Ruprecht-Karls-Universität Heidelberg
Institut für Informatik
Lehrstuhl für Datenbanksysteme

Masterarbeit

Name: Jörg Hauser
Betreuer: Prof. Dr. Michael Gertz
Dr. Alexandros Stamatakis
Abgabedatum: November 4, 2011

Ich versichere, dass ich diese Bachelor-Arbeit selbstständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

Abgabedatum: November 4, 2011

Zusammenfassung

Abstract

Given n species there are $\prod_{i=3}^{n}(2i-5)$ different binary tree topologies of which potentially represents their evolutionary relation. Phylogenetics searches for the most likely arrangement of species within this rapidly growing number of possible trees topologies, employing DNA or amino acid sequences. A trivial approach would be to compute the likelihood of any possible tree – based on the given input alignment data – and pick that resulting the highest value. Anyway because the immense number of possible topologies this approach is infeasible even with today's computing power.

In addition to tree-topologies phylogenetic search-space complexity increases due to tree branch lengths and evolutionary models plus their parameters. One basic thing used to compute likelihood values are models of base or amino acid substitution for DNA and proteins respectively. These approximate the probability that one base or amino acid mutated into another after a given time. Especially for proteins (i.e. amino acids sequences) there exist various models of evolution. Depending on which model one choses for phylogenetic analyses the likelihood of the considered evolutionary trees may differ. One problem therefore is to find the model which best reflects the actual evolution of the given partition of DNA or proteins (i.e. the model which maximizes the likelihood within all possible combinations of tree topologies, branch lengths and models).

Although there is some information lost during the proteinbiosynthesis (since not all base substitution in the DNA also alter the resulting amino acid), sometimes in phylogenetic analyses trees are calculated based on protein alignment data in stead of DNA alignments. For example when there is assumed a long distance in between species – preferring protein-data over DNA is promising. Because of the limited alphabet in DNA (four possible states) back substitution (e.g. A $\rightarrow$ T $\rightarrow$ A) is more common than within the 20 "standard-"characters long alphabet of amino acids.

As already mentioned, the longer alphabet on the other hand caused the rise of many protein evolution models, of which each is represented by a matrix with 20x20 entries (DNA substitution matrices 4x4), which makes it even more expensive to estimate the model that best fits the evolutionary process of the given data from a computational point of view.

This thesis shall analyze the influence of different protein substitution models on the resulting phylogenies and propose heuristics to efficiently calculate the protein evolution model which maximizes the likelihood of the resulting evulutionary trees.

Contents

1 Introduction

1.1 Motivation and Contribution

Partitions of genes can evolve at different speed and with varying properties. Thus it is important to choose the best model to partition assignment during phylogenetic inference. Unfortunately first it is still unknown, which gene partitions mutated according to which model of evolution, and second there are plenty of scientifically justified models available. Thus it is difficult to motivate the assignment applied. In many of todays publications, models appear to be chosen arbitrary. Because every combination of models maybe yields a different tree all of them should be considered. With m^n combinations of m models to n partitions, however it is not efficient to check any assignment with a growing number of partitions.

If one further considers, that a varying model mapping can have impacts on the overall resulting trees, ideally the tree topology and its branch lengths should be optimized alongside. Only if all possible states of topology, branch lengths and model mapping are considered, overall maximum likelihood can be achieved. However luckily, tree topology does not influence model mapping reasonably, at least for only one partition [5].

This works main goal is to provide an overview of technical impacts on the model partition mapping and proposes some heuristics to quickly asses model assignments. In the end at least the question if it matters for real world data shall be answered (do the trees differ with optimal model assignment differ reasonable from those with an arbitrary mapping).

1.2 Structure

2 Phylogenetics Introduction

The aim of this chapters is to provide the needed preliminaries and an overview of phylogenetics. Although strictly speaking there is no related work, as there was not much focus on optimal model assignment until now, the last section briefly compares to Posada and Buckley and Darriba et al. approaches in [4] and [1] to detect a statistically well suited model for unpartitioned DNA and AA alignments respectively.

