Development of New Methods for Inferring and Evaluating Phylogenetic Trees

TOBIAS HILL
Dissertation presented at Uppsala University to be publicly examined in B7:101a, BMC, Husargatan 3, Uppsala, Friday, March 9, 2007 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract


Inferring phylogeny is a difficult computational problem. Heuristics are necessary to minimize the time spent evaluating non-optimal trees. In paper I, we developed an approach for heuristic searching, using a genetic algorithm. Genetic algorithms mimic the natural selections ability to solve complex problems. The algorithm can reduce the time required for weighted maximum parsimony phylogenetic inference using protein sequences, especially for data sets involving large number of taxa.

Evaluating and comparing the ability of phylogenetic methods to infer the correct topology is complex. In paper II, we developed software that determines the minimum subtree prune and regraft (SPR) distance between binary trees to ease the process. The minimum SPR distance can be used to measure the incongruence between trees inferred using different methods. Given a known topology the methods could be evaluated on their ability to infer the correct phylogeny given specific data.

The minimum SPR software the intermediate trees that separate two binary trees. In paper III we developed software that given a set of incongruent trees determines the median SPR consensus tree i.e. the tree that explains the trees with a minimum of SPR operations. We investigated the median SPR consensus tree and its possible interpretation as a species tree given a set of gene trees. We used a set of o-proteobacteria gene trees to test the ability of the algorithm to infer a species tree and compared it to previous studies. The results show that the algorithm can successfully reconstruct a species tree.

Expressed sequence tag (EST) data is important in determining intron-exon boundaries, single nucleotide polymorphism and the coding sequence of genes. In paper IV we aligned ESTs to the genome to evaluate the quality of EST data. The results show that many ESTs are contaminated by vector sequences and low quality regions. The reliability of EST data is largely determined by the clustering of the ESTs and the association of the clusters to the correct portion of genome. We investigate the performance of EST clustering using the genome as template compared to previously existing methods using pair-wise alignments. The results show that using the genome as guidance improves the resulting EST clusters in respect to the extent ESTs originating from the same transcriptional unit are separated into disjunct clusters.

Keywords: Evolution, Phylogeny, SPR, Genetic Algorithm, Tree metrics

Tobias Hill, Department of Neuroscience, Box 593, Uppsala University, SE-75124 Uppsala, Sweden

© Tobias Hill 2007

ISSN 1651-6206
ISBN 978-91-554-6799-7
urn:nbn:se:uu:diva-7501 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-7501)
Sic gorgiamus allos subjectatos nunc
List of publications and manuscripts

This doctoral thesis is based on the following papers:


II. Tobias Hill, Mikael Thollesson, Tommy Säfström, Andreas K. E. Vernersson, Robert Fredriksson and Helgi B. Shiöth. Determining the minimum number and nature of the SPR operations separating pairs of unrooted phylogenetic trees. Manuscript

III. Tobias Hill, Mikael Thollesson, Robert Fredriksson and Helgi B. Schiöth. Inferring species trees from genomes with high horizontal gene transfer and recombination rates using minimum SPR distances. Manuscript

Contents

Introduction.....................................................................................................9
Evolution......................................................................................................9
  Genes ....................................................................................................9
  Gene families ....................................................................................10
Sequence data ........................................................................................11
  Expressed sequence tags ....................................................................11
  Sequence alignment ............................................................................12
Phylogeny .................................................................................................13
  The number of possible trees ............................................................13
  Inferring phylogenetic trees .............................................................15
  Distance based phylogenetic inference ..........................................15
  Character based phylogenetic inference .........................................16
  Searching for trees ............................................................................18
  Genetic algorithms ..........................................................................20
Comparing phylogenetic trees ...............................................................22
  The SPR distance ............................................................................22
  Trees within trees ............................................................................24
  Median SPR consensus ....................................................................25
Research aims ..........................................................................................27
Results and discussion ............................................................................28
  Paper I ..............................................................................................28
  Paper II ............................................................................................29
  Paper III ............................................................................................29
  Paper IV ...........................................................................................30
Conclusions ...............................................................................................31
Future perspectives ..................................................................................32
Acknowledgements ..................................................................................34
References .................................................................................................36
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>BLAT</td>
<td>BLAST Like Alignment Tool</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit</td>
</tr>
<tr>
<td>EMBL</td>
<td>The European Molecular Biology Laboratory</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed Sequence Tag</td>
</tr>
<tr>
<td>FPT</td>
<td>Fixed Parameter Tractable</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic Algorithm</td>
</tr>
<tr>
<td>GAML</td>
<td>Genetic Algorithm for Maximum Likelihood</td>
</tr>
<tr>
<td>GEC</td>
<td>Genomic EST Clustering</td>
</tr>
<tr>
<td>GPCR</td>
<td>G-protein coupled receptor</td>
</tr>
<tr>
<td>HGT</td>
<td>Horizontal Gene Transfer</td>
</tr>
<tr>
<td>LCA</td>
<td>Last Common Ancestor</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MP</td>
<td>Maximum Parsimony</td>
</tr>
<tr>
<td>MYA</td>
<td>Million Years Ago</td>
</tr>
<tr>
<td>NCBI</td>
<td>The National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NJ</td>
<td>Neighbor Joining</td>
</tr>
<tr>
<td>NNI</td>
<td>Nearest Neighbor Interchange</td>
</tr>
<tr>
<td>NP</td>
<td>Non-deterministic Polynomial</td>
</tr>
<tr>
<td>OTU</td>
<td>Operational Taxonomical Unit</td>
</tr>
<tr>
<td>PAM</td>
<td>Percent Accepted Mutation</td>
</tr>
<tr>
<td>RNG</td>
<td>Random Number Generator</td>
</tr>
<tr>
<td>SPR</td>
<td>Subtree Prune and Rgraft</td>
</tr>
<tr>
<td>TBR</td>
<td>Tree Bisection and Reconnection</td>
</tr>
<tr>
<td>TNT</td>
<td>Tree analysis using New Technology</td>
</tr>
<tr>
<td>UPGMA</td>
<td>Un-weighted Pair Group Method with Arithmetic Mean</td>
</tr>
<tr>
<td>UTR</td>
<td>UnTranslated Regions</td>
</tr>
<tr>
<td>WGD</td>
<td>Whole Genome Duplication</td>
</tr>
</tbody>
</table>
Introduction

