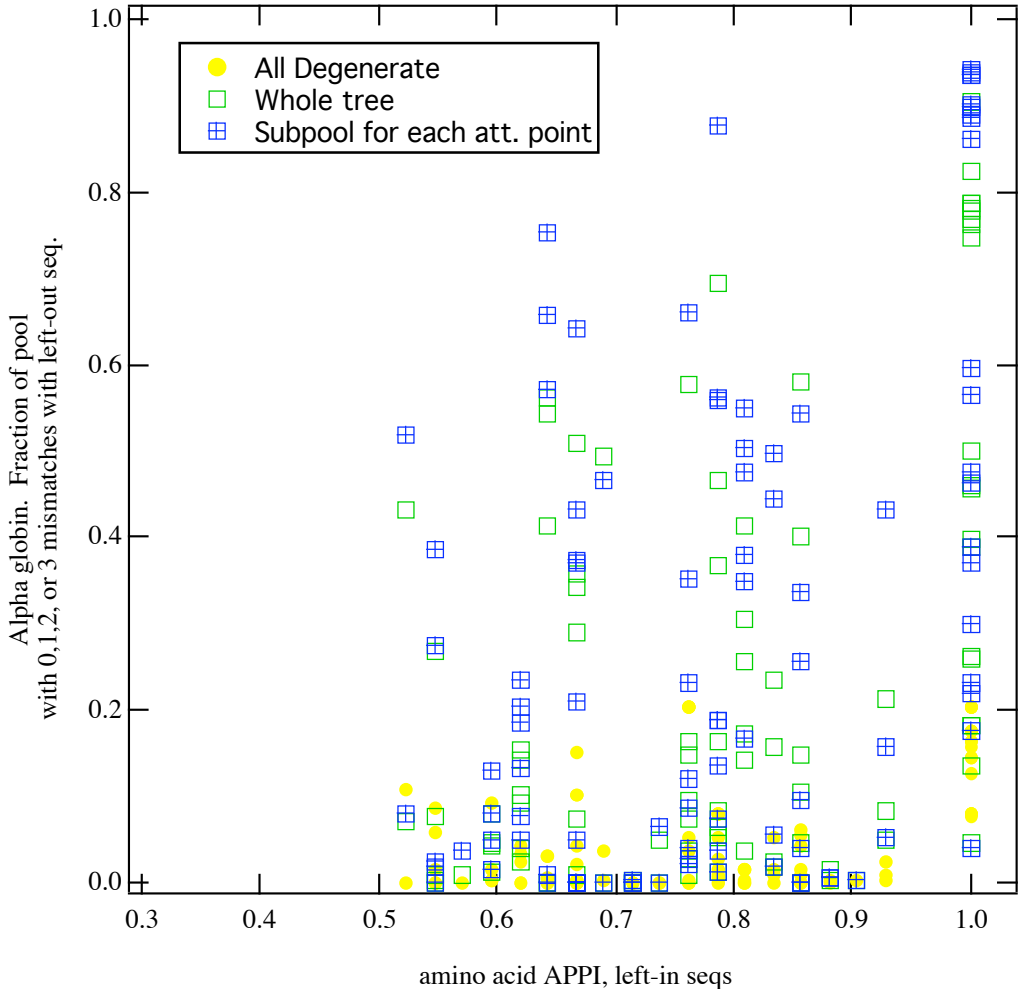

Figure 4.2. Comparison of 3 pool construction methods on the alpha globin alignment. The amino acid APPI is calculated based only on the left-in set of sequences.

Comparison of the three pool construction methods for the alpha globin gene

Figure 4.2 shows how each of the three pool construction methods performed on each left-in subset of sequences, for each 7-aa segment of the alignment. There are three markers on the graph for each of the 20 7-aa segments of the alignments.

This Figure can be used to address the question of if the "whole-tree" pool construction method is better, or if the "subpools" pool construction method is better. The answer that the graph gives is that there is not evidence in the data to indicate that either is superior to the other. Their distributions of performance measures look about the same.

The yellow dots in Figure 4.2 show the scores of the "all-degenerate" pool construction method pools. Based just on observing Figure 4.2, they clearly have lower performance measure values, on average, than the pools constructed by the evolutionary model based methods. But these low scores are due probably to the fact that these pools are all highly degenerate. That is, there are many more types of oligonucleotides in each pool than Sells and Chernoff recommend. These high pool sizes can be seen in Figure 4.3.

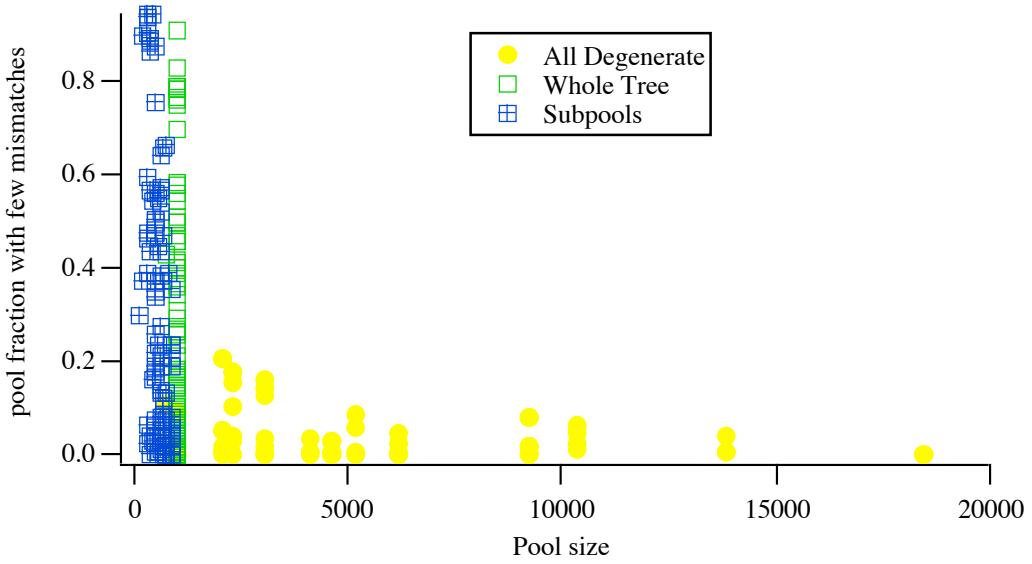

Figure 4.3. alpha globin pool performance measure vs pool size, for the three pool construction methods.

If a researcher were working with this alignment, he would adjust the pools, to leave out some codons, in order to be able to do the standard primer pool design method. But making adjustments like that require judgment calls, so the algorithm I use to represent the standard primer pool design method does not make those adjustments. The result is that the algorithm designs pools that do not meet the rules set out by Sells and Chernoff. So we can not use these data to do comparisons against the other pool construction methods.

Comparison of the 3 pool construction methods for beta globin

The tree for beta globin is the same as for alpha globin. The tree was fit with PAML using the two sequences concatenated. So refer to Figure 4.1 to see the codon tree for beta globin.

The goal of graphing the data in Figure 4.4 is to compare the performance of the standard primer pool construction method with the performance of the two evolutionary model based pool construction methods.

