Using a supertree approach to detect laterally transferred genes within Staphylococcus

Author: Dave Tang
Student number: 40043889
Email: s4004388@student.uq.edu.au

Principal Supervisor: Professor Mark Ragan
Associate Supervisor: Dr. Robert Beiko

Abstract

In a preliminary analysis, Kuroda and colleagues (2001) identified many putative lateral gene transfer (LGT) cases in *Staphylococcus* based on BLAST hits. The validation of LGT used in this study is based on the conflicting phylogenies of orthologous proteins and protein-coding genes to a reference phylogeny. The primary aims of this study are to identify relationships among orthologous proteins and protein-coding genes from 8 *Staphylococcus* isolates. These orthologous groups of proteins and protein coding genes can be used to infer a species history, which in turn can be used to identify possible cases of LGT.

A small subset of orthologous proteins and protein-coding genes were identified to be discordant with the reference topology of the 8 *Staphylococcus*. As there are other possible explanations of discordance, such as bad data and hidden paralogy, these other explanations were assessed before invoking LGT as the cause of discordance. However results showed that discordance was not related to these explanations.

Lastly there has been statistical support for putative LGT cases to be overrepresented by functions implicated in mobile and extrachromosomal elements and certain cellular processes, particularly pathogenicity-related genes. This is highly expected due to the known mechanisms of LGT in Staphylococcus.

Acknowledgements

The field of bioinformatics, especially in molecular evolution, was an extremely challenging task for myself. Without the people, whom I will shortly acknowledge, who have supported me throughout my Honours year I would not have been able to make any progress in understanding bioinformatics as a discipline and most importantly, I would not have been able to complete my thesis.

Firstly I would like to thank my principal supervisor, Prof. Mark Ragan, for offering me a position in his lab and giving me such a wonderful opportunity to learn about bioinformatics.

Secondly I would like to thank my associate supervisor, Dr. Robert Beiko, who despite my incompetence to understand the many disciplines in bioinformatics has been extremely patient and helpful towards my understanding about almost everything in bioinformatics and phylogenetics.

Thirdly, I have to thank Tim Harlow for offering me technical assistance with the computer side of things. Without his help, I would not have been able to complete much of the analyses presented within this thesis.

And lastly I would like to thank the people I bug around the lab asking for help with my perl scripts, which are CX and Simon. I would have asked Michael for help with my perl scripts but however he likes to use python.

Table of contents

Introduction

Page numbers

Methods

Result

Discussion

List of Figures

List of Tables

List of Abbreviations

General abbreviations

LGT = Lateral gene transfer
rRNA = ribosomal RNA
TIGR = The Institute for Genomic Research
NCBI = National Center for Biotechnology Information

Staphylococcus related abbreviations

S. aureus = *Staphylococcus aureus*
S. epidermidis = *Staphylococcus epidermidis*
SaPI = *S. aureus* pathogenicity islands
Mu50 = S.aureus_Mu50
MW2 = S.aureus_MW2
COL = S.aureus_COL
N315 = S.aureus_N315
ATCC = S.epidermis_ATCC12228
RP62A = S.epidermis_RP62A
MRSA252 = S.aureus_aureus_MRSA252
MSSA476 = S.aureus_aureus_MSSA476

Methodology abbreviations

BLAST = Basic local alignment search tool
RC/s = Representative cluster/s
MRC/s = Maximally Representative cluster/s
MRP = Matrix Representation with parsimony
dfit = Most similar supertree method
qfit = Maximum Quartet fit
sfit = Maximum Split fit
PP = posterior probability
ML = Maximum likelihood
NJ = Neighbour joining
MCMC = Monte Carlo Markov Chain
COG = Clusters of Orthologous Groups of proteins

Introduction

Prokaryotic evolution and lateral genetic transfer

Prokaryotes are small (cell size from 0.2-750 µm; genome size between 500 and 10,000 kilobases) unicellular organisms that propagate themselves primarily by binary fission. It was commonly assumed that prokaryotic gene transfer was mainly vertical (i.e. mother to daughter cells) due to their replication mechanism (Jain et al., 2002). Thus early work done on prokaryotic adaptation, evolution and speciation was mainly focused on clonality and periodic selection (Gogarten et al., 2002). Favourable mutations offering a selective advantage will accumulate and eventually fixate into the genome driving the evolution of prokaryotes (Morschhauser et al. 2000). Differential selection from environmental conditions will fixate different lineages into particular niches, thus diversifying prokaryotes.

To capture the diversity of prokaryotes, early taxonomic work was based on phenetic principles, which are relationships based on observable properties of organisms, was done (Jones and Sneath, 1970). Most of the taxonomic work done was based on phenetic relationships and this was the basis for describing bacterial phylogenies. The discovery that DNA and protein sequences could be interpreted as evolutionary documents (Zuckerkandl and Pauling, 1965) coupled with pioneering work done with ribosomal RNA (rRNA) by Carl Woese (Woese and Fox G, 1977) provided the first useful evolutionary classification of microbes as well as a universal tree of life (Eisen, 2000). The 16S rRNA tree is the gold standard for inferring phylogenetic relationships, especially in prokaryotes but has been the subject of much controversy. The main argument against the 16S rRNA tree is whether a single gene tree can represent the evolution of species (Doolittle, 1999) especially when more and more evidence is coming in from completed genome sequences for genes being transferred between organisms. If genes have been transferred across taxa instead of following a vertical transmission, then constructing a phylogeny of the genes does not give a picture of the organismal history. Some have even argued that no hierarchical universal classification can be constructed due to lateral gene transfer (LGT) (Doolittle, 1999). In order to construct an overview of how prokaryotes have diversified, it is necessary to understand the implications of LGT.

The mechanisms of lateral gene transfer

The processes that facilitate LGT have been established for quite some time (Sneath and Jones, 1970), and consist of three principal mechanisms: transformation, conjugation and transduction (Jain et al., 2002). Transformation is the process where a competent prokaryote takes up free DNA that is available in their immediate surrounding. This mechanism of gene transfer is restricted for prokaryotes that are not naturally competent, for example *S. aureus* are not naturally competent (Holden et al., 2004); however, if the prokaryote is competent, transformation has the potential to transmit DNA between very distantly related organisms (Ochman et al., 2000). Conjugation is the process that is equivalent to bacterial sex where a tube-like structure called a pilus joins a donor and an acceptor together allowing the transfer of genetic material through the pilus (Jain et al., 2002). DNA is transferred by either a self-transmissible or mobilization plasmid and is known to mediate transfers between domains, for example between bacteria and plants (Ochman et al., 2000). Transduction is the movement of genes via viruses and the process of transduction is limited by the ability of the virus to infect the organism. There are broad host range viruses that have the capability of facilitating the lateral movement of DNA across vast phylogenetic distances (Jain et al., 2002). Although these mechanisms of LGT are able to transfer DNA across vast evolutionary distances, they must be incorporated into the recipient's genome and only if it offers a selective advantage will be it maintained and kept (Ochman et al., 2000).