Evolution

Evolution is the process by which the complexity and diversity of life has come into place. Evolution refers to all the changes that have transformed life from the earliest beginning to the forms that we observe today (Campbell, 1993). Charles Darwin introduced the term natural selection as the principal mechanism by which evolution works. Natural selection means that the organism that is best adapted to their environment has a higher probability of producing an above average number of offspring. The next generation would then contain a higher number of individuals with the favorable traits. The favorable traits will spread through the population due to the above average offspring of those who carry them. The favorable traits will become norm in the population until other even more successful adaptations appear.

For natural selection to work, traits need to be inherited from one generation to the next. If the offspring is a totally random combination of traits that does not depend on the parents, natural selection would not work. Traits are currently inherited through DNA passed from one generation to the next. Earlier in evolutionary history traits might have been passed by other simpler molecules or the actual molecule itself might have been the trait to inherit (Chen, et al., 2005). In animals the DNA molecules are passed to the next generation in two ways, the main DNA molecules that contains the information needed to construct and regulate the new organism and the mitochondrial DNA that is used to build the mitochondria which serves as the power plant of the cells.

Genes

Genes are the unit by which traits are inherited. Genes are encoded in the DNA and susceptible to various mechanisms that are considered to underlie evolution. Novel gene functionality is generally created by copying existing genes rather than evolving completely new genes. The 2R theory for genome duplication proposes that two rounds of whole genome duplications (WGD) occurred early in vertebrate history (Ohno, 1970). These duplications took place more than 400 million years ago (MYA) resulting in up to four copies of each gene originating from a common ancestor of vertebrates and invertebrates such as Drosophila (Dehal and Boore, 2005; Larhammar, et al., 2002;
Two mechanisms have been proposed to result in WGD, allopolyploidy and autopolyploidy (Wolfe, 2001). Autopolyploid species have multiple copies of chromosomes derived from a single species while allopolyploid organisms have chromosomes derived from different species as the result of the fusion of two fertilized oocytes (Chenuil, et al., 1999). There are studies that see only one round of duplication (McLysaght, et al., 2002) and others that attribute the duplications found to limited segment duplications (Friedman and Hughes, 2003). Recent studies have shown that limited segment duplications occur with a rate sufficient to duplicate an entire genome in 100 million years in eukaryotes (Lynch, et al., 2001). Limited segment duplications are usually generated by unequal crossing-over resulting in a region of chromosome with several copies of the ancestral gene. The result of duplication is two or more functional copies of the ancestral gene.

Only a small number of the duplicated genes will undergo neofunctionalization (the gene acquires novel functionality that is preserved by natural selection) or subfunctionalization (the pair of genes partition the ancestral function) in time to avoid nonfunctionalization by deleterious mutations within a few million years (Force, et al., 1999).

Point mutation is another mechanism that may change genes and thereby their effect on the organism that carries them. A point mutation is the replacement of one ribonucleic acid in the DNA by another or an insertion/deletion. Mutations will in most cases degrade the fitness of the gene and natural selection will not favor them. A favorable mutation however will spread through the population and thus over time becomes fixated in the genome.

The complex relationship between the organisms is the result of duplication, mutation and other mechanisms. Genes are commonly found in different species as homologues (corresponding genes in different species).

Gene families

Gene duplication with neofunctionalization of one gene copy creates two closely related genes whose DNA is almost identical. Further duplications of the two initial copies and further neofunctionalization increase the size of the gene family. Given time large gene families encompassing very differing functionality may arise. An example is the G protein-coupled receptor (GPCR) gene family. With around 800 human family members it is one of the largest gene families. They all share a seven transmembrane helix structure and their general function is to mediate signals from the outside of the cell to the inside. GPCRs can be found in a wide variety of organisms including plants, yeast, insects and vertebrates. It remains to be determined if all GPCRs share a common ancestor, however there are subgroups that clearly share the same ancestor (Fredriksson and Schioth, 2005).
Sequence data

Genome sequencing projects increase the amount of genomic sequence data at an exponential rate. The first complete microbial genome sequence, Hoemophilus Influenzae was released in 1995 (Fleischmann, et al., 1995). Currently there are 176 eukaryotic genome sequencing projects in progress with 23 complete and 124 in assembly stage according to (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi). The vast amount of sequence data being made available to researchers has increased the need for automation i.e. computer science. Sequence data is made publicly accessible from online databases such as The National Center for Biotechnology Information (NCBI) GenBenk (Bilofsky, et al., 1986) and The European Molecular Biology Laboratory (EMBL) (Hamm and Cameron, 1986).

Searching the sequence data for sequence similarity is made possible by the development of software such as the basic local alignment search tool (BLAST) (Altschul, et al., 1990), the BLAST like alignment tool (BLAT) (Kent, 2002) and HMMER (Krogh, et al., 1994) etc.

The availability of sequence data and the possibility to perform similarity searches has made it possible to quickly and accurately predict the existence and position of novel gene family members in species with sequenced genomes. Successful identification of the position of a gene in the genome does not however give instant access to the coding regions of the gene or information on possible splice variants. This information may however be found using other sources of sequence data.