As mentioned above, because of degeneracy issues, many of the automatically constructed primer pools have too many (according to Sells and Chernoff) types of primers in them. The scores of these pools are indicated by yellow dots in the figures. The performance measures of the pools that do meet the Sells and Chernoff criteria are plotted in red, and a red dotted line is drawn indicating the score (performance measure). So this dashed line represents the performance of the standard primer pool design method. (These dashed lines are equivalent to the vertical dashed lines on Figure 4.5.)

First consider the performance of the "whole-tree" method (green squares) compared to the standard method (red horizontal lines). While all of the whole tree method points are greater than the lower red line, the points about evenly straddle above and below the upper red line. Basically, no clear pattern or relative performance trend is evident, between the "whole-tree"

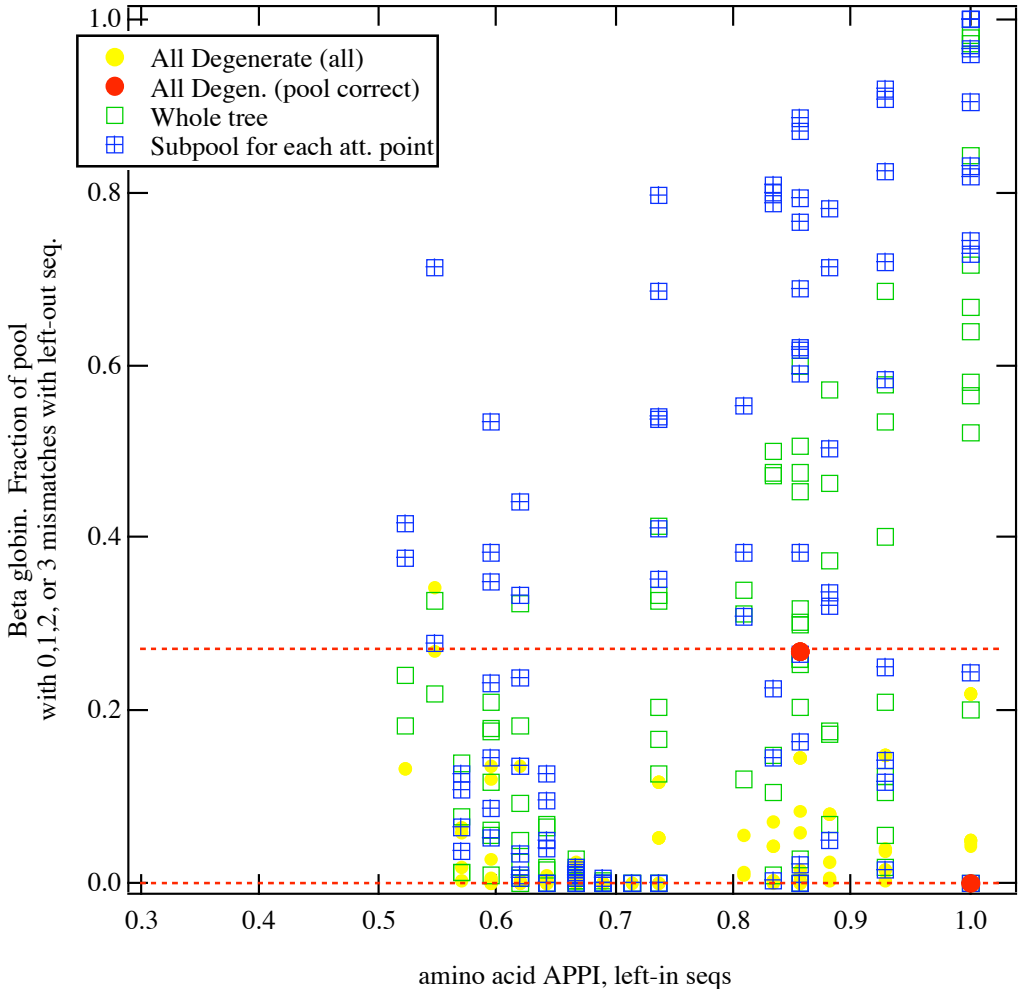

Figure 4.4. Comparison of three pool construction methods. Beta globin alignment. The horizontal red lines are the "pool fraction" performance measures for the pools designed by the "all degenerate" pool construction method that meet the Sells and Chernoff criteria for pools (see text).

pool construction method pools and the standard method pools. This process of simple graphical inspection to notice trends in the data relies on a clear effect being shouted out by the data. A clear trend would have to be strong, for example if all the green squares lay above both red lines. There is no trend like that in this data.

The same absence of clear pattern can be noted in the relation of the "subpools" data (blue squares) and the standard method data.

Figure 4.5 shows the same data as in Figure 4.4 in a different way. Figure 4.5 is a graph of three histograms, showing the distributions of the "whole tree" performance measures (green squares on Figure 4.4), the "subpool" performance measures (blue squares), and the standard method performance measures (yellow dots). Each histogram contains a total of 100 points (one for each combination of segment and left-out sequence in the beta globin alignment.) Figure 4.5 also has a vertical dashed line, equivalent to the horizontal dashed line in Figure 4.4, showing the pool fraction performance measure for the "all degenerate" designed pool that meets the standard primer design method criteria.

Figure 4.5 is useful because in it one can see that the distributions of the two evolutionary model based pool construction methods are quite similar. One can conclude from it that the data shows no evidence that one of these methods is superior to the other.

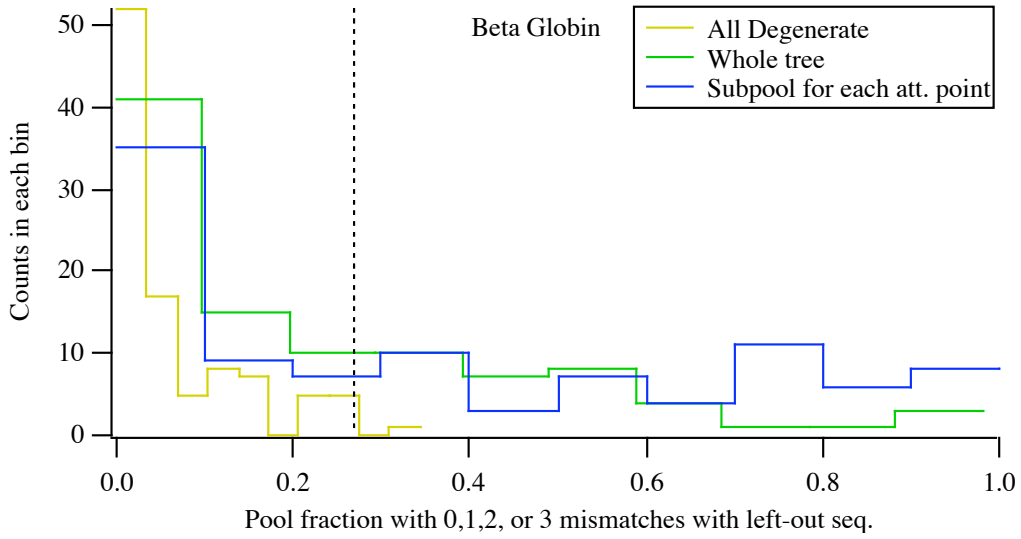

Figure 4.5. Three histograms for the beta globin data.

