

Adaptive molecular evolution in the *cryptosporidium parvum* I and II
genomes

by

Bradley A. Wangia

A project submitted in partial fulfillment
of the requirements for the degree of

Master of Science in Computer Science

Tufts University

2004

Approved by _____

Program Authorized to Offer Degree _____

Date _____

TUFTS UNIVERSITY

ABSTRACT

Adaptive molecular evolution in *Cryptosporidium Parvum* I and II
by Bradley A. Wangia

Co-advisors: Dr. Anselm Blumer & Dr. Lenore Cowen
Department of Computer Science

The protozoan parasite *Cryptosporidium Parvum* causes cryptosporidiosis in humans and various other animals. Type I of the parasite infects humans only and type II infects both humans and other animals. It is postulated that the C.Parvum I parasite may have lost its ability to infect other animals or type II may have gained the ability to infect animals along the evolutionary path. Such changes would be evident to statistical analysis at the genetic level. In this paper I explore the C.Parvum genome for this evidence using currently available statistical methods. I use Wall et.al.'s reciprocal smallest distance (rsd) algorithm to find putative orthologs between the C.Parvum I & II genomes. Their algorithm relies on global sequence alignment and maximum likelihood estimation of evolutionary distances and is a substantial improvement over reciprocal best blast hit for finding orthologs. In addition the distances found by rsd are directly proportional to the evolutionary distance between the two sequences. I use these distances to determine the most likely sequence pairs to have adapted under diversifying selection. A web interface is provided to explore our results. I hypothesize that after further analysis, these sequences will code for genes that will prove to be critical in enabling host specificity and therefore would be good targets for drug discovery.

TABLE OF CONTENTS

Neutral theory of molecular evolution.....	iv
Reciprocal best distance algorithm	vi
Pylogenetic Analysis of Maximum likelihood	viii
Cryptosporidium parvum orthologs identified.....	xii
Conclusion.....	xiii
Glossary.....	xiv
Bibliography	xvi

LIST OF FIGURES

<i>Number</i>	<i>Page</i>
A model of codon substitution	IX
Likelihood and Bayes	XVI

ACKNOWLEDGMENTS

The author wishes to acknowledge Dr. Anselm Blumer, and Dr. Lenore Cowen in the computer science department at Tufts University for their advisory role for this paper. I offer special thanks to Dr. Giovanni Widmer at the Department of Biomedical Sciences, Tufts University School of Veterinary Medicine for the formulation of the problem and review of the results.

I would also like to thank Katee Bates, Arlene, Christian and Shaneè Wangia for their concern, questions, and encouragement that kept me going during the course of this work.

Chapter 1

NEUTRAL THEORY OF MOLECULAR EVOLUTION

The work presented in this paper is heavily reliant on the neutral theory of molecular evolution developed from Kimura and Lewontin's publications. I therefore present a review of the theory here as obtained from wikipedia, an online encyclopedia.

When one compares the genomes of existing species, or looks between a species and its forebears, the vast majority of single-nucleotide differences are selectively "neutral." That is, these differences do not influence the fitness of either the species or the individuals who make up the species. As a result, the theory regards these genome features as neither subject to, nor explicable by, natural selection. This view is based in part on the genetic code, according to which sequences of three nucleotides (codons) may differ and yet encode the same amino acid (GCC and GCA both encode alanine, for example). Consequently, many potential single-nucleotide changes are in effect "silent" or "unexpressed". Such changes are presumed to have little or no biological effect.

A second assertion or hypothesis of the neutral theory is that most evolutionary change is the result of genetic drift acting on neutral alleles. A new allele arises typically through the spontaneous mutation of a single nucleotide within the sequence of a gene. In single-celled organisms, such an event immediately contributes a new allele to the population, and this allele is subject to drift. In sexually reproducing, multicellular organisms, the nucleotide substitution must arise within one of the many sex cells that an individual carries. Then only if that sex cell participates in the genesis of an embryo and offspring does the mutation

contribute a new allele to the population. Neutral substitutions create new neutral alleles.

Through drift, these new alleles may become more common within the population. They may subsequently decline and disappear, or in rare cases they may become "fixed"--meaning that the substitution they carry becomes a universal feature of the population or species. When an allele carrying one of these new substitutions becomes fixed, the effect is to add a substitution to the sequence of the previously fixed allele. In this way, neutral substitutions tend to accumulate, and genomes tend to evolve.