Expressed sequence tags

Genes encoded in DNA are expressed in different amounts in different cells, tissues and organs. In eukaryotes the genes are transcribed into precursor messenger RNA (mRNA) which is in turn spliced to remove introns. The remaining mRNA is used as a template during a process called translation that synthesizes a protein. Because the spliced mRNA found in a cell contains only the information needed for translation and not intra-gene regions or introns it can be used to identify gene coding regions in the genome. To produce ESTs the mRNA is converted into complementary DNA (cDNA) to allow sequencing. The cDNA is sequenced either randomly or directionally to produce a single sequence read called expressed sequence tag (EST). ESTs are short sub-sequences of transcribed spliced nucleotide sequence typically 300-700 nucleotides long (Adams, et al., 1991). 5’ESTs usually contains parts of the protein coding sequence while the 3’ESTs are more likely to contain only untranslated regions (UTRs) of the gene.

ESTs are stored in databases such as NCBI dbEST which currently contains data from 1263 species with over 7.8 million human sequences (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).
The quality of EST data is highly variable due to the different sequencing techniques used to produce them and also because they originate from different cell types or tissues. EST sequences have a relatively high base calling error, 1-3% (Hillier, et al., 1996). Besides sequencing errors there are other sources that complicate the use of ESTs, genomic contamination i.e. the EST contains DNA from either the originating species or other sources, unspliced or incompletely spliced mRNA result in ESTs with little or no information on splice sites. Furthermore the number of ESTs in the database would be relative to the amount of mRNA found in the cell or tissue. Abundantly expressed genes would be overrepresented while genes with low expression would be scarce. However current techniques such as normalization or subtraction are applied to compensate for differences in expression levels (Marchtin and Pardee, 2000).

Sequence alignment
Sequence alignments are used to determine regions of DNA, RNA or protein sequences that are similar due to evolutionary, functional or structural relationships. A pair of sequences is typically aligned by inserting gaps in the sequences so that identical characters occupy the same position in both sequences. The resulting pair-wise alignment is presented as a two row matrix with the identical characters in columns. Pair-wise alignments in turn can be used to create a multiple alignment using progressive alignment construction. For a number of sequences all possible pair-wise alignments are created. The pair-wise alignments are used to create a distance matrix from which a distance based tree is created. The branching order of the tree is then used to determine the order in which the sequences are added to the multiple alignment. The sequences with the shortest pair-wise distance are aligned first and are thereafter considered as one so that gaps created between them cannot be deleted from the final multiple alignment. The multiple alignment is extended by repeating the step of adding the sequences with the shortest pair-wise distance according to the guide tree to the alignment. Score matrices optimized for different evolutionary distances are used to aid the alignment process. Multiple alignments are used in many facets of bioinformatics and software such as ClustalW (Thompson, et al., 1994) and T-Coffee (Notredame, et al., 2000) is used extensively.
Phylogeny

Phylogeny is the study of relationships. It aims to determine the relationships among species, molecules or traits. Phylogeny previously dealt with morphological features like size, color, number of legs etc. Current studies primarily rely on other sources of information such as DNA, RNA and protein sequences e.g. genes. The degree of conservation between the genes is used to compare them. A high degree of conservation indicates that a short evolutionary time has passed since the genes diverged. Molecular phylogeny is a key method for determining the orthologue/parologue relationship of genes which is crucial for the early prediction of the putative role of a newly discovered gene. The evolutionary history of the genes is visualized as a tree with the leafs representing the extant species or genes (also known as operational taxonomical unit OTU). The internal nodes of the tree represent the ancestral species or genes of the OTUs (also known as last common ancestor LCA).

There are a number of different tree representations. The tree may be rooted or unrooted. In rooted trees an internal node is selected by some criterion to be the ancestor of the other nodes in the tree and consequently used as the root of the tree. In an unrooted tree the internal relationship is shown without indicating which of the internal nodes that appeared first. The trees may be bi or multifurcating. In the multifurcating case the internal nodes are allowed to have more than two child nodes while in the bifurcating case the number of children is exactly two for each internal node. The tree may contain branch lengths. Branch-lengths indicate the amount of time that has passed since each branching event. Assuming a molecular clock i.e. a constant rate of evolutionary progress a tree with branch lengths could be presented along a time axis. The tree with branch lengths and the time axis would then present us the time the different genes appeared and the time separating them.

The number of possible trees

Being able to infer the phylogenetic tree that best explains the observed molecular data is based on the assumption that we can tell trees apart. A criterion is needed to evaluate the different trees in order to find the tree that score highest according to the criterion. Given such a criterion finding the optimal phylogenetic tree would simply involve generating all possible trees and evaluating them according to the criterion. However, the number of possible rooted, labeled binary trees for a given number of leafs is very large.
\[
\frac{(2n - 3)!}{(2^{n-1}(n-1)!)}
\]

*Equation 1.* The number of possible rooted labeled binary trees. \( n \) is the number of taxa included in the analysis.

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Number of possible trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>10395</td>
</tr>
<tr>
<td>10</td>
<td>34459425</td>
</tr>
<tr>
<td>11</td>
<td>654729075</td>
</tr>
<tr>
<td>12</td>
<td>13749310575</td>
</tr>
<tr>
<td>13</td>
<td>316234143225</td>
</tr>
<tr>
<td>14</td>
<td>7905853580625</td>
</tr>
</tbody>
</table>

The large number of possible trees prohibits an exhaustive search for the tree that best explains the observed molecular data. The practical limit for evaluating all possible trees is roughly 12 taxa. With increasing computational power provided by faster computers the practical limit will move upwards but ever so slowly. Molecular phylogenetic reconstruction is a difficult combinatorial problem and it is in fact non-polynomial (NP) complete (Cook, 1971; Day, 1987). The brute force approach could be used in cases with very few taxa but if we wish to solve larger problems we need to employ other techniques.
Inferring phylogenetic trees

Numerous methods have been proposed to solve the problem of inferring phylogenetic trees from sequence data. They can be broadly divided into two categories, distance and character based. Both methods use sequence data in the form of a multiple alignment to infer trees.