Compare pool construction methods using the G3PD alignment data

The analysis in this chapter is done using codon evolutionary models. So the G3PD phylogenetic tree has the same topology shown in Figure 2.1, but branch lengths are different. The branch lengths are now in units of expected nucleotide substitutions per codon. Figure 4.6 shows the correct branch lengths for the Chapter 4 analysis.

((((O68075: 1.98438, (G3PC_ALCEU: 0.61338, G3P_PSEAE: 0.93723): 0.11075): 0.63644, (G3P2_RHOSH: 0.81243, G3P_XANFL: 0.52343): 0.67738): 0.30750, G3P_ZYMMO: 1.72735): 1.55838, (G3P_ARCFU: 2.11991, ((G3P_PYRWO: 1.51403, (((G3P_METBR: 0.71299, G3P_METTH: 0.93675): 0.55817, G3P_METFE: 0.79636): 0.81618, G3P_METJA: 1.67450): 0.33596): 0.49739, G3P_SULSO: 2.61710): 0.16558): 9.90060, ((G3P_USTMA: 0.43124, (G3P_PHARH: 0.56698, (G3P_LYOSH: 0.73343, G3P_PHACH: 0.33290): 0.23387): 0.13856): 0.14735, (((G3P_COCHE: 0.16232, G3P_ERYGR: 1.02113): 0.16709, G3P1_TRIKO: 1.24626): 0.13456, (G3P_MONAN: 0.43364, G3P2_TRIKO: 0.39220): 0.00000): 0.10937): 0.56635);

Figure 4.6. Newick representation of the G3PD tree with branch lengths in units of expected nucleotide substitutions per codon.

The axes on Figure 4.7 have ranges similar to Figures 4.2 and 4.4 (the analogous graphs for the globin genes) to allow comparison with those graphs. From this comparison one can see that the distribution of pool fraction performance measures is more tightly packed, and has a smaller average, than the analogous distribution for the globin genes (above). The G3PD distribution is very similar to the analogous rpL20 distribution (not shown).

Figure 4.8 is three histograms showing the distributions of the data points in Figure 4.7. One can see from inspection of Figures 4.7 and 4.8 that the “whole tree” pool construction method and the “subpools” pool construction method again created pools whose distributions of performance measures are very similar. And both of the evolutionary model based pool construction methods have performances in the same range, and with similar

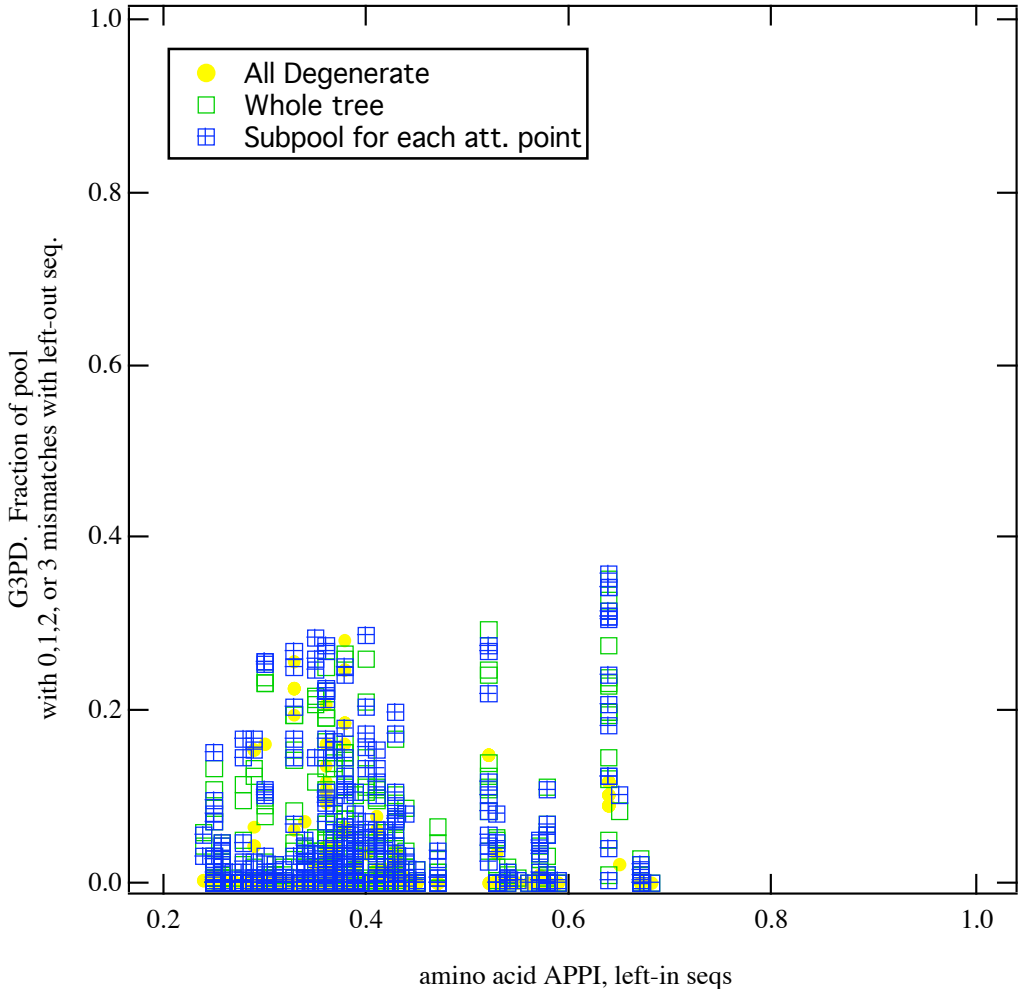

Figure 4.7. G3PD pool performance measures vs. APPI.

distributions to, the performances of the pools designed by the standard primer pool design method.

The rpL20 tree and model for codons

The topology of the r. p. L20 tree used in Chapter 4, the tree for codon models, was taken from the amino acid based tree. The other codon-realm evolutionary model parameters

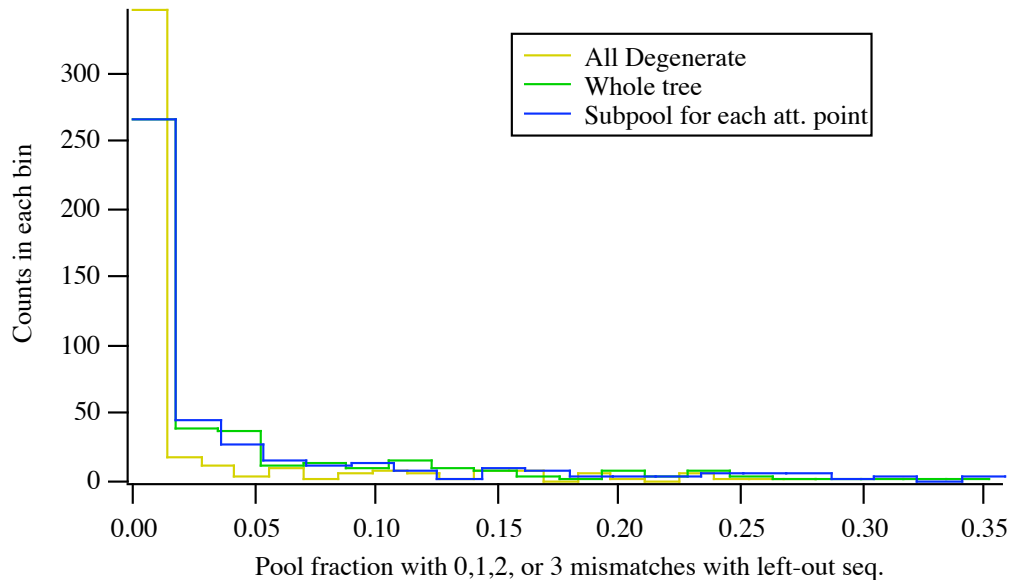

Figure 4.8. 3 Histograms of "pool fraction" performance measure for G3PD data.