Humans are interested in the origin of life and especially man kind since a long time. Phylogenetics plays a major role in developing knowledge about rise of life, as it describes the task of reconstructing evolutionary relationship of organisms. These organisms can either be entire species (e.g. humans and apes evolved from a common ancestor) or individuals (e.g. this person was an ancestor of this child). It is obvious that knowledge of evolution offers a start point for classification. Thus phylogenetics is an essential part of biological systematics. Most often the goal is to organize taxa (sets of organisms or individuals) in phylogenetic trees. Within this thesis the terms taxa and species are often used synonymously, referring to taxa. Even though there are alternative strategies available (e.g. phylogenetic networks) to describe evolutionary relationship, these will not be discussed in this thesis, because tree structures are the common practice.

In a phylogenetic tree branches (edges) represent the amount of evolution among the connected taxa (e.g time or divergence). At the leaf nodes (also tips) observed alignments of the taxa are placed. The inner nodes are what makes inference of phylogenetic trees computationally expensive, as they represent hypothetic ancestors. Usually they are not observable anymore. Moreover a binary tree structure is usually assumed. These can either be rooted, at the most recent common ancestor of the tips, or unrooted (i.e. not making any statement about the initial organism). The option to omit the root first reduces computational needs, but second needs for a time reversible model of evolution (Section 2.3).

Historically phylogenetic trees were inferred morphologically, i.e. observable properties of the organisms serve to classify them and reconstruct evolution. Nowadays molecular phylogenetics is a frequently used alternative. Correlations as well as differences between both methods have been observed. Thus it's still an ongoing debate

whether, and if yes, then why one approach is more accurate than the other. Although both approaches seem to be of valuable from a biologists point of view, only molecular phylogenetics will be discussed in this work.

2.1 Molecular Phylogenetics

The task of molecular phylogenetics is reconstructing evolutionary relationship between taxa based on their genetic code (often also called phylogenomics, combining genomics and phylogenetics). Although most approaches in theory can be applied to many types of data of organisms, frequently either deoxyribonucleic acid (DNA) or amino acid (AA) alignment data is used. To abstract from chemical properties both are represented as sequences of characters abbreviating their names. For DNA four characters A,C,G and T (adenine, cytosine, guanine and thymine) can occur. For proteins there are twenty standard amino acids, which are enumerated in Listing 2.1.

```
A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val
```

Listing 2.1: Amino acids and their corresponding abbreviations.

2.1.1 Sequence Alignment

Evolution does not only consist of mutations of genes, insertions and deletions can also occur. Thus, not necessarily all organisms share the same amount of genetic information. This results in character sequences of different lengths. To compare sequences of different lengths, alignments must be calculated, firsthand. The quality of alignment results are especially important to molecular phylogeny inference, as they are the only concrete input to infer evolution. Therefore results can only be of good quality, if the alignments were so.

2.1.2 Data Partitions

and their linkage over time

1. where does one derive partitioned AA data from
2. What stand the partitions for
3. Trees are calculated based on sites (i.e. parts of partitions)
4. different evolutionary models may suit best for different partitions

5. but partitions are linked by time

2.1.3 Phylogenetic Tree Inference

Parsimony

Similar to Levenshtein-Distanz and Hamming distance ...

Distance Matrices

Maximum Likelihood

[2]

Bayesian methods

2.2 Amino Acid vs. Deoxyribonucleic Acid

Although DNA encodes for AA, there occur differences in resulting trees depending on which data one uses. On first glance one could guess that this must be due to mistakes. But biological process is not simple and there occurs a lot of information processing within organisms. As already mentioned, theoretically all intermediate molecular structures between DNA and AA, can be used to infer evolutionary trees, the majority of studies, however, uses either DNA or AA alignments. DNA is transformed multiple times in order to contribute to organisms' forming. In fact protein sequences were the first sequences derived. Overall one can say, that shape of species is made up of proteins (i.e. sequences of AA) rather than DNA. To bridge the gap between DNA in cells' core, DNA is transcripted into messenger RNA (mRNA), whic is carried to cytosol. Afterwards mRNA is translated into the corresponding amino acids. The process from DNA to AA is called protein biosynthesis.

A good portion of the organic information processing has been recovered so far. It's known, that codons (triplets of DNA) encode for one amino acid. For three bases (DNA) there are $3^4 = 81$ possible states. These encode for 20 common AAs plus some codons which are needed for information processing in organisms (start-/stop codons). In this process, therefore information is lost, as 81 states encode for about 20 afterwards. Protein data is frequently applied, anyway.