Distance based phylogenetic inference

Distance based methods use pairwise distances to infer trees. The distances are commonly based on a multiple alignment. A pairwise distance matrix is calculated by some metric often including a score matrix. The general idea is to group sequences on their pairwise evolutionary distance thus completely sidestepping the tedious work of examining all possible topologies that could explain the observed data. The sequences with the shortest distance are grouped together and their entries in the distance matrix are removed to be replaced by the node that connects them together. The distance matrix is recalculated with the new node as a mean of the nodes it connects. Distance based methods include unweighted pair group method with arithmetic mean (UPGMA), neighbor-joining (NJ) (see Figure 1) etc. (Saitou and Nei, 1987; Sneath and Snokal, 1973).

![Neighbor joining (NJ)](http://www.icp.ucl.ac.be/~opperd/private/neighbor.html)

Figure 1. a) All OTUs are placed at equal distance to each other in a star shaped tree. The data in the distance matrix (not shown) is used to calculate the net divergence for each OTU. The net divergence is then used to calculate a new distance matrix that is used to choose the two OTUs with the smallest distance, in this case A and C. b) A new internal node U is created, the distances from A and C to U are calculated using the net divergence. The distance matrix is updated i.e. A and C are removed and U is inserted, the remaining distances are updated. c) The procedure is repeated until all distances are removed from the matrix. Figure modified from (http://www.icp.ucl.ac.be/~opperd/private/neighbor.html).
Character based phylogenetic inference

Character based methods in contrast to the distance based methods map the history of the observed sequence data onto a proposed tree. The mapping is then evaluated according to some criterion. For the optimal mapping to be found all possible topologies with all possible mappings need to be evaluated. Due to the high number of possible topologies and mappings an exhaustive evaluation is prohibited for more than 12 taxa. Finding an answer in reasonable time requires settling for a possibly suboptimal solution.

Figure 2. The multiple alignment show four genes and the short sequences that represent them. For four OTUs there are three possible unrooted trees. With such a small number of possible trees an exhaustive search can be performed. The three possible trees are shown together with the sequences that would be observed at the internal nodes. For each of the possible trees the number of mutations (m) is calculated. The tree with the lowest number of mutations, in this case $m = 4$ is the most parsimonious solution. Figure modified from (http://www.icp.ucl.ac.be/~opperd/private/parsimony.html).
There are a number of criterions proposed for evaluating the proposed trees, among them maximum parsimony (MP), maximum likelihood (ML) and Bayesian methods are commonly used (Felsenstein, 1982; Fitch, 1971; Ronquist and Huelsenbeck, 2003; Strimmer and Von Haeseler, 1996).

MP seeks to minimize the number of evolutionary events i.e. mutations, insertions or deletions that are needed to explain the observed sequence data. The sequences are mapped onto a tree and each column in the multiple alignment is evaluated (see Figure 2). Only informative sites in the alignment contribute to the MP analysis. Sites that are conserved among all sequences do not contain any information that can be used to determine their relationship. Protpars from the Phylip (Felsenstein, 2005) software package and Tree analysis using New Technology (TNT) (Goloboff, 1999) are commonly used to infer MP. Weighted MP also seeks to minimize the number of evolutionary events, however the events are not ranked equally as in unweighted MP. An evolutionary model is used to weight the transitions from one amino acid to another. Changes that have a small impact on the properties of a protein are weighted lighter than changes that radically alter the properties.

ML evaluates the likelihood that a proposed tree would yield the characters that we observe. Each column in the multiple alignment is mapped to a tree and the likelihood is calculated. The likelihood of the tree is the product of the likelihood of each column. The likelihood of an individual column is based on an evolutionary model. Each substitution has a certain probability. All possible combinations of nucleotides or amino acids at the internal nodes are tried and the resulting likelihoods are summed together (see Figure 3). DnaML and ProML from the Phylip (Felsenstein, 2005) software package,
PhyML (Guindon and Gascuel, 2003), TreePuzzle (Strimmer and Von Haeseler, 1996), GA-mt (Katoh, et al., 2001) and GAML (Brauer, et al., 2002) are a small selection of the software available for ML inference.

Searching for trees

The large amount of possible trees prohibits an exhaustive search for the optimal solution. We are forced to rely on heuristic search methods for finding a good enough solution. The general approach to finding the solution is to start with a random tree, perform small changes on it and evaluate the resulting trees to see if an improvement has been made. If improvement has been made we continue making small changes to the improved tree until no further improvements can be made and a local optimum has been found. There is however no guarantee that the local optimum is the global optimum. This heuristic is called hill-climbing as the algorithm always travels up towards the highest increase in score.

The small changes to the tree can be made by a number of different mechanisms, nearest neighbor interchange (NNI), subtree prune and regraft (SPR), tree bisection and reconnection (TBR) or tree-fusing (Goloboff, 1999) etc.

A NNI operation is performed by removing the internal branch that connects four subtrees. The four subtrees are the recombined in one of the three possible ways (see Figure 4).

SPR operations involve pruning a subtree from the tree and the removal of the internal branch that attached the subtree to the tree. The pruned subtree is then regrafted at any of the internal branches in the tree (see Figure 5).

A TBR operations starts by bisecting a tree into two subtrees, the internal branches in both subtrees that connected the two together are removed. The two subtrees can then be reconnected by forming a new internal branch between any of the internal branches in the two subtrees (see Figure 6).
Figure 4. a) An unrooted tree consisting of four subtrees with the internal branch that connects them. b) The internal branch is dissolved. c) The internal branch is recreated in two different ways creating the two possible NNI neighbors of the tree in a). Figure modified from (Felsenstein, 2004).