(codon stationary probabilities, branch lengths, and Yang's kappa and "dN/dS ratio") were calculated by PAML and then used in my likelihood calculations. This PAML fit was done with one "dN/dS" value for all the branches of the tree. Figure 4.9 shows the branchlengths PAML arrived at in this fit.

(RL20_BACST: 1.20685, (RK20_CYAPA: 0.84316, (RK20_CHLRE: 1.12600, (RK20_TOBAC: 0.00000, (RK20_EPIVI: 0.63489, (RK20_MAIZE: 0.56267, (RK20_PINTH: 0.23446, (RK20_MARPO: 0.37430, (RK20_EUGGR: 1.56263, RK20_ASTLO: 1.43981): 2.68826): 0.71787): 0.82321): 0.18438): 0.00000): 2.70042): 0.29249): 0.93973, RL20_ECOLI: 1.11272);

Figure 4.9. Newick representation of the phylogenetic tree for the ribosomal protein L20 alignment. This shows the relationship of aligned codon sequences. Branch lengths are in units of expected nucleotide substitutions per codon.

The plot of performance measures for pools created by the three pool construction methods from the ribosomal protein L20 sequences, vs. APPI, is not shown because the conclusions one can draw from it are similar to the conclusions reached from Figure 4.7 the analogous graph for the G3PD data. The ribosomal protein L20 sequences graph gives no evidence that

either the "whole tree" pool construction method or the "subpools" pool construction method creates pools that perform better than the other method's pools. Also, as was the case for the G3PD results, both evolutionary model based pool construction methods have performances in the same range, and with similar distributions to, the performances of the standard method pools.

Does pool size affect the "fraction of pool with few mismatches" performance measure?

In this section the word "size" means the number of different types, or sequences, of oligonucleotides in a pool. It does not mean the number of oligonucleotide molecules in a PCR reaction.

The Chapter 4 data can address this question: Does a smaller pool size increase performance values, as measured by the "pool fraction" performance measure? Figures 4.10, 4.11, 4.12 and 4.13 plot performance (on the y axis) against pool size (on the x axis) for the alpha globin, beta globin, G3PD, and r.p. L20 alignments, respectively.

The Figure 4.13 shows all the data points in the r.p. L20 data set. . Figures 4.10, 4.11, and 4.12 cut off the pool size axis at a maximum of 1024, so many points in those data sets, mostly for the "all degenerate" method designed pools, fall off of those graphs, to the right. Figure 4.12 is included to show a wider picture of the data. (As does Figure 4.3 earlier in the chapter.)

One conclusion about pool size that Figure 4.13 allows one to reach, is that pools with very large sizes (greater than 2000) perform worse than pools smaller than that amount. This result is consistent with the Sells and Chernoff rule to not make primer pools of a size larger

than 1024.

Within the smaller range of data points, shown in Figures 4.0, 4.11, and 4.12, there seems to be some population of points with low pool size (size around 200-300) and high performance measures, in the beta globin data. This population is not apparent in the G3PD and r. p. L20 data.

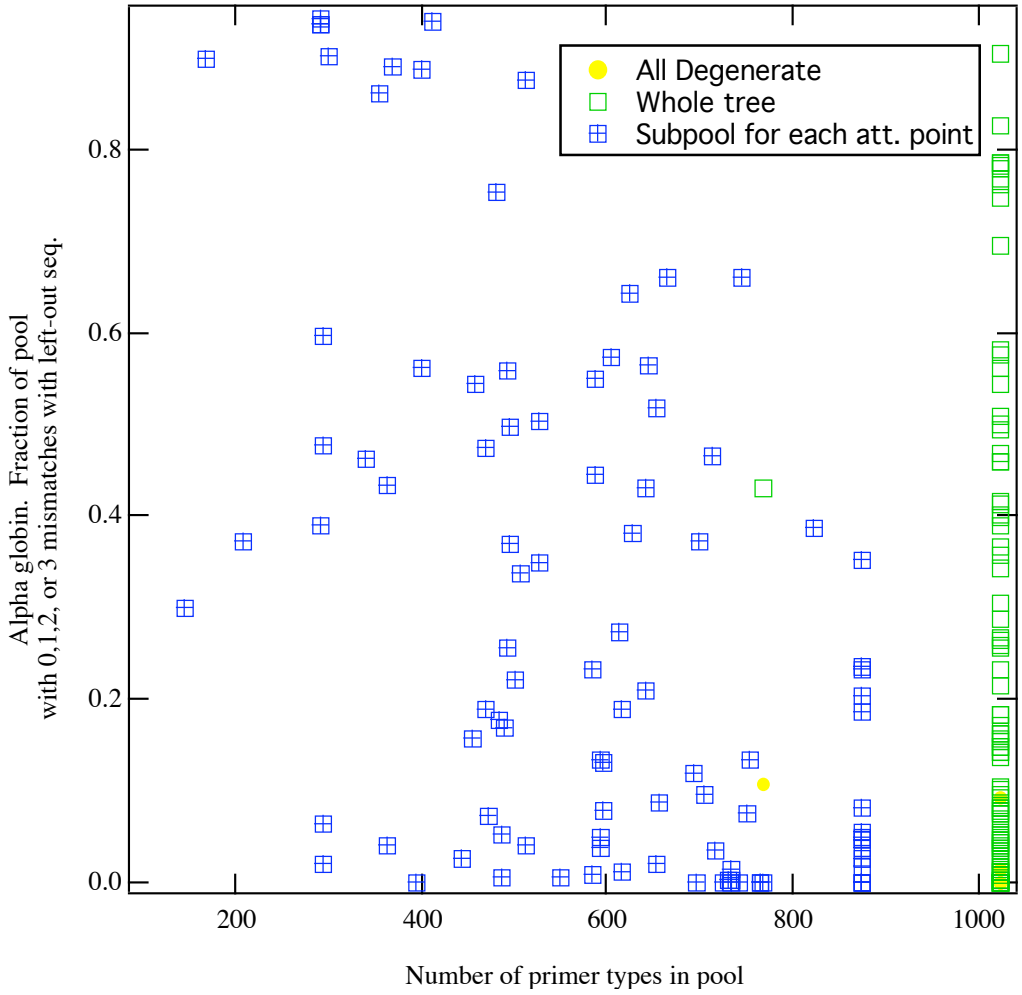

Figure 4.10. Pool fraction vs. Pool size, for alpha globin gene.