The *Staphylococcus*

In the *Staphylococcus* genus, there are at least 30 identified species, however among these, only 2 species are of clinical importance; *Staphylococcus aureus* (*S. aureus*) and *Staphylococcus epidermidis* (*S. epidermidis*) (Zhang et al., 2003). *S. aureus* is the most common cause of hospital-acquired infection, causing clinical disease in 2% of all patient admissions and is also becoming increasingly resistant to antibiotics (Lindsay and Holden, 2004). Besides their importance in the hospital, *S. aureus* isolates have also been identified in the general community causing a wide range of infections (Baba et al., 2002). *S. epidermidis* on the other hand is a normal inhabitant of human skin and mucous membranes that rarely causes infections in healthy individuals (Zhang et al., 2003). However over the years *S. epidermidis* has emerged as one of the major pathogens in hospitals (Zhang et al., 2003), often causing infections associated with implanted medical devices (Gill et al., 2005).

Currently 6 *S. aureus* genomes and 2 *S. epidermidis* genomes have been sequenced, giving a complete overview of the genome content of these isolates. Comparative genomics done amongst *S. aureus* (Lindsay and Holden, 2004) and among *S. aureus* and *S. epidermidis* (Gill et al., 2005) have shown a large core of genes to be consistent among the *Staphylococcus*, with a larger proportion of shared genes within strains of *S. aureus*. The remaining part of the genome is termed the accessory genome, and accounts for ~25% of any *S. aureus* genome.

Hiramatsu and colleagues (Hiramatsu et al., 2004) stated that genes in the core genome are inherited vertically and genes in the accessory region are acquired by LGT. This is plausible since the accessory genome consists of mobile (or once mobile) genetic elements, including bacteriophages, pathogenicity islands, chromosomal cassettes, genomic islands, plasmids and transposons, which transfer horizontally between strains (Lindsay and Holden, 2004). Mobile genetic elements are segments of DNA that encode enzymes and other proteins that mediate the movement of DNA within genomes or between bacterial cells (Frost et al., 2005). What is alarming is that within these regions are genes associated with virulence or resistance functions, thus the spread of these elements laterally amongst other organisms has clear clinical implications. For example the pathogenicity islands (SaPI), which carry superantigen genes implicated in toxic shock and food poisoning poise a clinical problem if transferred to other strains of bacteria (Lindsay and Holden, 2004).

LGT has played a vital role shaping the virulence and resistance of *Staphylococcus* (Gill et al., 2005) hence it is an important mechanism to understand. A phylogenetic approach is used in this study to identify laterally transferred genes within *Staphylococcus*. As most of the work done to characterise LGT in *Staphylococcus* is based on the methods stated above, using a phylogenetic method to infer LGT in *Staphylococcus* is a first. As phylogenetic reconstruction methods are the only reliable way to infer historical events from sequences (Eisen, 2000) and can be supported statistically, this approach is a novel and much more rigorous method than current methods used to detect LGT.

Detecting laterally transferred genes

Methods for detecting LGT can be broken into two main groups: surrogate methods and phylogenetic methods (Ragan, 2001a). Surrogate methods aim to detect genes that have an unexpected phyletic distribution or contain atypical sequence compositions. An unexpected phyletic distribution can be observed when a genome contains genes that are found in a phylogenetically distant organism, but are missing from more-similar species. Since bacteria have specific genomic signatures associated with their sequences, such as codon usage bias and GC content bias, regions with differing signatures can be proposed to have a foreign origin (Ochman et al., 2000). There are two disadvantages of surrogate methods; firstly there is a lack of consistency between surrogate methods of detecting putative LGT cases (Ragan, 2001a) and secondly phylogenetic conclusions shouldn't be made on observations of similarity due to its inappropriateness to phylogenetics (Doolittle, 1999).

A phylogenetic approach for detecting LGT is based on incongruence between different trees. A necessary prerequisite for phylogenetic approaches is the use of orthologues. Figure 1 illustrates 3 definitions; homologues are those that are related by a common ancestor; paralogues are related by gene duplications; and orthologues are related by speciation. Hence orthology can be placed into context with a tree since they both illustrate a vertical signal. If the tree of one orthologous family is topologically incongruent with that of a second, then the families have incompatible genealogical histories, possibly due to LGT (Ragan, 2001b). As phylogenetic trees take into account evolutionary signals from sequences and have well established methods of statistical confidence, such as the bootstrap, they offer a systematic approach to detecting LGT (Ragan, 2001a).

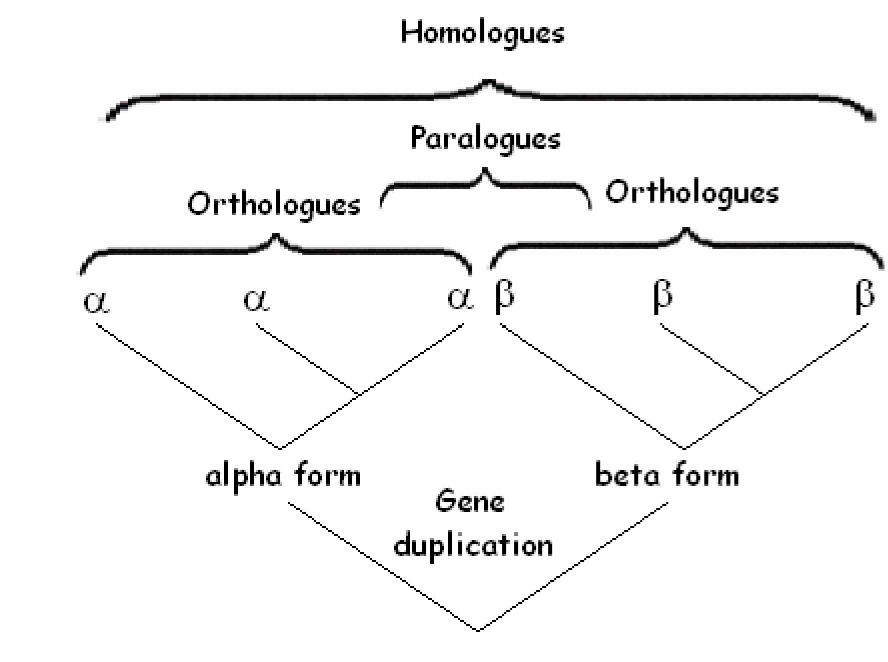

Figure 1. Illustration of homology, paralogy and orthology

Determining orthology

Generally if sequences have similarities that are statistically significant, they are often assumed to be homologs. The basic local alignment search tool (BLAST: Altschul et al., 1990) is a commonly used tool that provides statistical properties on the similarities of two sequences. The expectation (E) value within BLAST is a statistical figure that shows how often the observed similarity between a pair of sequences is expected to occur by chance. A low E value is thus desired since the chance of observing the sequence similarity is small. Another useful value provided in a BLAST search is the bit score (S’), which is a derived from the raw alignment score (S) in which the statistical properties of the scoring system have been taken into account (i.e. a normalised score). The raw alignment score (S) is a calculated with the summation of scores for each aligned position and the scores for gaps in the alignment.