Figure 5. a) An unrooted tree with seven OTUs. b) An internal branch is dissolved i.e. the subtree containing E, F and G is pruned. c) The surplus internal branch in the upper tree is removed. d) The arrows indicate the positions in the upper tree that the lower subtree could be regrafted to. The lower subtree could be regrafted to the middle section of the upper tree but that would result in a tree identical to a). Figure modified from (Felsenstein, 2004).

Figure 6. a) An unrooted tree with seven OTUs. b) An internal branch is dissolved i.e. the two subtrees are bisected. c) The surplus internal nodes in both subtrees is removed. d) One of the possible reconnections is shown. All internal branches in both subtrees can be reconnected with each other, allowing for a very large number of possible trees. Figure modified from (Felsenstein, 2004).
Genetic algorithms

Genetic algorithms (GA) (Holland, 1975) mimic the adaptations of evolving populations by including processes analogous to mutation, recombination and selection to solve real world problems (Davis, 1991; Mitchell, 1996). Each individual in the evolving population is a potential solution to a problem and the individual’s fitness is determined by how well it solves the problem.

GAs have been used to find near optimal solutions to NP-hard problems in many different areas. Finding solutions to the static problem of scheduling times and locations for student exams (Burke and Newall, 1999), as well as the dynamic scheduling needed by the industry to manage the ever changing conditions of hardware failure, changed delivery dates and other unforeseen events (Jensen, 2003). GAs have been applied to biological problems such as multi-class prediction for the analysis of gene expression data (Oui and Patrick, 2003). GAs are extremely well suited to NP-hard problems and many GA applications have been very successful in such areas.

GAs have been applied to phylogenetic inference (Goloboff, 1999; Katoh et al., 2001; Lemmon and Milinkovitch, 2002; Lewis, 1998; Matsuda, 1996; Nixon, 1999). In the phylogenetic implementation of the GA an individual is usually represented by a tree topology. Like other heuristic search methods, GAs does not guarantee finding the optimal solution. The parameters chosen for the GA affect the extent of the search space searched and the efficacy of the search.

The GA used to search for the most parsimonious weighted tree is initialized with a single population of n individual rooted binary trees with random topologies. The fitness of each individual is the weighted parsimony score of the tree. The individual fitness is the basis for the rank ordering of the population. Being able to rank the population is vital to the function of the GA. The selection scheme that increases the fitness over time relies on the ranking of individuals so that more fit individuals have a higher change of parenting the next generation.

The probability of any given individual being selected to contribute to the next generation is determined by is rank. The probability of leaving an offspring in the next generation is defined to be $i/\Sigma n$, where $i$ is the position of the individual in the sorted rank list and $n$ is the number of individuals in the population. This is known as rank roulette selection (Mitchell, 1996). The rank roulette selection is used to ensure that regardless of the difference in fitness the probability of selection is based on the rank and not the particular fitness. This means that the individual with the lowest fitness will have probability $1/\Sigma n$ of being selected and the highest fitness will have probability $n/\Sigma n$.

Elitism is used to ensure that the individuals with highest fitness remain unchanged in the population. The new generation is extended with a fixed
number of unchanged individuals from the parental generation chosen based on their high rank.

A mutation is performed by randomly changing the topology of the tree. Two subtrees are selected randomly and exchanged for each other. With luck the change will improve the fitness of the individual. All individuals in the offspring generation except those selected using elitism are subject to the possibility of mutation or recombination i.e. not all individuals in the offspring generation are necessarily topologically different from their parents. In a recombination the characteristics of two individuals are mixed to create a new individual. One individual is chosen from the parental generation using rank roulette selection. A copy of the individual is inserted into the offspring generation. Another individual is chosen whose characteristics will be mixed with the individual in the offspring generation. A subtree with three or more OTUs in the offspring individual is randomly selected and pruned. The pruned subtree is recursively scanned and all OTUs are registered. The individual from the parental generation is scanned recursively and as the OTUs listed from the pruned subtree are located they are added to a new subtree. The order of the OTUs in the parental tree determines the order of the OTUs in the subtree that is inserted into the offspring tree at the position where the old subtree was removed. Recombination is an important difference between GAs and other heuristic search methods as it allows the sharing of potentially good portions of two distinct trees.
Comparing phylogenetic trees

Researchers are often presented with trees that are topologically different even if they represent the same sequence data. This can be attributed to a number of factors, different inference methods, different evolutionary models used, varying the parameters for the heuristic search etc. Determining the distance between different topologies becomes necessary to evaluate the trees. Do the differing methods and parameters agree on how the topologies should look?

Phylogenetic trees can be compared using tree distances. An number of different tree distances have been proposed, the symmetric difference (Robinson and Foulds, 1981), the branch score distance (Kuhner and Felsenstein, 1994), NNI, SP R and TBR. The symmetric difference and the branch score distance are similar in that they compare the two trees by identifying the branches that are not present in both trees.

The distance between two trees can instead be defined as the number of local rearrangements needed to transform one tree into the other. A smaller number of operations separating two trees would mean that they are closer together. The local rearrangements used are the same that are used to heuristically search for MP and ML trees, NNI, SP R and TBR. The SP R operation offers an advantage due to its ability to model complex biological events such as horizontal gene transfer (HGT) and recombination (Allen and Steel, 2001; McFadden and Gilson, 1995; Stahl, 1987; Syvanen, 1985).

The SP R distance

The complexity involved in calculating the minimum SP R distance remains to be determined, however it has been conjectured that the parameterized SP R distance is fixed parameter tractable (FPT) (Allen and Steel, 2001). The time to determine the SP R distance would be dependent on the distance rather than the number of OTUs in the trees. To accomplish this, the trees are reduced, i.e. the sections of the trees that are identical are removed and replaced by markers allowing them to be identified. This means that the sections of the trees that are not needed to determine the SP R distance are removed without affecting the SP R distance. The reduction is done by repeatedly applying the following rule defined in (Allen and Steel, 2001).