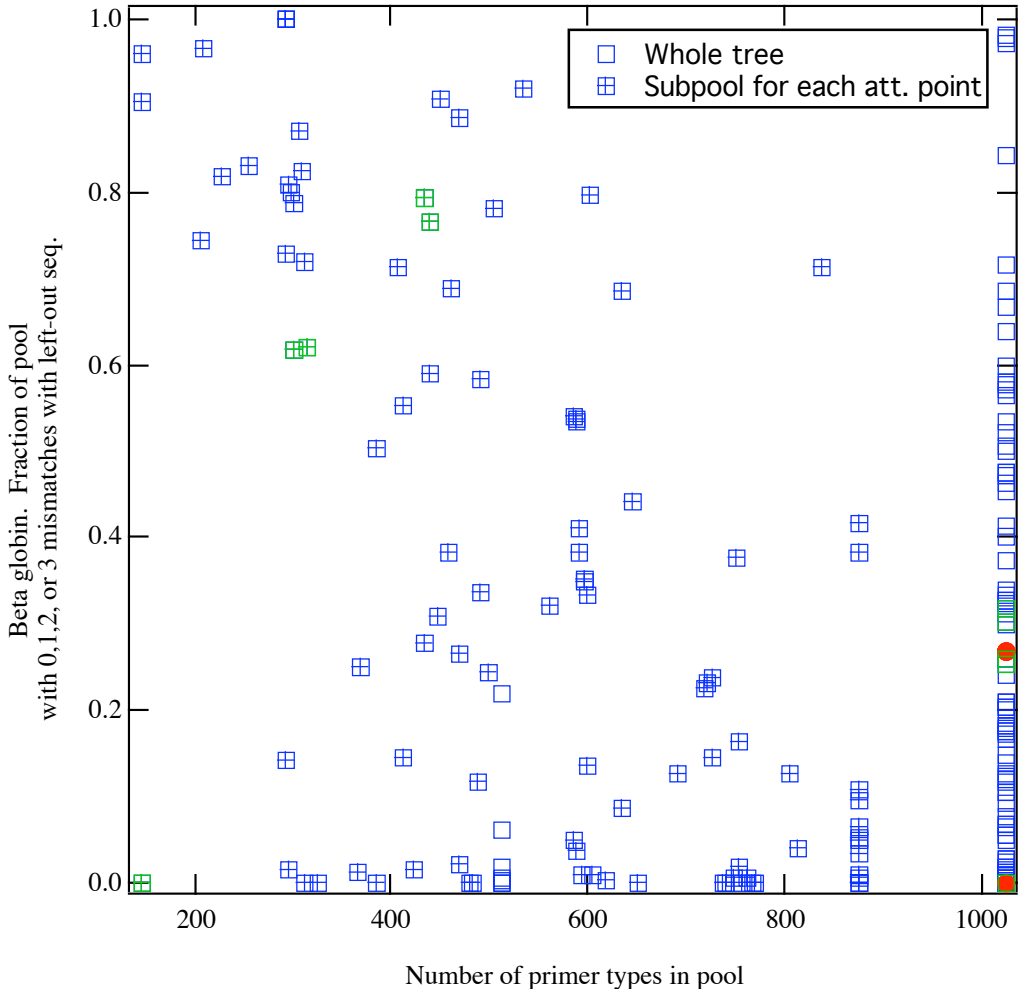

Figure 4.11 Influence of pool size on performance in beta globin. The fraction of each pool with 3 or fewer mismatches on the nucleotide level, between the oligos in the pool and the left out sequence, is graphed vs pool size.

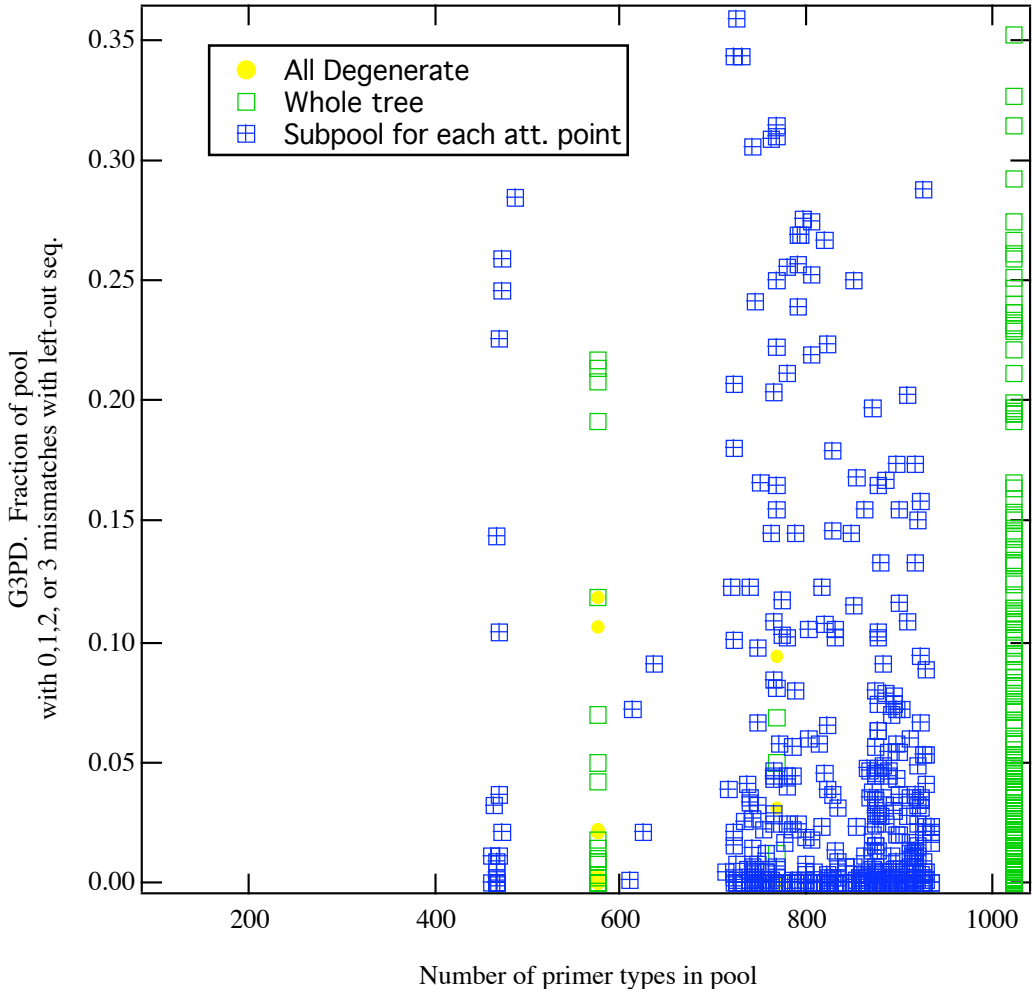

Figure 4.12. Pool size and performance, G3PD data.