According to the mathematics of drift, when looking between two species or two isolated populations, most of their single-nucleotide differences can be assumed to have accumulated at the same rate as individuals with mutations are born. This latter rate, it has been argued, is predictable from the error rate of the enzymes that carry out DNA replication--enzymes that have been well studied and are highly conserved across all species. Thus, the neutral theory is the foundation of the molecular clock technique, which evolutionary molecular biologists use to measure how much time has passed since species diverged from a common ancestor. While the mutation rate is no longer considered a constant, diverse and more sophisticated clock techniques have emerged.

Based on this theory, I applied an implementation of D. P. Wall et. al.'s reciprocal smallest distance algorithm to the genomes of *cryptosporidium parvum* I and II. In the next few chapters, I discuss the rsd methods used, present an overview of the definitive test for diversifying selection at a site and then review the results of my analysis and conclude with additional work that need to be done.

Chapter 2

RECIPROCAL SMALLEST DISTANCE ALGORITHM

According to Wall et.al., comparisons of evolutionary rates of proteins in the absence of a normalizing molecular clock, must be based upon comparisons between orthologs; sequences that diverged from each other at the species split. Wall et. al.'s procedure for detecting putative orthologs improves on reciprocal blast hits.

Their method employs blast as a first step, starting with the subject genome, S , and a protein sequence, q belonging to the query genome Q . A set of hits, H , exceeding a predefined significance threshold (I used $E < 10e-20$) is obtained. Then using ClustalW each protein sequence in H is aligned separately with the original sequence q . If the alignable region of the two sequences exceeds a threshold fraction of the alignments total length (used 0.2) the program PAML is used to obtain a maximum likelihood estimate of the number of amino acid substitutions separating the two protein sequences, given an empirical acid substitution rate matrix. The model under which a maximum likelihood estimate is obtained may include variation in evolutionary rate among protein sites, and for more distant comparisons I used the default a gamma distribution with shape parameter $\alpha = 1.53$. Of all sequences in H for which an evolutionary distance is estimated, only s , the sequence yielding the shortest distance is retained. This sequence s , is then used for a reciprocal blast against genome Q , retrieving a set of high scoring hits, L . If any hit from L is the original query sequence, q , then the distance between q and s is retrieved from the set of smallest distances calculated previously. The remaining hits from L are then separately aligned with s , and maximum likelihood distance estimates are calculated for these pairs as previously described. If the protein sequence from L producing the shortest

distance to j is the original query sequence, q , it is assumed that a true orthologous pair has been found and their evolutionary distance is retained. The protein evolutionary distances obtained from the maximum likelihood method are directly proportional to relative evolutionary distances.

Chapter 3

PYLOGENETIC ANALYSIS OF MAXIMUM LIKELIHOOD

The latest release of the software package PAML developed and distributed by Ziheng Yang, is the implementation of his 2000 publication, “Statistical methods for detecting molecular adaptation”, with Joseph Bielawski. In that paper Yang and Bielawski point out that analyses that averaged evolutionary rates over sites and time had little power. They then went on to propose methods that are designed to detect positive selection at individual sites and lineages. I review their methods here in an attempt to give a definitive test for positive selection. The methods described here were not used to identify candidate pair but may be the basis for future work. Yang’s PAML software is used as is and no methods are added for this analysis.

Traditional methods of ortholog detection compared nonsynonymous substitution, a nucleotide substitution that changes the encoded amino acid, rate and the Synonymous substitution, a nucleotide substitution that does not change the encoded amino acid, in protein coding sequences. These methods regarded a nonsynonymous rate elevated above the synonymous rate as evidence of Darwinian selection. These rates are defined in the context of comparing two DNA sequences, with D_s and D_n as the synonymous and nonsynonymous substitutions per site respectively. The ration $\omega = D_n/D_s$ measures the difference between the two rates.

Yang et. al. present this ratio in the context of the codon substitution model shown below.

Box 2. A model of codon substitution

The codon is considered the unit of evolution. The substitution rate from codons i to j ($i \neq j$) is given as:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition.} \end{cases}$$

Parameter κ is the transition/transversion rate ratio, π_j is the equilibrium frequency of codon j and $\omega (= d_N/d_S)$ measures the selective pressure on the protein. The q_{ij} are relative rates because time and rate are confounded in such an analysis. Given the rate matrix $Q = \{q_{ij}\}$, the transition probability matrix over time t is calculated as:

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

where $p_{ij}(t)$ is the probability that codon i becomes codon j after time t . Likelihood calculation on a phylogeny involves summing over all possible codons in extinct ancestors (internal nodes of the tree). After Refs 16,18,27,79.