- Replace any pendant subtree that occurs identically in both trees by a single leaf with a new label

The SP R neighborhood of a tree is defined to be all possible trees that can be created by performing one SP R operation on a specific tree excluding those that recreate the original tree. The SP R distance is found by applying the above rule on both trees until no further reductions are made. One of the
trees is selected to be the starting point the other tree to be the reference tree. The SPR neighborhood of the starting point tree is created. All neighborhood trees are compared to the reference tree and if one of them is found to be identical the SPR distance have been determined to be one. If none of the neighborhood is identical to the reference tree they are all subjected to a round of reduction using the above rules. The neighborhood trees are then sorted based on the number of reduced OTUs, with the most reductions placed first. Starting with the most reduced neighbor the procedure is repeated, i.e. all neighbors are created and evaluated. The repetitions continue until a solution is found (see Figure 7). Once a solution has been found an upper bound can be applied to the remaining calculations, i.e. solutions that require higher or equal amounts of SPR operations are not interesting, we need only consider solutions that can improve on the already found solution. Still all solutions that could possibly improve on the found solution needs to be evaluated and this is what consume the bulk of the time spent searching for the minimum SPR distance.

Figure 7. Calculating the minimum SPR distance between a) and d). The sections of the trees that are identical are reduced to avoid unnecessary calculations, resulting in trees b) and c). Tree b) is then used as starting point. All subtrees are pruned and regrafted in all possible positions not resulting in a tree identical to b). Some of the resulting neighbors are shown as smaller trees. The neighbors are compared to the goal tree and if any of them is identical to the goal tree the solution is found. If not the neighbors are scanned and all sections that are identical to the goal tree d) are reduced. The trees are sorted based on the number of reduced nodes. In this case the encircled tree can be reduced the most as the subtree containing A and E can be found in c). The tree with the highest number of reduced nodes is more likely to generate a solution among its neighbors and is placed first among the sorted neighbors. The procedure is repeated with the first neighbor in the sorted list as the starting point. In this case the desired tree is found by moving B from the subtree A, E and B to the FG subtree. Two SPR operations is the minimum distance between a) and d) with the encircled trees as intermediate tree.
Trees within trees

The evolution of genes is linked to the organism that they reside within in the same way as the organisms are linked to the geographical areas they reside in. A number of genes sampled from a set of species have a direct connection to the evolution of the species (see Figure 8).

Figure 8. The evolutionary trees of genes are associated with the species in which they have evolved. a) a tree representing the evolution of three species. Contained within the tree is the path of gene and its evolution. There have been a number of duplications and a sorting event, i.e. a gene has been silenced. b) a tree showing the relationship between the a geographical area and an initial specie. As the geographical area divides a speciation events occurs when the two populations become geographically separated. Further speciation occurs with further subdivision of the geographical area. Figure modified from (Page and Charleston, 1998).

Figure 9. Duplication events may confuse the process of inferring a species tree from a gene tree. Should only one of the duplicated genes be sampled from each species, i.e. gene S from specie v and z and gene T from specie x, the inferred species tree would be the incorrect tree seen in b). c) A sorting event and a subsequent HGT might in a similar fashion affect the inferred species tree. A gene is sorted after the speciation creating x and y. The sorted gene is then reintroduced from z by HGT. The inferred species tree would again be the incorrect tree seen in b). Figure modified from (Page and Charleston, 1998).
That does not however mean that the phylogenetic trees of the gene agree with the species tree. The gene may have undergone duplication, recombination, HGT or other evolutionary mechanisms that would produce a gene tree that is incongruent with the species tree (see Figure 9).

Inferring the relationship between genes and species or species and geographical areas may be complex. Acquiring sequence data from the interesting species may be less than trivial. Difficulties creating reliable phylogenetic trees given complex sequence data further complicate matters. Processes like recombination and HGT often result in incongruent gene trees. A number of methods have been proposed to infer the species tree from gene trees.

- Creating multiple alignment based on individual genes found in the studied species. The alignments are then concatenated into a large alignment from which a phylogenetic tree is inferred using standard phylogenetic techniques (Gadagkar, et al., 2005).
- Bayesian analysis with a Markov Chain Monte Carlo method incorporating duplications and extinctions (Arvestad, et al., 2003).
- Reconciled trees which minimize the number of duplications needed to reconcile gene trees with a species tree (Charleston, 1998; Page, 1998; Page and Charleston, 1998).
- Uninode coding which infers species trees by analyze paralogues (Simmons, et al., 2000).

We propose that the median SPR consensus tree describe below can be interpreted as a species tree when applied to gene trees. The median SPR consensus tree explains the observed gene trees with a minimum number of recombination and/or HGT events.

**Median SPR consensus**

The median SPR consensus tree is defined to be the tree that has the minimum combined SPR distance to a set of trees sharing the same OTUs. The median SPR consensus tree may be one of the trees in the set or another tree with the same set of OTUs. Finding the median SPR consensus tree involves creating all possible trees for a given set of OTUs and determining the pairwise SPR distance for each possible tree to all trees in the original set. Creating all possible trees is prohibited by the rapid increase in possible trees with an increase in OTUs (see Equation 1). An algorithm for finding an approximation to the median SPR consensus tree is therefore suggested.

An approximated median SPR tree can be found using the algorithm described above for calculating the minimum SPR distance (see Figure 10). The pairwise distance between all trees in the set is calculated. The distances are stored in a matrix where the sum of each column is the combined SPR
distance to all trees in the set. For each pairwise distance larger than one there are a number of intermediate trees that connect the pair of trees through SPR operations. If the pairwise distance is three there are two intermediate trees. For each intermediate tree a column is added to the SPR distance matrix. The pairwise SPR distance from each intermediate tree to the trees in the set are calculated and entered into the matrix. When all intermediate trees are evaluated the distances in each column is summed and the tree with the smallest combined distance is considered the median SPR consensus tree.