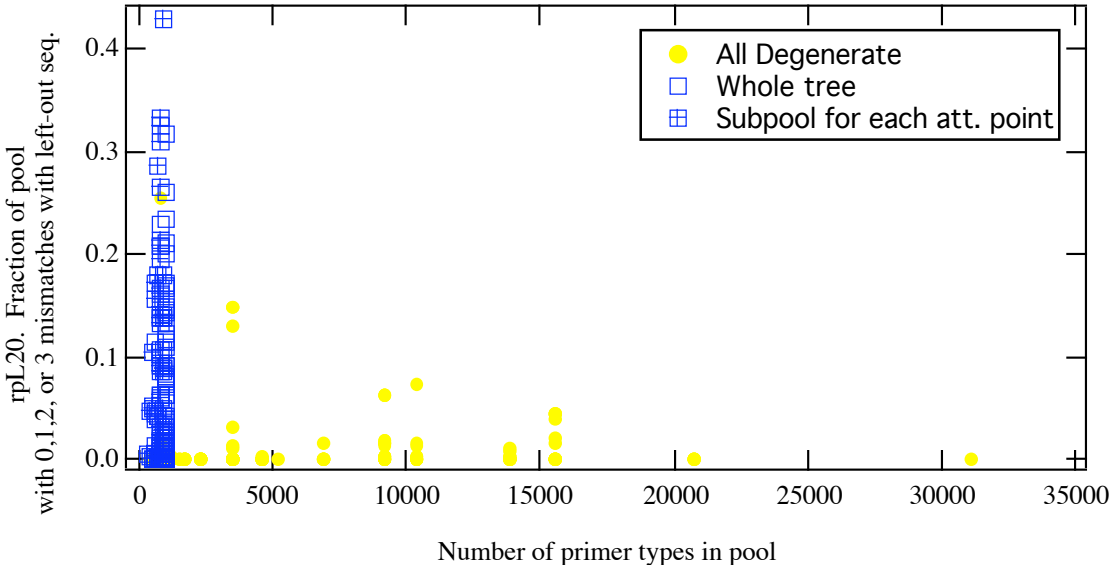

Figure 4.13. Pool size and performance, ribosomal protein L50 data.

Chapter 5

Conclusions from this research

Chapter 2

In Chapter 2, I employed three different ways of comparing performance measures calculated by the three prediction methods. Those three ways of comparing were: comparing the groups of average P(LO_seq|LI_set)'s with the Wilcoxon-Nemenyi-McDonald-Thompson test; comparing the number of cases in which the P(LO_seq|LI_set)'s for one method were greater than the corresponding P(LO_seq|LI_set)'s of the other method (by inspection of the distribution of differences, and by sign test), and comparing the number of cases in which the P(LO_seq|LI_set) for one method was more that 0.01 greater than the corresponding P(LO_seq|LI_set) for the other method.

The pair of prediction methods that showed the strongest and most consistent differences between them, by these means of comparison, was the multisite evolutionary model based method and the standard primer design method (represented by the multinomial distribution). No results of any of the comparisons (see Table 2.5 and the section following it) of these two prediction methods indicated that the standard method had better performance measures. Although in some of the cases there was no significant difference between the two methods. The Wilcoxon-Nemenyi-McDonald-Thompson procedure found a difference between the two methods for all three alignments. These consistent results are evidence that the multisite evolutionary model method predicts better than the standard primer design method, as represented by the multinomial distribution.

For the other two pairs of methods (standard method vs. single site evolutionary model method results summarized in Table 2.4 and the section following it, multisite method vs. single site method summarized in Table 2.6 and the section following it) some results indicated one way and some results indicated another way. For both of these comparisons,

the preponderance of the evidence pointed in the direction of the evolutionary model method as being better than the standard method, and the multisite evolutionary model method as being better than the single site method. But some of the cases contradicted this direction, making it difficult to reach the conclusion strongly. Repeating these tests on a larger number of alignments might allow one to clearly distinguish a difference in performance.

An interpretation of the three methods' relative performances that is consistent with the results of Chapter 2 is that the single site method is intermediate in performance, between the multinomial method and the multisite evolutionary model based method, but that its performance can not be strongly distinguished from either based on the data from three alignments.

The result showing the multisite evolutionary model method's clear advantage over the standard method, indicates, perhaps, that the most promising direction for future research into evolutionary model methods will be further development of methods that identify correlations between occurrence of states at different sites. This interpretation would direct mathematicians' future efforts toward improving multisite methods, and lead one to think that evolutionary models that apply the assumption of independent evolution between sites in a way that misses this information to have less potential.

A possible application of evolutionary models: improve the CODEHOP method

The comparisons in Chapter 2 are all about predicting a single sequence (i.e. not constructing a pool of sequences, as in Chapter 4). Researchers always screens with pools of primers, never a single primer sequence. So the comparisons in Chapter 2 are useful because they help one decide if a prediction method is a better one to use to construct a pool. But the methods would not be used in the lab exactly how they are used in Chapter 2 —to design just one primer. There is an exception to that rule though. The CODEHOP primer design procedure of Rose et al. (1998), described in detail in Chapter 1, uses a single primer

sequence as one part of every oligonucleotide in the pools it designs. Rose et al. want to use just one sequence for this part of their primers because doing so allows them to make smaller-sized pools. The CODEHOP algorithm determines this single sequence simply by using the most common codon (according to a codon usage table specified by the user) coding for the most common amino acid at that position in the amino acid sequence. That algorithm is similar to a multinomial method used in the codon realm. The results from Chapter 2 indicate that, when the comparison is done in the amino acid realm, the multisite evolutionary model method predicts a single sequence better than the multinomial method does. If the multisite method were shown to have the same performance advantage in the codon realm, then the multisite evolutionary model method would be a preferable method for designing the 5', non-degenerate part of the CODEHOP primers. The CODEHOP program is a widely used for primer design, as shown by the count of citations reported in Chapter 1. So this improvement, if carried out, could benefit many researchers.

Chapter 3 —cluster information

The prediction method described and tested in Chapter 3, that makes use of cluster information in a phylogenetic tree, performed better than the standard primer design method.

I would expect models that make use of information in clusters within a tree to have potential for even better prediction of related sequences. Realizing this potential would involve solving some interesting problems, such as determining the best algorithm for defining clusters within a tree.

This computational result will, like all the results of this dissertation, have to be followed up on in the laboratory. The lab is the ultimate test to confirm that method is useful for screening for homologous genes.

Chapter 3 —results over varying branch lengths

The largely consistent results shown in Figures 3.2, 3.3, 3.4, and 3.5 indicate that the results seen in this dissertation are probably applicable to trees with average branch lengths (or, similarly, alignments with different APPI's) different than the three alignments studied in this dissertation. This is an important result because I want to be able to claim that the observations I make in these three alignments are characteristic of what one would see in many alignments. This results shown in those figures supports the generality of my observations.

An interesting result seen in this section of Chapter 3 is the apparent switch, depending on APPI, in performance between the standard method (represented by the multinomial distribution) and the single site evolutionary model method. Although no formal statistical comparison was done, the multinomial method has higher performance measures than the single site evolutionary model method when tested on alignments with higher (above 0.4) APPI's. But in alignments with low (less than 0.4) APPI's, the single site evolutionary model based method has higher performance measures. If this result holds true in future research, for many different data sets, a rule of thumb a researcher could apply when screening for genes would be: use the standard method for some levels of identity of sequences, use evolutionary methods for other levels. (Although in these graphs the multisite evolutionary model method always has higher performance measures than the other two prediction methods, .so the best rule of thumb would be to always use the multisite method.)

Chapter 4 —pool construction— conclusions

The Chapter 4 data does not indicate that either the "whole-tree" multisite evolutionary model based pool construction method or the "one subpool per attachment point" multisite evolutionary model based pool construction method is better (based on the fraction of pool with few mismatches performance measure) than the standard pool construction method.