Figure1: A model of codon substitution

If an amino acid change is neutral, no net change in evolutionary fitness is achieved but the change can be fixed by chance, it will be fixed at the same rate as a synonymous mutation with $\omega = 1$. If the amino acid change is deleterious to fitness, natural selection will reduce its fixation rate in the population and therefore $\omega < 1$. Only when amino acid change offers a selective advantage is it fixed at a higher rate than a synonymous mutation, with $\omega > 1$. Therefore, a ω ratio significantly higher than one is convincing evidence of diversifying selection.

Yang and Bielawski use maximum likelihood based on explicit models of codon substitution to estimate D_n and D_s . Parameters in the model such as sequence divergence τ , transition/transversion rate ratio κ and the D_n/D_s ratio ω are estimated from the data and are used to calculate D_n and D_s according to their definitions. A major feature of their method is that the model is formulated at the

level of instantaneous rates, there is no possibility for multiple changes, and probability theory accomplishes all difficult tasks in one step: estimating mutational parameters such as κ ; correcting for multiple hits; and weighting pathways of change between codons.

Statistical tests can be used to test whether Dn is significantly higher than Ds . Here a likelihood ratio test is used. The null model has $\omega = 1$, whereas the alternative model estimates ω as a free parameter. Twice the log-likelihood difference between the two models is compared with a χ^2 distribution with one degree of freedom to test whether ω is different from one.

Yang et. al. contrast their method of detecting amino acid sites under Darwinian selection with previous methods. Most methods assume that all amino acid sites are under the same selective pressure, with the same ω ratio. This analysis would effectively average the ω ratios across all sites and positive selection is detected only if that average is > 1 . In their opinion this was too conservative a test for positive selection because many sites might be under strong purifying selection owing to functional constraints, with the ω close to zero. They reviewed recent work that tried to solve the problem but concluded that it would still be impractical to use one ω parameter for each site. Their approach used a statistical distribution to describe the variation of ω among sites. The test for positive selection then involves two major steps: first, to test whether sites exist where $\omega > 1$, which was achieved by likelihood-ratio test comparing a model that does not allow for such sites with a more general model that does; and second to use the Bayes theorem to identify positively selected sites when they exist. Sites having high posterior probabilities for site classes with $\omega > 1$ are potential targets for diversifying selection.

Yang reviews a previous implementation of likelihood-ratio test based on two simple models. The null model, M1 (neutral) assumes a class of conserved sites with $\omega = 0$ and another class of neutral sites with $\omega = 1$. The alternative model, M2 (selection), adds a third class of sites with ω estimated from the data. If M2 fits the data significantly better than M1 and the estimated ω ratio for the third class is > 1 , then some sites are under diversifying selection. However, this model lacked some power because M1 does not account for sites with $0 < \omega < 1$, and the third class in M2 is forced to account for such new sites. To cater for these, they implemented new models. For example, the beta distribution (M7 beta) is a flexible null model with $0 < \omega < 1$, and can be compared with an alternative that adds an additional site class with ω estimated. A general discrete model (M3) was also implemented. These models identify positive selection in six out of ten genes the author analyzed. The analysis used here used a flexible M3 model.

Chapter 4

CRYPTOSPORIDIUM PARVUM ORTHOLOGS IDENTIFIED

The distances obtained represent the maximum likelihood estimate of the number of amino acid substitutions separating the two protein sequences. C.Parvum I & II sequences shared the same time of divergence and therefore that protein evolutionary distance would be proportional to relative evolutionary rates. This would make the orthologs with the highest evolutionary distance, the best candidates for diversifying selection and therefore for host specificity.

I used data mining methods to cluster the genome wide results. The results of the analysis clustered into those pairs with ML distances greater than 0.05, those between 0.01 and 0.05, those with less than 0.01. I hypothesized that the first cluster would be of most interesting as it's pairs, represented related sequences that had split at the species level and had the highest likelihood of amino acid substitutions. Such sequences would be the most likely candidates for finding genes responsible for host specificity.