The best reciprocal BLAST hit criterion is often used to identify orthologous sequences. A best reciprocal match is one where two proteins are the best BLAST hit of each other. However detecting orthologs is not as trivial as taking the best reciprocal BLAST hit. If gene duplications occurred in each of the given two genomes subsequent to their divergence, only a many-to-many relationship will adequately describe orthologs (Tatusov et al., 1997). A many-to-many relationship is needed to get an overall picture of best BLAST hits, so that redundancies can be removed (i.e. those that are paralogs).

Multiple sequence alignment

ClustalW (Thompson et al., 1994) and T-Coffee (Notredame et al., 2000) are two commonly used programs, out of the many programs that are available, for making multiple global sequence alignments. These two programs use heuristics to calculate alignments since calculating all possibilities of a multiple sequence alignment will be computationally intractable. A benchmark of multiple sequence alignment programs

(Raghava et al., 2003) showed that T-Coffee and ClustalW both align reference sequences correctly with around 90% accuracy based on how well the programs would identify the conserved motifs in the benchmark data.

Multiple sequence alignments are starting points for studying molecular evolution and for inferring phylogenetic trees. They provide an overview of how sequences have evolved through time by providing an alignment of the sequences showing conserved residues, point mutations and insertions and deletions, which are represented with gaps in the sequence. The conservation of residues is due to structural or functional importance. Non-functional regions can freely evolve and hence correspond to different residues among sequences. Different programs use different algorithms to calculate the best sequence alignment. However when a sequence is highly conserved many programs will often provide good alignments (Raghava et al., 2003). Thus it is useful to compare alignments produced from different programs. The default settings of alignment programs are often calibrated to work for most cases of DNA and protein alignments: since sequence divergence is not expected to be high within genus *Staphylococcus*, inferred alignments should be less dependent on the choice of program and parameters than in 'challenging' benchmark databases such as BAliBASE (Thompson et al., 2005).

Inferring phylogenetic trees

Phylogenetic trees are either constructed or a tree search is performed looking for tree that optimises an optimality criterion based on observations of the data such as sequence alignments. The outcome is a tree that represents a hypothesis of the evolutionary relatedness of characters, such as sequences. Four commonly used methods are mainly employed in reconstructing phylogenies; neighbour joining (NJ), parsimony, maximum likelihood (ML) and Bayesian methods (Holder and Lewis, 2003). Procedures used to derive phylogenetic trees implicitly rely on a number of assumptions and approximations.

NJ is a relatively fast algorithm that constructs most trees in a matter of seconds. The algorithm converts the DNA and protein sequences into a distance matrix, which represents an estimate of the evolutionary distance between sequences. However because information is compressed into distances, a lot of information is not considered, thus NJ only works relatively well with much more similar alignments (Holder and Lewis, 2003). NJ differs from the other methods in that it constructs trees whereas the parsimony, ML and Bayesian methods searches for an optimal tree based on a criterion. In parsimony the optimality criterion is the tree with the fewest number of mutations that could possibly produce the data is considered the correct tree. Long-branch attraction, a phylogenetic artefact, is a major problem for parsimony as it does not take into account convergent evolution for similarities between long-branch lengths, hence grouping them as the most closely related (Holder and Lewis, 2003). Unequal substitution rates are a major problem with parsimony as no model of substitution is used (Gribaldo and Philippe, 2002).

In ML, a hypothesis (*e.g.*, a phylogenetic tree with branch lengths) is judged by how well it predicts the observed data; the tree that has the highest probability of producing the observed sequences is preferred (Holder and Lewis, 2003). This is the likelihood of the tree which is calculated by the probability of the data given the model and tree hypothesis. Despite being regarded as the most appealing way to estimate phylogenies (Holder and Lewis, 2003), ML is extremely slow as all possible trees are exhaustively searched for.

An evolutionary model is used to describe the different probabilities of change from one character to another. The simplest such models are the Jukes-Cantor model (Jukes and Cantor, 1969) for nucleotide sequence and the Poisson model (Bishop and Friday 1987) for amino acid sequences. Both models assume that substitution rates and character frequencies are equal. Amino acids model are much more complex than nucleotide models due to a much more characters, 20 for amino acids and 4 for nucleotides. As such amino acid models are built with a large training set of proteins to determine the frequencies of substitution (Lio and Goldman, 1998). Models are used in some phylogenetic analyses such as ML and Bayesian methods and they are important for calculation of likelihood functions.

Bayesian inference of phylogeny uses an optimality criterion, which maximizes the posterior probability (Huelsenbeck et al., 2001) and is used as the confidence value in Bayesian phylogeny estimations. The posterior probability (PP) for a hypothesis is proportional to the likelihood multiplied by the prior probability of that hypothesis. As such ML and Bayesian methods are related in that they both use likelihood functions. Bayesian inference of phylogeny usually uses an algorithm for estimating the PP, which is commonly the Monte Carlo Markov Chain (MCMC) (Huelsenback et al., 2001). The MCMC algorithm samples trees and the posterior probability of a tree is the amount of times that tree has been sampled. Thus if a tree or node in a tree is sampled more often than others (i.e. a higher PP value) then more confidence is associated with that particular tree. The PP is generally regarded as an estimate of the likelihood function and by trying to maximise the PP one is indirectly trying to maximise the likelihood, which is similar to the ML analyses (Holder and Lewis, 2003), though Bayesian analyses are much faster than ML methods. The basic idea is to construct a Markov chain that has as its state space the parameters of the statistical model and a stationary distribution that is the posterior probability distribution of the parameters. The MCMC involves two steps: 1. a new tree is proposed by stochastically perturbing the current tree; 2. this tree is then either accepted or rejected based on the posterior probability then step 1 is repeated (Huelsenbeck et al., 2001). New trees with lower likelihoods that are proposed than the current tree are accepted only a proportion (p) of the time, where p is the ratio of the posterior of the proposed tree compared to the posterior of the current tree (Holder and Lewis, 2003), hence proposed trees with a huge drop in posterior probability are less likely to be accepted. Convergence is reached when proposed trees have similar posteriors as the current tree. Trees used in the start of the MCMC run usually have low posterior probabilities and are discarded as burn-in.

A Phylogenetic Search for LGT in *Staphylococcus*

The sequencing of *Staphylococcus* species has revealed that LGT has played an important role with the evolution of virulence and antibiotic resistance in *Staphylococcus*. Virulence and antibiotic resistance genes offer selective advantage to the organism and as such have been fixated into the genome. Closely related organisms are thought to transfer genes more often than transfers between distantly related organisms due to compatibility issues and hence it is interesting to examine the amount of gene transfer has occurred among the *Staphylococcus*.