Figure 10. An approximation algorithm for the calculation of the median SPR consensus tree, i.e. the tree that given a set of trees that share the same OTUs explains the observed trees with a minimum number of SPR operations. In this case the set of trees is A through F. a) The algorithm starts by calculating the pairwise minimum SPR distance between all trees in the set. a) shown here is the minimum SPR calculation between B and C with a distance of 2 and an intermediate tree named BC1. The intermediate trees from each calculation are stored for future use. The resulting SPR distances are stored in the matrix enclosed by a unbroken line. b) For each of the intermediate trees the distance to each of the trees in the original set A – F the minimum SPR distance is calculated. The SPR distances are stored in the matrix enclosed by a dotted line. Here only a small selection of the intermediate trees is shown in the matrix. c) The columns in the two matrices are then summed together to determine which tree has the minimum SPR distance to the other trees. In this case AC2 would be the median SPR consensus tree.
Research aims

The aims of this thesis were to:

- Develop and evaluate a genetic algorithm to search for the most parsimonious weighted rooted binary tree (Paper I)

- Develop an algorithm that determines the minimum SPR distance between two unrooted binary trees. Evaluate the algorithm with respect to performance and compare it to other tree distance measures such as Robinson-Foulds (Paper II)

- Define the median SPR consensus tree and develop an algorithm to calculate an approximation to it using the minimum SPR algorithm developed in Paper II (Paper II)

- Cluster ESTs using the genome as template and compare the results to clustering techniques using pairwise alignments for clustering (Paper IV)
Results and discussion

Paper I

We developed an algorithm that searches for the most parsimonious weighted rooted binary tree using a genetic algorithm. The genetic algorithm used is affected by a number of parameters. The size of the population, the number of trees that remains unchanged from one generation to the next, the location of the topological changes and the probability of a specific tree being mutated or recombined. These parameters were evaluated with regard to the fitness value reach and the time required reaching it. One parameter at a time was varied with the others fixed. The individual runs were terminated based on the increase in fitness per generation.

The results show that larger population increases the fitness at an increased time to calculation termination. The number of individuals that remain unchanged from one generation to the next should be between 30 and 70 percent to reach the highest fitness. Mutation and recombination probabilities between 50 and 70 percent showed the fastest increase in fitness.

To test the ability of the algorithm to infer a tree from a real world example we selected a set of human adhesion like G protein-coupled receptors (Fredriksson, et al., 2003). The data set was analyzed using software from the Phylip package (Felsenstein, 2005) as well as the genetic algorithm. The results show that seven main groups can be distinguished within all four trees. There are minor differences within the main groups, however these differences do not single out the genetic algorithm but are evenly distributed.

The genetic algorithm is not guaranteed to return the same result from two consecutive runs. This could be considered a strength as well as a weakness. The strength lies with the ability to find a number of differing highly scoring topologies that explains the observed data. If a number of differing topologies scores equally this indicates that other means of analysis are necessary to determine the relationship. The non deterministic behavior could be considered a weakness in cases where one is only interested in the single highest scoring tree found by for example hill climbing techniques.
Paper II

We developed an algorithm that determines the minimum number of subtree prune (SPR) and regraft operations separating unrooted binary trees. The impact of the number of leafs in the trees and the number of SPR operations separating them was evaluated using both real world data (Thollesson, 2000) as well as artificially constructed trees. The number of SPR operations separating the trees has a major impact on the time required to find the minimum solution while the number of taxa only seem to influence the time needed when the SPR distance increases. The time required to find the minimum solution varies greatly, with 5 SPR operations and 25 leafs in the trees the time varies from 4 seconds to 139000 seconds with a median of 2006 seconds. The artificial as well as the real world data show that the time needed increases with an increased SPR distance. However the time required to find the minimum distance to the real world trees was on average significantly shorter. This observation could be explained by a less random behavior in the real world data. Simulated trees are created by randomly pruning sub-trees and regrafting them at randomly selected locations in the tree. This does not seem to be the behavior of real world data nor is it expected to be. Simulating trees with real life characteristic SPR operations require further studies.

Tree metrics can be used to evaluate the performance of phylogenetic inference methods. If the trees created by the methods are close, i.e. separated by a small number of SPR operations the methods can be said to be in relative agreement on the topology. We compared the SPR distance the Robinson-Foulds (Robinson and Foulds, 1981) distance. In (Thollesson, 2000) a number of different maximum parsimony weighting schemes where used to produce a set of trees. The Robinson-Foulds and SPR pair-wise distances between the trees in the set were calculated. The two methods generally agree on the distances with a few exceptions.

The Robinson-Foulds distance has the advantage of being relatively easy to calculate, it also allows for comparison of non-binary trees. Calculating the SPR distance on the other hand may be very costly and the current implementation only allows binary trees. The two distances differ in another important aspect. The SPR distance shows the intermediate solutions that connect the two trees. This allows us to trace the individual SPR operations and perform further analysis on the intermediate trees.

Paper III

Here we introduce the median SPR consensus tree and its interpretation as a species tree. Using the algorithm developed in paper II we search for a tree that explains a set of trees with a minimum of SPR operations. Finding the
minimum SPR operations separating two trees can be very time consuming. To investigate if we could decrease the time spent without major impact on the outcome we observed at which time during calculations that the minimum and second best solution were found. Results show that it is very likely to find the minimum or second best solution within the first 20 seconds of calculation.

In (Boussau, et al., 2004) a α-proteobacteria species tree was inferred using a conserved section of the genome. This species tree was used as a reference to test the ability of the median SPR consensus approach to infer a species tree from incongruent gene trees. The results show that the species trees inferred from the conserved region of genome and the incongruent gene trees agree on the evolution of the species.