The highest ranking ortholog pairs found were

C.PARVUM 1 SEQUENCE	C.PARVUM 2 SEQUENCE	DISTANCE
CpT1H_3616-3-234-656	CpT2IOWA_VII_1-6-89165-88707	4.742
CpT1H_3648-5-4364-3888	CpT2IOWA_VIII_2-1-4300-4893	4.1734
CpT1H_3246-6-3454-2228	CpT2IOWA_IV_1-4-268280-266937	2.7109
CpT2IOWA_I_1-1-501961-502272	CpT1H_3198-2-2591-3007	2.4502
CpT1H_3823-4-11915-11559	CpT2IOWA_VIII_2-2-841517-841831	1.8306
CpT1H_3818-5-6941-6495	CpT2IOWA_VIII_1-2-175691-176158	1.7724
CpT1H_3553-6-3998-3531	CpT2IOWA_IV_2-6-510770-510228	1.2849
CpT2IOWA_Contig_7-3-12327-13157	CpT1H_3319-2-3251-3994	1.1394
CpT1H_3280-5-495-166	CpT2IOWA_VIII_2-6-716309-715998	1.1392
CpT1H_3818-6-6757-6071	CpT2IOWA_VIII_1-2-168275-168895	1.0645
CpT1H_3852-5-19595-19248	CpT2IOWA_IV_2-5-473508-473194	1.0619

A complete listing of the results can be found at

<http://www.cs.tufts.edu/~bwangia/mproject/results/>

An XML capable web browser is required (**Internet Explorer 5.0+**, **Netscape 6+**). I decided to use XML as it affords the ability to run clustering methods easily.

Chapter 5

CONCLUSION

The validity of my hypotheses that some of the sequences in cluster 1 could code for genes that contribute to host specificity is unconfirmed and needs further investigation. While I attempted to run blast searches for the top ranking sequences, no results for the possible function of most of these sequences was found. This may not be a discouraging result since the genomes in question are not annotated and the functions of the genes may not be available in popular databases yet. For these results to be fully reviewed the genes that these sequences code for would need to be identified and their function determined. As demonstrated in chapter 2, methods exist to provide stronger analysis for positive selection.

While these sequences showed promise as a source of future work on host specificity, their expression and regulation could play an important role in the complex interaction between host and parasite that determine host specificity. Other assumptions like the role that gene interaction plays in the survival of the parasites in a specific host would also need to be investigated.

This project took advantage of software packages developed specifically for certain tasks. These components were customized and tied together then applied to the cryptosporidium genome. The use of such component software will be crucial in developing timely and effective solutions to different problems in computational biology.

GLOSSARY

Codon-usage bias. Unequal codon frequencies in a gene.

Equilibrium frequency.

$$\Pi_{ijk} = \frac{\pi_i \pi_j \pi_k}{1 - \sum_{\text{stop codons}} \pi_1 \pi_2 \pi_3}$$

Equilibrium frequency of codon ijk is the product of observed frequencies of nucleotides which compose the codon, normalized by a factor to account for the presence of stop codons

Mutation. Alteration in the nucleotide sequence.

Types of mutations

Enhances fitness. Likely to be rapidly fixed – 'positive' selection.

Deleterious to fitness. likely to be rapidly lost – 'negative' or 'purifying' selection.

Neutral. no net change in fitness, but can be fixed by chance ('random genetic drift').

Mutation Fixation. the tendency of a mutation to become a permanent part of the gene pool of a population.

Deletions/Insertions. ('indels') are removal/addition of one or more nucleotides.

Duplications. copying and insertion of a nucleotide sequence (a major driving force for diversification of genetic information).

Nonsynonymous substitution. a nucleotide substitution that changes the encoded amino acid.

Prior probability. the probability of an event (such as a site belonging to a site class) before the collection of data.

Positive selection. darwinian selection fixing advantageous mutations with positive selective coefficients. The term is used interchangeably with molecular adaptation and adaptive molecular evolution.

Posterior probability. the probability of an event conditional on the observed data, which reflects both the prior assumption and information in the data.

Purifying selection. natural selection against deleterious mutations with negative selective coefficients. The term is used interchangeably with negative selection or selective constraints.

Synonymous substitution. a nucleotide substitution that does not change the encoded amino acid.

Transition/transversion rate bias. unequal substitution rates between nucleotides, with a higher rate for transitions (changes between T and C and between A and G) than transversions (all other changes).

Transversion. A point mutation in which a purine (adenine and guanine) is replaced by a pyrimidine (cytosine, and thymine), or a pyrimidine is replaced by a purine.

Transition. A point mutation in which a pyrimidine is replaced by another pyrimidine, or a purine is replaced by another purine.