This project aims to identify genes exchanged within the Staphylococci and the patterns of gene transfer based on incongruent phylogenetic trees. The use of phylogenetics offers a systematic basis for detecting LGT and allows the use of rigorous statistics for support. A phylogenetic approach to studying LGT in *Staphylococcus* has not been previously and this project aims to use this approach to estimating LGT.

Methods

The pipeline applied in this project is similar to one developed for an analysis of 144 prokaryotic genomes (Beiko et al., 2005a) where 22,432 putatively orthologous sets of proteins were used to construct a supertree. A similar pipeline was used in this project to investigate 8 genomes belonging to the genus *Staphylococcus* (including four not examined in the previous analysis).

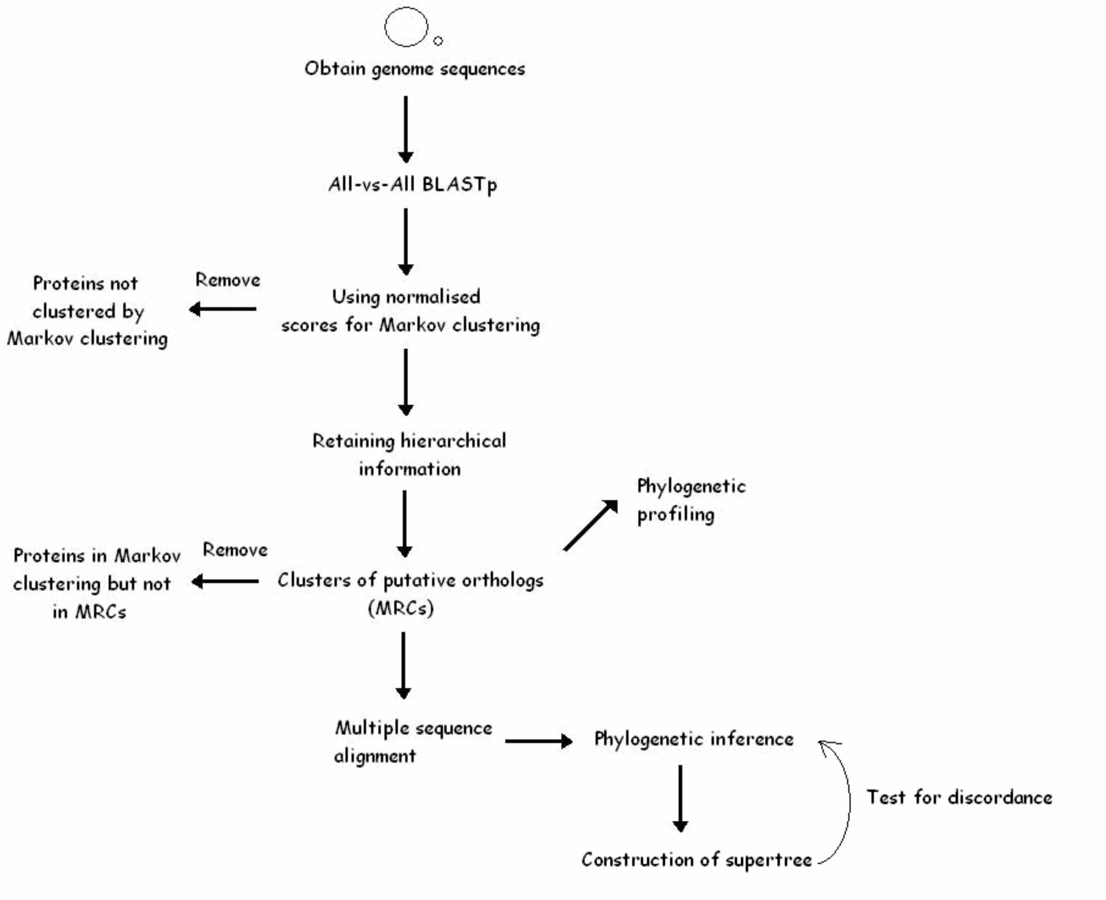
Obtain genome sequences
All-vs-All BLASTp
Proteins not clustered by Markov clustering
Remove
Using normalised scores for Markov clustering
Retaining hierarchical information
Phylogenetic profiling
Proteins in Markov clustering but not in MRCs
Remove
Clusters of putative orthologs (MRCs)
Multiple sequence alignment
Phylogenetic inference
Test for discordance
Construction of supertree

Figure 2. Pipeline of the project

The nucleotide and amino acid genome sequences of all 8 *Staphylococcus* isolates publicly available as of 25th February 2005 were downloaded from the National Center for Biotechnology Information (NCBI). The set of 8 included 6 *Staphylococcus aureus* species; S.aureus_aureus_MRSA252, S.aureus_aureus_MSSA476, S.aureus_COL, S.aureus_Mu50, S.aureus_MW2, S.aureus_N315 and 2 *Staphylococcus epidermis* species; S.epidermis_ATCC12228, and S.epidermis_RP62A. 2 more *Staphylococcus* genomes are publicity available from NCBI as of 19th October 2005;*Staphylococcus saprophyticus subsp. saprophyticus* and *Staphylococcus haemolyticus JCSC1435.* These genomes are not represented in the current analyses, but can provide further context for *S. aureus* in future work.

Determination of homologous and orthologous sets of proteins

The 8 genome sequences (i.e. main chromosome and plasmid sequences if available) of *Staphylococcus* encoded a predicted total of 20,865 proteins and an all-*vs*-all BLASTp (Altschul et al., 1997) was carried out on these proteins. An expectation score (e) of ≤ 1.0 x 10^{-3} was used as the cut-off for all pairwise comparisons of the proteins: proteins with statistical similarity greater than this threshold were treated as unrelated.

In an all-*vs*-all BLASTp, all proteins from each genome are BLASTed against every other protein from each genome including itself (i.e. 64 comparisons for 8 genomes). A normalised BLASTp score was then derived and used for the Markov clustering (MCL) algorithm with an inflation parameter (I) of 1.10 (See Harlow et al., 2004 for calculation of normalised BLASTp score and for parameter selection). The clusters of proteins determined under the MCL algorithm are putative homologs. In a Markov cluster every protein in space is connected by normalised BLASTp scores. The connectivity is determined at what limit the threshold is set and a Markov cluster does not need to be completely connected to exist. At different thresholds certain cluster sizes are possible. In figure 3, a threshold of 0.99 would connect 2 proteins by a single connection; at a threshold of 0.85, 3 proteins are connected by 2 connections and at threshold 0.66 a cluster of 4 proteins is possible.