The median SPR consensus method interpretation as a species trees does require that all gene trees included in the analysis have the same set of OTUs. This limits the applications of the algorithm as biological data is often incomplete i.e. a specific gene may not be present in one of the species included in the study and thereby disqualifying the data from the other species regarding that gene from being included. This limitation could be remedied with a generalized definition of the SPR operation to allow arbitrary sets of OTUs.

Paper IV

Human ESTs from dbEST were used to investigate how EST clustering performs when the genome is used as template compared to pairwise sequence alignment. The Genomic EST Clustering (GEC) method was developed to perform the genomic template clustering. An important feature when clustering EST is the number of disjunct clusters that result from the same transcriptional unit, i.e. each gene should be represented by a minimum number of clusters. Using the GEC method resulted in 80.4 % of the RefSeq genes represented by a single EST cluster compared to 25 % using the UniGene clustering method. The UniGene2 dataset was released in January 2005 and it is based on a method similar to GEC with additional information from mRNA and cDNA (http://www.ncbi.nlm.nih.gov/UniGene/help.cgi?item=build2). The UniGene2 dataset significantly reduced the number of resulting clusters compared to both UniGene and the GEC. The breath of expression of the EST was estimated using the 18145 sequences prefixed NM in RefSeq. 97.4 % or 17678 of the sequences found in RefSeq were considered matches against the genome.
Conclusions

Paper I: We showed that a GA with weighted MP criterion as fitness measure can be used to infer phylogenetic trees from amino acid sequences.

Paper II: We developed an algorithm and implemented software based on it that determines the minimum SPR distances between unrooted binary trees.

Paper III: We defined the median SPR consensus tree. We developed software to calculate an approximation to the median SPR consensus tree and showed that the approximation can be interpreted as a species tree based on incongruent gene trees.

Paper IV: The position of ESTs on the human genome was determined using BLAT. The positions were used to cluster the ESTs based on positional overlap. We showed that clustering ESTs based on positional data outperforms EST clustering based on pairwise alignments with respect to the number of disjunct cluster originating from a transcriptional unit.
Future perspectives

We have studied a GA and its ability to search for a weighted MP tree based on protein sequence data. The GA was able to infer phylogeny on par with other commonly used software. There are however a number of ways that we could possibly increase the current performance and accuracy of the GA.

GAs are prime candidates for parallelization. The tasks performed in the GA are inherently parallel. Individuals or entire populations could be handled as individual objects. In conjunction with the development of CPUs with the ability to perform calculations in parallel this provides an opportunity to greatly increase the performance of the phylogenetic inference.

The MetaPIGA (Lemmon and Milinkovitch, 2002) algorithm with several populations evolving side by side with the ability to exchange traits between could further increase the efficiency of the GA.

We have developed and studied an algorithm that calculates the minimum SPR distance between unrooted binary trees. The definition of an SPR operation states that a subtree may be pruned anywhere in the tree and subsequently regrafted anywhere in the tree. The number of SPR operations the separate two trees does not depend on the distance between points of pruning and regrafting within the tree. It would be interesting to evaluate the behavior of an extended definition of the SPR operation that does take distance into account.

Further extending the SPR definition to include trees with differing sets of OTUs would increase the ability of the algorithm to handle the often incomplete nature of biological data.

We have studied the median SPR consensus tree and its interpretation as a species tree. The median SPR consensus tree is the tree with the minimum SPR distance to all trees in a set sharing the same OTUs. The median SPR consensus tree will depend on the distribution of topologies in the set of trees. If a large proportion of trees have identical or highly similar topologies this will force the median SPR consensus tree topology towards that topology. If there are several trees that share the same topology they could be replaced by a single tree to avoid lengthy calculations and undue influence over the outcome. It would be interesting to incorporate a clustering technique into the algorithm. Trees that have nearly identical topology could be grouped into a cluster and a median SPR consensus tree could be calculated.
for the cluster. The group of nearly identical trees would then be replaced by the median SPR consensus tree representing the group. This could potentially shorten the calculation time as well as avoid biasing the input data.
Acknowledgements

I would like to thank past and present members of Helgi Schiöth, Dan Larhammar and Lars Oreland’s groups at the Department of Neuroscience for their friendship and imparted knowledge.

I would like to thank:

Helgi Schiöth, for giving me the opportunity to work in your group and pushing me on.
Robert Fredriksson, for always taking the time to hear my questions and almost infallibly providing me with an answer.
Mikael Thollesson, for great ideas and relevant critique.
Dan Larhammar, for interesting courses and discussions.

Everyone who contributed to this thesis.

Colleagues and students, Andor, for all the hard work and friendship. Tommy and Andreas, for friendship and showing me how coffee really should be drunk: excessively. Christian, for the lessons in computer science and countless late night sessions. Thomas, for great discussions and the introduction to the space elevator. David, for discussions regarding the large, small and sometimes theological. Tomas, for teaching me about evolution and Finspång and for being a great friend. Malin, for all the crazy stories. Kalle, for showing me how real sleeping is done. Paula, for great talks and night swimming. Tatjana, for the flowers. Torun, for the best pictures taken at BMC. Hille, for your great sense of humor and your ability to see beyond the blubber. Helena, for always talking first. Christina, for not bruising easily. Görel, for sharing my experience of teaching the physicians to be. Susanne, for putting up with me disturbing the (relative) peace. Hanna-Linn, for the energy. Josefin, for still being here after sharing a room with me. Aneta, for all the laughs. Bengt for having us at his island.

Past and present members of Helgilab, especially, Thora, Linn, Johan, Majd, Jenny and Smita. Thanks to members of Danlab and Orelab especially, Cecilia, Jenny, Camilla, Theresa and Earl.

Ulla, Birgitta, Marita and Lena for help and shelter from the bureaucracy.
My friends and family, for support and reminding me that there are things beside BMC. For patiently listening to my work related ravings and not slapping me during the last years.

Jessica, my love.
References


A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)