Nor does the data indicate which of the evolutionary model based methods is the better of the two. I am surprised by this result. I would have predicted that the subpool method would have higher performance measures than the whole-tree method. The subpool method has the advantage of including primers that are the highest scoring sequences for their branch of the tree, not the entire tree. I would have guessed that pools composed of these primers would cover the sequence space (that the target sequence lives in) better than the "whole-tree" method pools do, and thus result in a better "pool fraction with few mismatches" performance.

The Chapter 4 data shows that pools with sizes above 1024 do not perform as well as pools with sizes below that threshold. This result is a confirmation that the size chosen by Sells and Chernoff is probably a good cutoff. The data does not indicate that, among pools with sizes smaller than the size 1024, there is a correlation of smaller pools sizes and better performances. There is a hint of such a correlation in the beta globin data, but not in the other alignments' data.

These Chapter 4 results are consistent across all three alignments studied, even though the performance measure distributions varied a good deal between the different alignments.

Directions for future research

The most important direction for future research is to test some of these methods in the lab. That is, to design primers using an evolutionary model based method, and to use those primers to screen of a DNA library. (And to compare the results of this screen with a screen of the same DNA library by a standard method designed pool of primers.) This dissertation is aimed at helping experimentalists find related genes. So the ultimate test of the methods presented here is to see if the methods really do help in the lab.

Possible improvement to the method: codon bias.

If a researcher had access to information about the codon preferences of a particular species, those could be incorporated into the calculations of probabilities of related sequences. Codon preference information would be useful if the researcher is looking for a homologous gene in one specific species, not several species. Primers designed in this way would be targeted to work in that specific species, and would be, perhaps, less likely to work on a different species.

Improvements to the evolutionary methods

A way to improve the method would be to use models that have site-specific and branch specific substitution matrices. These models would probably assign more accurate probabilities to possible new sequences. There would be an additional computational cost to using these models, but that is not a major barrier.

Use information from other alignments

A lot of uncertainty is added to the calculations that the evolutionary models carry out by the absence of information about the attachment point and branch length to the generalized new species. If it were possible to identify more precisely where the correct attachment point is, and more precisely the branch length from the attachment point to the new species, one could probably make better predictions. It would be possible to make estimates about these model parameters if the target species is known, and if one made use of information from other alignments that contain a sequence from the target species. (The given-find problem that this dissertation addresses, posed in Chapter 1, describes a situation in which one does not have a particular target species in mind, or wants to look for related genes in multiple species. So the question here would be different.) Using the other alignment, one could infer a species tree that would include the target species. This tree would give one a good first estimate of the

attachment point and branch length. Such a method will eliminate the uncertainty introduced into predictions by integrating over different possible attachment points.

References

Ausubel, Frederick M., Roger Brent, Robert E. Kingston, David D. Moore, J.G. Seidman, John A. Smith, Kevin Struhl, editors. 2002. "Short Protocols in Molecular Biology. 5th Edition." John Wiley and Sons.

Bartl, Simona. 1997. "Amplification using degenerate primers with multiple inosines to isolate genes with minimal sequence similarity." Methods in Molecular Biology, Volume 67: from Molecular Cloning to Genetic Engineering. Bruce Alan White, Editor. Humana Press. pp. 451-457.

Bassett, Carole L., Timothy S. Artlip, Ann M. Callahan. 2002. "Characterization of the peach homologue of the ethylene receptor, PpETR1, reveals some unusual features regarding transcript processing." Planta. 215:697-688.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. 2002. "The Pfam Protein Families Database." Nucleic Acids Research 30:276-280

Chen, M. H. C., G. H. Liin, H. Y. Gong, C. Y. Lee, C. Y. Chang, T. T. Chen, J. L. Wu. 1998. "Cloning and characterization of insulin-like growth factor I cDNA from black seabream (*Acanthopagrus schlegeli*)" Zoological Studies. 37:213-221.

Dayhoff, M.O., R.M. Schwartz, and BC Orcutt. 1978. "A model of evolutionary change in proteins." p345. Atlas of Protein Sequence and Structure 1978.

Donehower, Lawrence A., Robert C. Bohannon, richard J. Ford, and Richard A. Gibbs. 1990. "The use of primers from highly conserved *pol* regions to identify uncharacterized retroviruses by the polymerase chain reaction." Journal of Virological Methods. 28:33-46.

Dopazo, Joaquín, and Francisco Sobrino. 1993. "A computer program for the design of PCR primers for diagnosis of highly variable genomes." Journal of Virological Methods. 41:157-166.

Felsenstein, J. (1981) "Evolutionary trees from DNA sequences: a maximum likelihood approach." J. Mol. Evol., 17:368-376

Felsenstein, J. 1986. "The Newick tree format". http://evolution.genetics.washington.edu/phylip/newicktree.html

Gibbs, Adrian, John Armstrong, Anne M. Mackenzie, Georg F. Weiller. 1998. "The GPRIME package: computer programs for identifying the best regions of aligned genes to target in nucleic acid hybridisation-based diagnostic tests, and their use with plant viruses." Journal of Virological Methods. 74:67-76.

Gorelenkov, V., A. Antipov, S. Lejnine, N. Daraselia, and A. Yuryev. 2001. "Set of novel tools for PCR primer design." Biotechniques 31:1326-1330.

Hauser, John R. 1996. "Handling floating-point exceptions in numeric programs." ACM Transactions on Programming Languages and Systems. 18:139-174.

Hollander, Myles, and Douglas A. Wolfe. 1999. "Nonparametric Statistical Methods." Second Edition. John Wiley and Sons Inc. New York.

Jones, David T.; William R. Taylor; and Janet M. Thornton. 1992. "The rapid generation of mutation data matricies from protein sequences". Computer Applications in the Biological Sciences. 8:275-282.

Jones, Carol E., Toni M. Fleming, Don A. Cowan, Jennifer A. Littlechild, and Peter W. Piper. 1995. "The phosphoglycerate kinase and glyceraldehyde-3-phosphate dehydrogenase genes from the thermophilic archaeon *Sulfolobus solfactaricus* overlap by 8 bp." European Journal of Biochemistry. 233:800-808.

Kirimura, Kohtaro, Masashi Yoda, and Shoji Usami. 1999. "Cloning and expression of the cDNA encoding an alternative oxidase gene from *Aspergillus niger* WU-2223L." Current Genetics. 34:472-447.

Löffert, Dirk, Susan Karger, Margret Berkenkopf, Nicole Seip, and Jie Kang. 1997. "PCR optimization: primer design." QIAGEN News. (newsletter.) 1997(5):1-4.

Mitsuhashi, Masato. 1996a. "Technical report: part 1. Basic requirements for designing optimal oligonucleotide probe sequences." Journal of Clinical Laboratory Analysis. 10:277-284.

Mitsuhashi, Masato. 1996b. "Technical report: part 2. Basic requirements for designing optimal PCR primers." Journal of Clinical Laboratory Analysis. 10:285-293.

Montpetit, Michael L., Sharon Cassol, Teresa Salas, and Michael V. O'Shaughnessy. 1992. "OLIGOSCAN: A computer program to assist in the design of PCR primers homologous to multiple DNA sequences." Journal of Virological Methods. 36:119-128.

Meyer, Carl Dean. 2000. "Matrix analysis and applied linear algebra" Philadelphia. Society for Industrial and Applied Mathematics.