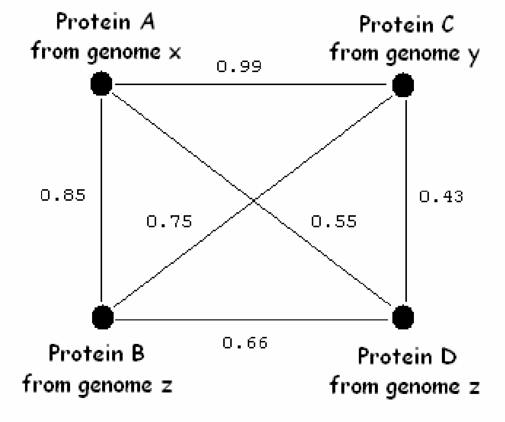

Figure 3. Representation of a Markov cluster with 4 proteins each connected by their normalised BLASTp scores

Single linkage clustering was then applied to further separate the homologous groups of proteins into putative orthologs. Single linkage clustering is based on representative clusters (RCs), which are clusters where each protein in the cluster must be from a different genome. In Figure 4 the letters represent different genomes from which homologous sequences in a Markov cluster have been obtained. Hence RCs would be AB, ABC, and ABCDE, since every sequence in these clusters comes from a different genome. A Maximally representative cluster (MRC) is a RC that has the maximum amount of proteins in the cluster while still satisfying the criterion of one protein from one genome, thus the MRC in figure 4 is ABCDEF. The highest threshold at which this MRC can form is slightly above 0.60. The lowest threshold at which this MRC can exist is just when gene 2 from genome D branches off. The cluster support is determined by the maximum – minimum threshold. These MRCs are groups of putative orthologs, and the sets that were separately subjected to sequence alignment and phylogenetic analysis.

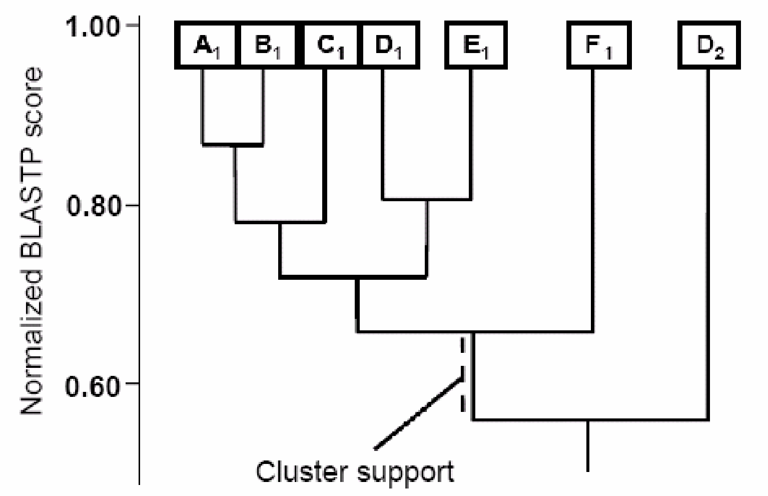

Figure 4. Single linkage clustering showing RCs and MRCs

Sequence alignment

T-Coffee (Notredame et al., 2000) version 1.37 and ClustalW (Thompson et al., 1994) version 1.83 were used to perform the multiple sequence alignments on the protein sequences in each MRC. Both programs were run under their respective default settings and their individual alignments were compared. A comparison of the alignments produced by T-Coffee and ClustalW was done by comparing the relative positions of individual residues of each alignment.

Calculation of alignment similarity between T-Coffee and ClustalW

A comparison of the alignments produced by the different programs is based on comparing aligned residues in each column. The column residues in T-Coffee alignments were compared to the column residues in ClustalW alignments. If the residue aligned by T-Coffee is the same residue aligned by ClustalW, a score of 1 is added to a cumulative score. The sum of column similarities is then divided by the mean of the self score to produce a normalised score that can be used to compare scores for different alignments (Beiko et al., 2005b).

Alignment percent identity based on overall pairwise score

The overall pairwise percent identity of an alignment is calculated by comparing each sequence in a given MRC to every other sequence. All positions in the first sequence are compared to all positions in the second sequence. The identity score between a pair of sequences is incremented by 1 for each alignment column in which the residues from the two sequences are identical. With 4 sequences there would be 6 pairwise comparisons; each pairwise comparison would give a percentage identity between the two sequences (identity score / length of sequence * 100); all percentage identities are added up and divided by the possible amount of pairwise comparisons, which in the case of 4 sequences is 6, yielding the overall pairwise score.

Example:

Sequences:

A	AGTC
B	AGTC
C	ATTC
D	AATC

	B	C	D
A	100	75	75
B		75	75
C			75

The overall score = 100 + 75 + 75 + 75 + 75 + 75 / 6

= 79.2%

These overall pairwise scores provide a simple measure of the amount of sequence divergence within an alignment.

The protein alignments done with T-COFFEE were used to guide the alignment of nucleotide sequences. Once a protein alignment was done, the respective codon coding for the amino acid in the protein alignment was reverse transcribed back to the triplet code. An actual alignment of nucleotide sequences was not done, since direct nucleotide alignment does not respect the codon assignments of nucleotides and can lead to many frameshifts induced by gaps of length 1 or 2. The overall pairwise score of nucleotide alignments was done using the same scheme above.

Phylogenetic inference

Protein sequences in each MRC aligned using T-Coffee and the aligned nucleotide sequence in each MRC were used for phylogenetic inference. MrBayes (Huelsenbeck and Ronquist 2001) was used for Bayesian inference of phylogeny and amino acid parameter settings used were determined through extensive calibration experiments done on prokaryotic protein data sets (Beiko et al., submitted). Main settings include the choice of phylogenetic model and the number of generations to run the analysis.

The tree-like relationships of sequences proposed by MrBayes can be analysed in the form of bipartitions. A bipartition forms two relationships, which can be viewed as everything to the "left" and everything to the "right" of a given branch in the tree. For example with four sequences A, B, C and D, there are 7 possible bipartitions as seen in figure 5.

1. A | BCD 2. B | ACD 3. C | ABD 4. D | ABC

5. AB | CD 6. AC | BD 7. BC | AD

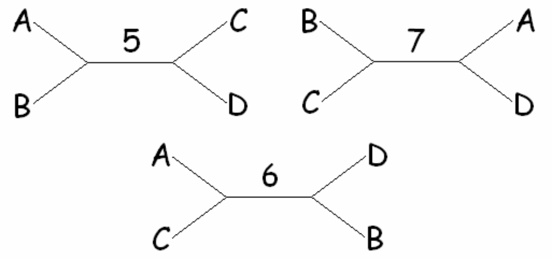

Figure 5. Representation of trees by bipartitions; bipartitions 1 – 4 are just the single branch versus every branch

Bipartitions 1 to 4 are trivial bipartitions because a single character on its own is not informative and will appear in every tree. There are 3 non-trivial bipartitions, which offer possible information to the relationships of the sequences.

A four-category discrete gamma approximation (Yang, 1994) was used to model rate variation across sites, as the tendency for some sites to undergo change is greater than other sites. This rate of changing from one character to another is modelled with the gamma distribution, sampling four points of the gamma distribution, which is approximated by four discrete categories of slowly and rapidly evolving sites. The alpha parameter of the gamma distribution defines the rates for within each category and the relative possibility of sites occurring in each category and was allowed to vary across a wide range of alpha values (0.1 to 50). A four-category invariant discrete gamma distribution was used for nucleotide sequences.