In fact the question of which data to use is disputed, as there must be less information available in proteins. To provide at least one reason for preferring protein data over DNA, the longer alphabet of proteins avoids backsubstitiution (A→T→A). Especially

for analyses regarding longer evolutionary distances, this can be beneficial, because not all DNA substitutions also alter resulting amino acids. For example if the codon ATA encoded for J and there was a substitution to AAA, which also encodes J and back to ATA again, this is a change (as mutation often has a negative conotation) not recognizable at protein level, as well as not essential at DNA level.

2.3 Models of Protein Evolution

A model of evolution is a simplified blueprint of how evolution could have been. This evolutionary history is described by a matrix specifying the probability of change of each base or amino acid into each other after a given amount of time. Change rates can for example be derived from empirical observations of closely related alignments, . There are many influences on AA models (amino acids' secondary structure as well as environmental stress can impact replacement rates), thus one must be aware that they remain models, although they become close to reality.

Having the ability to explicitly specify a model of evolution, anyhow, is beneficial in comparison to approaches with built in assumptions (e.g. Parsimony 2.1.3). On the other hand, this rises the need of detecting and specifying the model, which as already mentioned, is a non-trivial task.

2.3.1 Basics

Introduce AA models

2.3.2 Similarities

Detect similarities of those models (could be used to reduce computational costs)

2.4 Related Work

There is no related work, although there exist prior approaches to determine evolutionary models for DNA as well as AA data (ModelTest, ProtTest[1]). All of them are only applicable for data containing only one partition. ModelTest, as well as ProtTest, try to estimate the most appropriate model among a candidateset by using Akaike Information Criterion or Bayesian Information Criterion. In difference to these approaches the

[1] `http://darwin.uvigo.es/software/prottest.html`

one described here focuses on maximizing the likelihood only. Thus, a well-grounded proposal of which models best reflects real evolution can not be given. The returned assignments of models to alignment partitions maximizes the likelihood of resulting trees. A guarantee of closeness to real evolution is not possible for the others, as well. Though all approaches show to recover models correct, frequently for synthetic data.

3 Preliminaries

The former section introduced the task of inferring evolutionary trees from DNA or protein alignment data in general. Different models of evolution have been introduced. This chapter focuses on determining the "best" suiting model mapping for each alignment partition.

Although an increasing likelihood does not necessarily imply a model assignment that is closer to reality, as the likelihood method itself only finds the tree, that is most likely to produce the observed data – in contrast to the probability of the found tree or topology to be the true one. However, model assignments with maximum likelihood are assumed to be quite close to reality as they are likely to produce the observed data, and thus sufficient to get a simplified optimization problem.

3.1 Problem Description

Assume partitioned protein data for s species, that contains n partitions (as outlined in Figure 3.1). Further suppose there are m different protein substitution models available. Ideally the task of adequately assigning models to partitions is described by picking one model per partition, so that the log-likelihood of the resulting phylogenetic tree (which is to be inferred afterwards or retrieved during the process of model optimization) is maximized (i.e. among all possible tree topologies, branch lengths and model assignments return those maximizing the likelihood).

Unfortunately this ideal scenario seems to be computationally infeasible because of two facts. First it contains the already computationally expensive problem of maximizing the likelihood with fixed protein evolution models and varying tree topology, and second partitions are linked by "time", i.e. models can't be optimized on a per partition basis (The assignment of model m_1 to partition p_1 can thus worsen the overall result, so that the former optimal model m_2 for partition p_2 is not optimal any more).

For common binary phylogenetic trees there are $\prod_{i=3}^{s}(2i-5)$ possible topologies. The most naive approach, would have to test each of the m^n model combinations for every topology and optimize branch lengths accordingly to achieve likelihood values as close

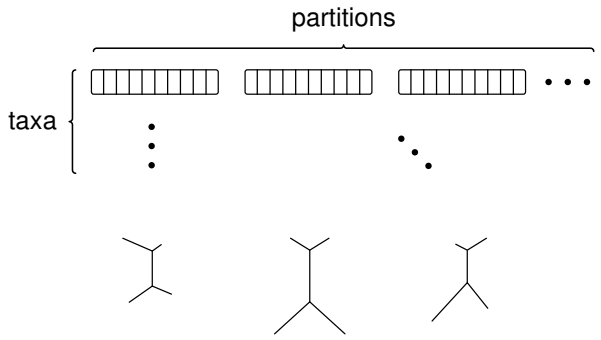

Figure 3.1: Concept and annotation of partitioned data-set. Partition optima may yield different branch lengths per partition. Tree topology is assumed to be fixed during model selection.

to optimum as possible.