Box 3. Likelihood and Bayes

The statistical-estimation theory used in the methods discussed in this review can be explained with the following simple hypothetical example. Suppose that a population is an admixture of two groups of people in the proportions 60% and 40%, and a certain disease occurs at a rate of 1% in Group I and of 0.1% in Group II. Suppose a random sample of 100 individuals is taken from the population, what is the probability that three of them carry the disease? The probability that a random individual carries the disease (D) is an average over the two groups (G_1 and G_2):

$$p = P(D) = P(G_1) \times P(D|G_1) + P(G_2) \times P(D|G_2) = 0.6 \times 0.01 + 0.4 \times 0.001 = 0.0064 \quad (1)$$

Similarly, the probability that an individual does not carry the disease is:

$$\begin{aligned} P(\bar{D}) &= P(G_1) \times P(\bar{D}|G_1) + P(G_2) \times P(\bar{D}|G_2) \\ &= 0.6 \times 0.99 + 0.4 \times 0.999 = 0.9936 = 1 - p \end{aligned} \quad (2)$$

The probability that three out of 100 individuals carry the disease is given by the binomial probability:

$$P = \frac{100!}{3! \times 97!} [p^3 (1-p)^{97}] = 0.0227 \quad (3)$$

If Eqn 3 involves an unknown parameter [such as the rate $P(D|G_i)$ in Group I], that parameter can be estimated by maximizing Eqn 3. In that case, Eqn 3 gives the probability of observing the data (sample) and is called the likelihood function.

The second question is to calculate the probability that an individual in the sample who carries the disease is from Group I. The Bayes theorem gives this probability as:

$$P(G_1|D) = P(G_1) \times P(D|G_1)/P(D) = 0.6 \times 0.01/0.0064 = 0.94 \quad (4)$$

Note that this is just the proportion of the contribution from Group I to $P(D)$ in Eqn 1. Thus, this individual is most likely to be from Group I. Similarly, a healthy individual in the sample is more likely to be from Group I than from Group II because

$$\begin{aligned} P(G_1|\bar{D}) &= P(G_1) \times P(\bar{D}|G_1)/P(\bar{D}) = 0.6 \times 0.99/0.9936 = 0.5978 \\ \text{and } P(G_2|\bar{D}) &= 1 - P(G_1|\bar{D}) = 0.4022 \end{aligned} \quad (5)$$

In methods for inferring sites under positive selection^{36,37}, we let D in the example be the data at a site and G_i be the i th site class with the d_N/d_S ratio ω_i . The probability of observing data at a site is then an average over the site classes (Eqn 1). The product of such probabilities over sites constitutes the likelihood (Eqn 3), from which we estimate any unknown parameters, such as the branch lengths and parameters in the ω distribution over sites. After the parameters are estimated, we use the Bayes theorem to calculate the probability that any site, given data at that site, is from each site class (Eqns 4 and 5).

Another straightforward application of the theory is ancestral sequence reconstruction; in this case, we replace G_i with a reconstruction (characters at interior nodes of the phylogeny) at a site. When we calculate the likelihood function, the probability of data at a site $P(D)$ is a sum over all possible ancestral reconstructions (G_i s) (Eqns 1 and 2). After parameters are estimated, the reconstruction that makes the greatest contribution to $P(D)$ is the most likely (Eqns 4 and 5)²⁴.

The Bayes method discussed here is known as the empirical Bayes, because it uses estimates of parameters and does not account for their sampling errors. This might be a concern if parameters are estimated from small samples or if the posterior probabilities are sensitive to parameter estimates. An alternative approach is the hierarchical Bayes method, which accounts for the uncertainty in unknown parameters by averaging over their prior distribution.

Note that the reconstructed ancestral sequences²⁴, as well as the inferred site classes in the site-class models^{36,37}, are pseudo data and involve systematic biases. To appreciate such biases, note that in the previous example, the Bayes calculations (Eqns 4 and 5) predict that each of the 100 individuals in the sample, healthy or sick, are from Group I. Although this is the best prediction, the accuracy is low. If such inferred group identities are used for further statistical analysis, misleading results might follow.

BIBLIOGRAPHY

- [1] Clark, A. et. al. *Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios*. Science, Vol. 302, December 2003.
- [2] Kimura, Motoo *Neutral theory of molecular evolution* Cambridge University Press, 1983.
- [3] Lewontin, Richard *The Genetic Basis of Evolutionary Change* Columbia University Press 1974.
- [4] Spencer V. Muse . *Estimating Synonymous and Nonsynonymous Substitution Rates*. Mol. Bio. Evo. 1995.
- [5] Wall, D. P., et al. *Detecting putative orthologs*. Bioinformatics Vol. 19 no 13 2003, pages 1710-1711.
- [6] Yang Ziheng, et. al . *Statistical Methods for detecting molecular adaptataion*. Science, Vol. 15, no 12. December 2000
- [7] Yang Ziheng and Rasmus Nielsen. *Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models*. Mol. Biol. Evol. 17(1):32-43. 2000