For amino acid models, MrBayes supports an amino acid model-switching method. A prior probability of 0.2 was set for the Jones (Jones et al., 1992), Dayhoff (Dayhoff, Schwartz and Orcutt, 1978), WAG (Whelan & Goldman 2001), VT (Muller and Vingron, 2000) and Blosum (Henikoff and Henikoff, 1992) amino acid models. This allows different empirical models to be sampled by MrBayes, instead of having every alignment conform to the same model. For the MCMCMC runs, 4 chains 3 heated and 1 cold were used with a 0.5 temperature for the heated chains. Branch length information was saved. And finally the MCMCMC runs were ran for 1,000,000 generations.

For nucleotide alignments, the General Time Reversible (GTR) (Tavaré, 1986) model was used and MCMCMC with the same parameter settings as amino acid runs was done for 2,000,000 generations. Since five parameters of substitution probabilities are needed as input for the GTR model compared to only one parameter choice necessary for amino acid runs, more generations were done for DNA.

Posterior probabilities are assigned to every possible bipartition in proportion to the number of times it is sampled. The number of times a tree is visited is dependent on the likelihood of the tree, thus the PP of a partition can be an estimate of the correct tree. Relating back to Figure 5, each of the 3 non-trivial bipartitions will be sampled at a frequency between 0.0 and 1.0 and this frequency is the PP support.

A burnin of 1000 was set for amino acid and nucleotide sequences, which discards 10% of the sampled trees. This is required since MrBayes randomly looks for trees in tree space so trees sampled in this first section have poor PP associated to them. The first generation is always sampled hence for amino acid runs, 10,001 trees will be used for summarizing and 20,001 for nucleotides.

A consensus tree for each MRC was produced using the 50% majority rule consensus approach (http://evolution.genetics.washington.edu/phylip/doc/consense.html). The consensus tree is constructed based on the PP support of every bipartition.

Reference tree construction

Clann version 2.0.1 (Creevey and McInerney, 2005) was used for supertree construction, which was considered the reference tree. There are four methods of supertree construction implemented into Clann; Matrix Representation with parsimony (MRP), Most similar supertree method (dfit), Maximum Quartet fit (qfit) and Maximum Split fit (sfit). All four methods were used to build supertrees from the protein and nucleotide data.

The consensus trees produced by MrBayes for the amino acid and nucleotide analyses were used as source trees for supertree construction. A 0.95 threshold of posterior probability (PP) support of bipartitions in the consensus tree was applied. Bayesian posterior probabilities have been shown to be extremely high when studying concatenated sequences (Suzuki et al., 2002), although the authors also postulated that this could occur with unconcatenated sequences. Thus a stricter (common practice uses thresholds ≥ 0.90 thresholds) PP threshold was applied in light of this evidence. Bipartitions with PP support less than 0.95 were collapsed, and fully unresolved consensus trees (i.e. a tree resembling a star topology) were removed prior to supertree construction. The branch lengths of each MRC tree were also removed. 2180 protein and 2061 nucleotide trees contained at least one bipartition with PP ≥ 0.95, and were used to build separate supertrees.

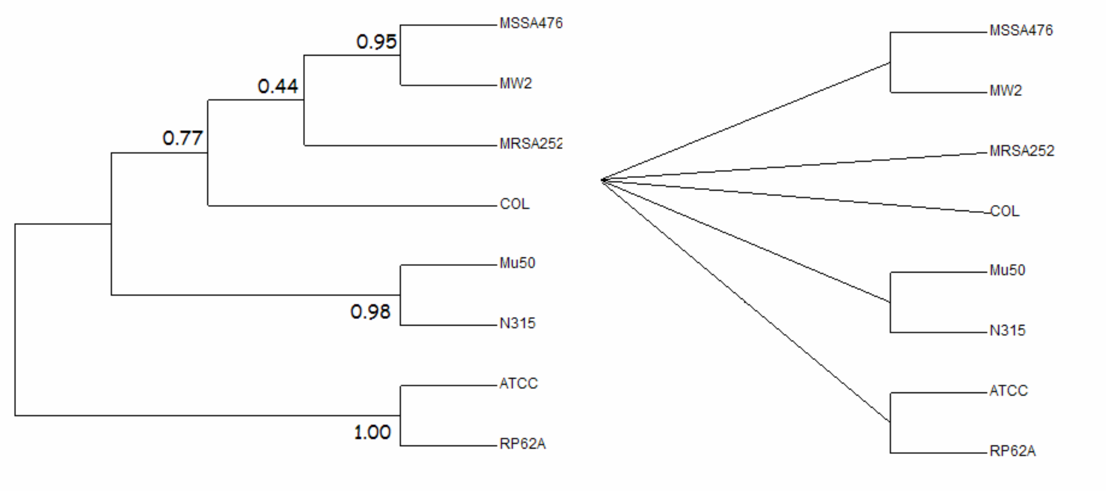

Figure 6. Collapsing of nodes less than a certain threshold. The tree on the left shows PP support for all nodes. The tree on the right collapses nodes that have a PP of ≤ 0.95, showing a partially resolved tree.

Additional analysis was done to determine the sensitivity of PP thresholds on the source trees. PP thresholds of 0.50, 0.90 and 1.00 were applied to source trees and trees with no resolution were removed from the source input into Clann. The supertree construction method with these alternative PPs was MRP.

Topological comparison of protein/gene trees with reference tree

Non-trivial bipartitions with PP ≥ 0.95 for each MRC were compared with the partitions in the MRP supertree. A partition with PP ≥ 0.95 that agrees with the supertree is a concordant MRC set. Conversely, a partition with PP ≥ 0.95 that disagrees with the supertree is a discordant MRC set, and is treated as a potential lateral genetic transfer case.

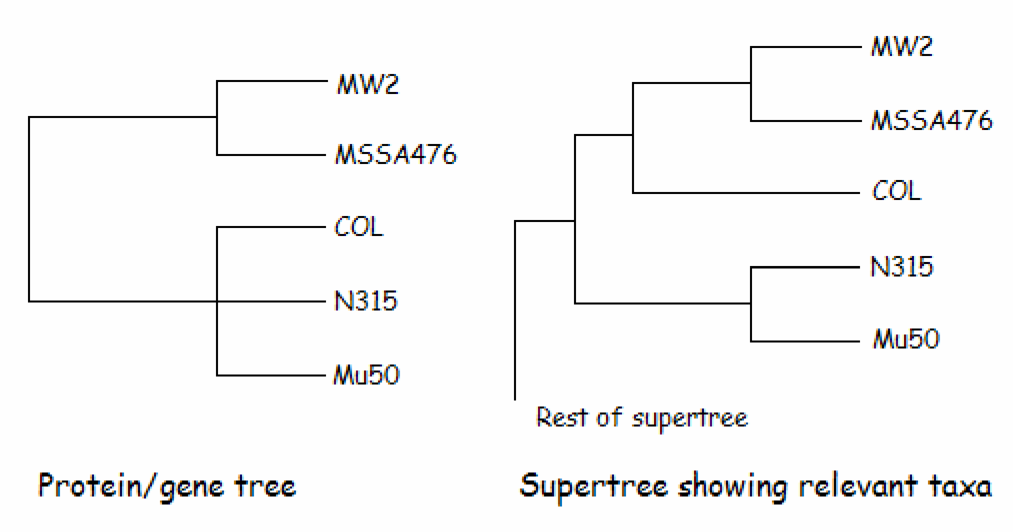

Figure 7. Protein/gene tree showing concordance and discordance nodes to that of the reference tree