3.1.1 Fixed Tree Topology

To reduce the complexity, a reasonable tree topology is assumed to be fixed. Of course this can be a major constraint, which causes results not to be valid for all circumstances. First Posada and Crandall in [5] for DNA evolution models found that tree topology does not influence results much, as long as it is a reasonable one, and second parsimony start trees are often used as starting point for maximum likelihood tree searches as well. Therefore assuming efficiently to compute parsimony trees as given seems to be an acceptable constraint to further evaluate optimization approaches. To ensure that the selected models yield better likelihood values and trees, results will be compared with traditional ones. In addition to that, runs with simulated data (i.e. the real tree is known) look promising.

Exhaustively solving the reduced instance of the model selection problem (i.e. enumerate and check all model combinations) is dominated by determining the ML value for each of the m^p combinations and estimate the joint branch lengths.

3.1.2 Example

The example alignments for four species (Listing 3.1) assumes that there were only three models of amino acid evolution (e.g. WAG, JTT and LG), and a AA alignment containing four taxa with two partitions (each should be assigned the best model). Of course at least two partitions are needed, to show the difficulty of partitioned data. Moreover four taxa were chosen, because this is the minimal number, where resulting trees can be of different topology, as for less taxa there exist at most one unrooted binary

Figure 3.2: Left three partitions with x sites, right one partition with $3 \times x$ sites

tree.

```
Spec1 KQQIILTYAATKGLKYLGLT
Spec2 KQHLLLYAAGQKGYKQLGLQ
Spec3 KEPVLLYYAREKGVTRLGYQ
Spec4 QEAVLLYFARQRGVTRLGYQ
```

Listing 3.1: Example

The following tree shall further exactly reflect the relatedness between these four species.

3.2 Per-Site Model Estimate

A per-partition model test could further be extended to per a site basis. This extension would even increase the number of possible models to $m^{\text{alignment length}}$. For example in Figure 3.2 with $m = 20$ there are 20^3 combinations for per partition optimization against 20^{30} per site (each partition contains 10 AA).

4 Algorithmic Approaches

This chapter provides an overview of the options one has to optimally map available models to alihnment partitions. Furthermore the impact on complexity and results will be discussed.

4.1 Search Strategies

4.1.1 Exhaustive Model Assessment

4.1.2 Naive Heuristic

Longer partitions have more influence on the overall likelihood than shorter ones. Thus optimizing models per partition and assigning the per partition optima is an intuitive starting point. Anyhow, branch lengths for the per partition optima are likely vary. Thus this initial mapping will not yield the global optimum for many data.

4.1.3 Simulated Annealing

4.1.4 Multi-Objective Optimization

4.2 Reducing Computational Costs

4.2.1 Classification of Models

Of a group of models, that is likely to yield similar results, only test one. Model options (+I, +G, +F) are not regarded anyway ...

4.2.2 Loosening of Algorithmic Requirements

Could a mixture of methods improve performance, without worsening quality (e.g. Removing alpha optimization after model assignment does not influence quality ...)

5 Experimental Setup and Results

The Expiriments run to evaluate the approaches of the former, will be discussed in this chapter. There are two questions, which form the crux:

1. does optimal model mapping really matter and
2. how realistic are the found assignments (compared to the statistical approach of Darriba et al.)

5.1 Tests

5.2 Result Evaluation

5.2.1 Evaluating Model Inference

For synthetic data, do the approaches detect the assumed model assignment?

Tree Similarity

5.2.2 Datasets

Biological Datasets

Simulated Data

5.2.3 Computation Infrastructure

5.3 Results

5.3.1 Accuracy of Model Matches

5.3.2 Impact on Tree Topologies

1. Does it really matter, how can one say if it matters...

- it matters if resulting trees change with optimal model assignment. Of course the resulting tree should have a better likelihood, otherwise it would not be the best assignment...
- Could it happen, that the best assignment in an additional tree search yields a different tree with worse likelihood?

5.3.3 Runtime Analysis

5.4 Discussion

6 Conclusions and Outlook

Bibliography

[1] D. Darriba, G.L. Taboada, R. Doallo, and D. Posada. Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164, 2011.

[2] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

[3] D. Posada and T.R. Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793, 2004.

[4] D. Posada and K.A. Crandall. Modeltest: testing the model of dna substitution. *Bioinformatics*, 14(9):817, 1998.

[5] D. Posada and K.A. Crandall. Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, 50(4):580, 2001.