Moulton, J. Kevin. 2002. Personal communication. Interviewed in January 2002, when Dr. Moulton was a post-doc in the lab of Brian Wiegmann, in the Entomology Department at NC State University.

Olsen, Gary. 1990. "'Newick's 8:45' Tree Format Standard" http://evolution.genetics.washington.edu/phylip/newick_doc.html.

Oshima, H., R. Miyazaki, Y Ohe, H. Hayashi, K. Kawamura, and S. Kikuyama. 2002. "Molecular cloning of putative gastric chitinase in the toad *Bufo japonicus*." Zoological Science. 19:293-297.

Pollard, J. H. 1977. "A handbook of numerical and statistical techniques." Cambridge University Press.

Rose, Timothy M., Emily R. Schultz, Jorja G. Henikoff, Shmuel Pietrokovski, Claire M. McCallum, and Steven Henikoff. 1998. "Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences." Nucleic Acids Research. 26:1628–1635.

Sells, Mary Ann and Johnathon Chernoff. 1995. "Polymerase chain reaction cloning of related genes." Methods in enzymology 254:184-195.

Sjölander, Kimmen., Kevin Karplus, Michael Brown, Richard Hughley, Anders Krogh, I. Saira Mian, and David Haussler. 1996. "Dirichlet Mixtures: A method for improved detectin of weak but significant protein sequence homology." Computer Applications in the Biosciences 12:327-345.

Swofford, D., G. Olsen, P. Waddell, and D. Hillis. 1996. Chapter 11 in "Molecular Systematics 2nd edition". D. M. Hillis, C. Moritz, & B. K. Mable, editors. Sinauer.

Venugopal, T. V. Anathy, T.J. Pandian, G.Z. Gong, and S. Mathavan. 2002. "Molecular cloning of growth hormone-encoding cDNA of an Indian major carp, *Labeo rohita*, and its expression in *Escherichia coli* and zebrafish."

Watanabe, Kazuya, Yukimo Kodama, Shigeaki Harayama. 2001. "Design and evaluation of PCR primers to amplify vacterial 16S ribosomal DNA fragments used for community fingerprinting." Journal of Microbiological Methods. 44:253-262.

Whelan, S.; Paul I. W. de Bakker; and N. Goldman. "PANDIT: A database of protein and associated nucleotide domains with inferred trees. Ms. in prep. 2003

Wichman, Holly A., and Ronald A. Van Den Bussche. 1992. "In search of retrotransposons: exploring the potential of the PCR." Biotechniques. 13:258-265.

Wünschiers, Röbbe, Kerstin Stangier, Horst Senger, and Rüdiger Schultz. 2001. "Moleculare evidence for a Fe-hydrogenase in the green alga *Scenedesmus obliquus*" Current Microbiology. 42:353-360.

Yang, Ziheng; Sudhir Kumar; and Masatoshi Nei. 1995. "A new method of inference of ancestral nucleotide and amino acid sequences." Genetics 141:1641-1650.

Yang, Z. 1997. "PAML: a program package for phylogenetic analysis by maximum likelihood." CABIOS 13:555-556. (Version 3.13. August 2002.) http://abacus.gene.ucl.ac.uk/software/paml.html

Appendix A
Derivation of model to predict related sequences given the species relationships

Chapters 1 and 2 include derivations of formulas to assign probabilities to homologous, related sequences, assuming one did not know the attachment point or branch length connecting the new species to the known tree. It would be interesting to look at the problem in which one assumes one does know the attachment point and branch length,[1] even though this is more information than a researcher would normally know.

Starting with

$$P(\square \mid X) = \frac{P(\square, X)}{P(X)}$$

Eq. A.1

In Equation A.1, $\square$ is the state of the rat node and X is the four known sequences. In Equation A.2, $\square$ is the state of the parent node to the rat node.

$$P(\square \mid X) = \frac{\sum_{\text{all } \square} \left[P(\square \mid \square) P(\square, X) \right]}{\sum_{\text{all possible } \square} P(X \mid \square) P(\square)}$$

Eq. A.2

[1] The problem of predicting related sequences is similar to the problem of tracing evolution along one branch, forward in time.

The denominator is the whole-tree likelihood. The variable $\square$ steps through all possible states of the root node of the tree. $\square \in \{\text{amino acids}\}$.

Now replace $P(\square|\square)$ with the expression for the transition probability.

$$P(\square \mid X) = \frac{\sum_{\text{all } \square} \left[t_{\square\square\ \square} \, P(\square, X) \right]}{\sum_{\text{all possible } \square} \mathrm{P}(\mathrm{X} \mid \square) P(\square)}$$

Eq. A.3

Expand the expression for $P(\square, X)$

$$P(\square \mid X) = \frac{\sum_{\text{all } \square} \left[t_{\square\square\ \square} \, P(X|\square) P(\square) \right]}{\sum_{\text{all possible } \square} \mathrm{P}(\mathrm{X} \mid \square) P(\square)}$$

Eq. A.4

And showing the equation in terms of the full length of the segment. In Equation A.5, S is the site.

$$P(\square_{\mathrm{FL}} \mid X_{\mathrm{FL}}) = \prod_{S=1}^{7} \left[\frac{\sum_{\text{all } \square_S} \left[t_{\square_S\ \square\ \square_S, \square} \, P(X_S|\square_S) P(\square_S) \right]}{\sum_{\text{all possible } \square} \mathrm{P}(\mathrm{X}_S|\square) P(\square)} \right]$$

Eq. A.5

In the numerator, $\alpha_S \in \{\text{amino acids}\}$. $P(\alpha_S)$ is the a priori, or stationary, probabilities. The branch length is represented by the variable t.

Example of calculations with this model

The desire to do the following calculation would arise if a researcher knew the other 4 sequences in the alpha globin alignment, and the attachment point and length of the branch to the rat sequence, but did not know the rat alpha globin sequence, and wanted to design primers to find it. The researcher would use Equation A.5 to calculate the probability of the left-out rat sequence, given the other 4 sequences as known, and given the inferred globin tree minus the rat branch, and given the rat branch length as 0.17121.

$$P(\text{rat\_node} = \text{VISANNK} \mid X_{\text{FL}}, t) = 4.26 \cdot 10^{-5}$$

It is possible to calculate this kind of probability for all sequences that are exactly one amino acid mismatch from the VISANNK sequence. There are $7 \cdot 19 = 133$ of these sequences. Their probabilities sum to 0.0111054.

There are $19^2 \binom{7}{2} = 7581$ sequences that are exactly two amino acid mismatches from the VISANNK sequence. The sum of their probabilities is 0.010617.

At one point while designing primers, the researcher might consider using a pool of primers that would detect the 7715 amino acid[2] sequences that are zero, one, or two amino acid mismatches —equivalent to a difficult to specify amount of nucleotide mismatching, due to

[2] Sells and Chernoff type pools cover one amino acid sequence. How a pool of oligonucleotide primers could cover so many sequences is unclear.

degeneracy— from the sequence VISANNK (the actual (left-out) rat amino acid sequence). The total probability (of being the related sequences) this model assigns to these 7715 amino acid sequences is 0.021765.[3]

[3] Equivalently, the model is telling the researcher that the probability that one of those sequences is not the hypothetical rat sequence is 1 - 0.021765 = 0.978235.