A bipartition in a MRC violates the supertree when the relationships in the MRC bipartition don't agree with bipartition in the supertree. Figure 7 shows a protein gene/tree with a concordant bipartition of MW2, MSSA476 | COL, N315, Mu50 since this agrees with the supertree partition of MW2, MSSA476 vs. every other taxon. However the bipartition of COL, N315, Mu50 | MW2, MSSA476 violates the bipartition of MW2, MSSA476, COL vs. every other taxon and is considered a discordant bipartition. Note that a single MRC can contain both concordant and discordant bipartitions such as the example shown in figure 7.

Discordant cases

Protein/nucleotide MRC trees that have bipartitions that are different from the bipartitions in the reference tree are discordant cases. These cases are putative laterally transferred characters from within the Staphylococci. Since a discordant bipartition can be formed artefactually by aspects of this project, all possible artefacts will be considered and assessed; discordances a result of the alignment problems; discordances due to low cluster support and one to consider in the future, discordances as a result of the phylogenetic inference method.

Functional characterisation

Assigning functional roles by COG annotations

The annotations done for each protein in the genomes do not provide much of a consensus. This is mainly due to different methods used for the sequencing and annotation of genomes applied by different sequencing centres. Thus functional characteristics used by Cluster of Orthologous Groups (COGs) (Tatusov et al., 1997) were applied to each MRC for functional characterisation. Note that orthologous information was not used from the COG database but just the annotation of proteins in genomes. There are 4 main functional categories broken down into 25 functions roles each assigned with a letter of the alphabet, called the COG letter. Some of these letters are not relevant to microbes, such as chromatin structure and dynamics since bacterial chromosomes lie freely in the cytoplasm.

Information storage and processing

(J) Translation, (A) RNA processing and modification, (K) Transcription,

(L) Replication, recombination and repair and (B) Chromatin structure and dynamics

Cellular processes and signalling

(D) Cell cycle control, mitosis and meiosis, (Y) Nuclear structure, (V) Defence mechanisms, (T) Signal transduction mechanisms, (M) Cell wall/membrane biogenesis, (N) Cell motility, (Z) Cytoskeleton, (W)

Extracellular structures, (U) Intracellular trafficking and secretion and (O) Posttranslational modification, protein turnover, chaperones

Metabolism

(C) Energy production and conversion, (G) Carbohydrate transport and metabolism, (E) Amino acid transport and metabolism, (F) Nucleotide transport and metabolism, (H) Coenzyme transport and metabolism, (I) Lipid transport and metabolism, (P) Inorganic ion transport and metabolism and (Q) Secondary metabolites biosynthesis, transport and catabolism

Poorly characterised

(R) General function prediction only and (S) Function unknown

The updated COG database (http://www.ncbi.nlm.nih.gov/COG/new/) contains clusters of orthologous genes from 66 genomes each given a functional role (Tatusov et al., 2003). Only the S. aureus N315 genome is within this restricted dataset. Thus only proteins from the N315 strain would have COG letters associated with it.

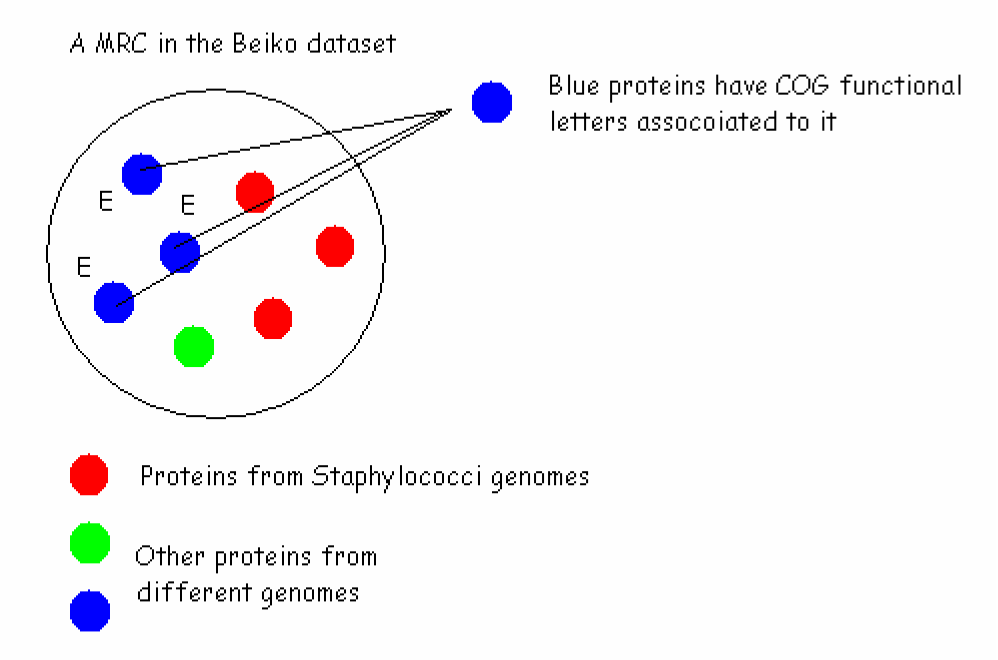
A MRC in the Beiko dataset
Blue proteins have COG functional letters assocoiated to it
E
E
E
Proteins from Staphylococci genomes
Other proteins from different genomes

Figure 8. Assigning COG letters from the Beiko et al. 2005a dataset to the *Staphylococcus* MRCs

The large dataset of 144 genomes analysed by Beiko et al. 2005a, contained three strains of *S. aureus* (N315, MW2 and Mu50) and one of *S. epidermis* (ATCC). In some cases, these genomes were represented in MRCs that also contained proteins from non-*Staphylococcus* genomes, which had COG annotations. Figure 8 shows a MRC that contains 3 proteins from the *Staphylococcus* genomes (shown in red) and other various proteins from other genomes. In this MRC the blue proteins have a COG letter E assigned to them. Hence the 3 proteins from the Staphylococci, are also associated the COG letter E. Any MRC with 2 functional roles assigned were discarded from the functional analysis. MRCs containing the Staphylococci may have no COG letter assigned to it were assigned as NULL cases.

The COG letters derived from the Beiko et al. (2005) dataset were then applied to the MRCs of 8 *Staphylococcus*. Not every protein of N315 formed MRCs in the Beiko et al. (2005) dataset, and hence both the derived COG letters from the Beiko et al. (2005) dataset and the COG letters assigned to every N315 protein were applied to the *Staphylococcus* MRCs.

Assigning functional roles by TIGR annotations

TIGR or The Institute for Genomic Research functional roles (Peterson et al., 2001) were also used to characterise the functions of the MRCs. There are 17 main categories, and sub categories within the 17 but they are not shown because only the main categories were used to characterise the Staphylococci MRCs. All 8 *Staphylococcus* genomes were re-annotated by TIGR and a functional category was assigned to proteins in the Staphylococci genomes. TIGR uses an annotation engine, which predicts ORFs and assigns functional assignments by performing a BLAST search against a wide range of databases (Peterson et al., 2001). Scores are analysed and if this score is above a cut-off then a TIGR Role Category is assigned to the protein (Peterson et al., 2001). Functional roles of each *Staphylococcus* protein re-annotated by TIGR were assigned to the MRCs.

TIGR functional categories

(1)Amino Acid biosynthesis, (2)Biosynthesis of co-factors prosthetic groups and carriers, (3)Cell envelope, (4)Cellular processes, (5)Central intermediary metabolism, (6)DNA metabolism, (7)Energy metabolism, (8)Fatty acid and phospholipid metabolism, (9)Mobile and extra chromosomal element functions, (10)Protein fate, (11)Protein synthesis, (12)Purines, pyrimidines, nucleosides, and nucleotides, (13)Regulatory functions, (14)Signal transduction, (15)Transcription, (16)Transport and binding proteins and (17)Unknown function

A Chi Squared test was done to see if there are functional roles significantly represented by discordant MRCs. Expected counts are those that are represented in all MRCs and was calculated by how often a functional role was present in concordant cases multiplied by the total count of discordant cases. Observed counts are those that are observed in the discordant MRCs. Chi squared tests were done for COG roles, COG categories and TIGR roles.

Phylogenetic profiling

Phylogenetic profiling takes the assumption that functionally related genes are distributed in an equal fashion (Pellegrini et al., 1999). This can be used as a quick estimate to seeing how related organisms are by the presence and absence of certain proteins. This approach offers an alternative to the phylogenetic estimation using Bayesian framework.

For all MRCs a binary code (i.e. 1 and 0) of 8 digits was used to indicate the presence or absence of a genome. The 1st position in the string of digits belongs to the ATCC strain, 2nd position = COL, 3rd position = MRSA252, 4th = MSSA476, 5th Mu50, 6th = MW2, 7th = N315 and 8th = RP62A (the order was chosen alphabetically). A string “11000101” indicates a MRC of size 4 with genomes from ATCC, COL, MW2 and RP62A.

Strings were produced for all MRCs of size 4 to 7, since MRCs of size 8 only have one possible combination. If a string occurs more than once, i.e. 01111110 a MRC with all 6 *S. aureus*, a tally of the occurrence was recorded. A phylogenetic profile can also be represented by gene loss events, e.g. a profile of 01111111 can be explained by a gene loss in ATCC.

Results

Clustering

The MCL algorithm clustered 19683 of the total predicted 20865 (94.3% coverage) protein coding genes in all 8 *Staphylococcus* genomes. The single linkage clustering criteria of Maximally Representative Clusters (MRCs) yielded 17923 proteins in 2470 clusters of size 4 to 8 (85.9% coverage).

There were a total of 1344 Markov clusters of size 4 (phylogenetically informative) to the largest Markov cluster of size 670. There were 377 clusters of size 9 and larger, representing 12633 proteins that have at least some paralogous relationships. Hence 1182/20865 (5.66%) proteins were not clustered by the Markov clustering (i.e. representing proteins with no apparent homologs). The 2942 proteins not included in a MRC (2942/20865 = 14.1%) could belong to MRCs of size 4 or smaller and thus offer no phylogenetic information since a group of 3 species only offers one possible relationship.

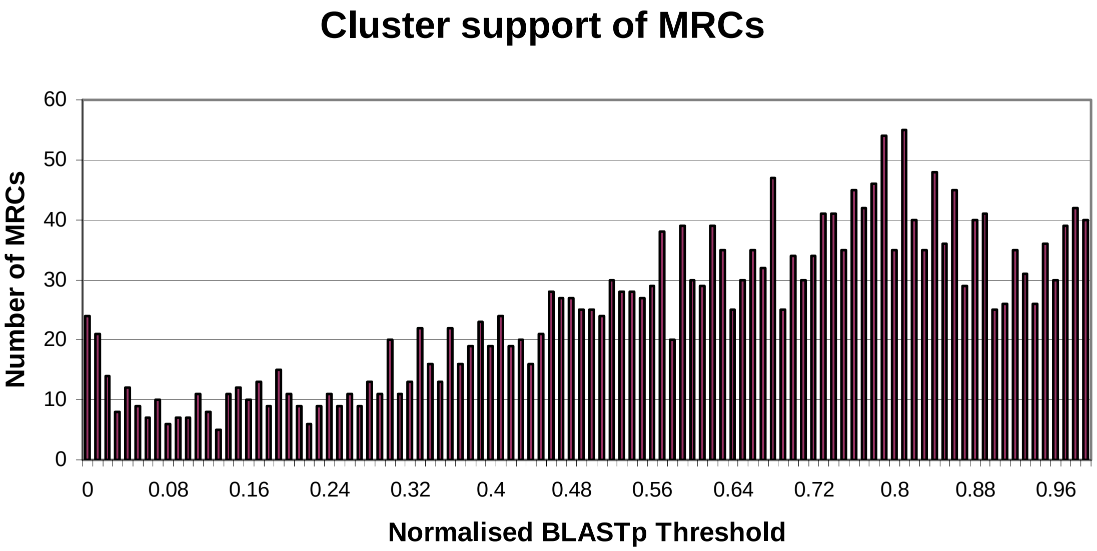

Figure 9. Distribution of the cluster support (maximum – minimum threshold) amongst all MRCs

Cluster support of all MRCs were generated and graphed, shown in figure 9. Results are skewed towards higher support values for clusters and reflected in the median distribution of cluster support of 0.67. A mean average assumes a normal distribution however the distribution of cluster support of MRCs is skewed towards the right and hence a median was used.

Multiple sequence alignment

Sequences contained within each MRC were aligned using ClustalW version 1.83 and T-coffee version 1.37 both running under default parameters. 2470 sequence alignments were produced by the respective programs and these were assigned a score based on the mean pairwise percent identity. All mean pairwise percent identity scores were collected and a mean was calculated. The average mean pairwise percent identity score was

89.22% for the 2470 ClustalW protein alignments and 89.14% for T-coffee protein alignments. The average pairwise percent identity score of the 2740 nucleotide alignments derived from the T-coffee protein alignments was 89.09% These percent identity scores reflect the similarity of all the sequences in MRCs.

Comparing ClustalW and T-coffee alignments based on the aligned residues in each column

All MRCs were aligned by ClustalW and T-Coffee and compared. A normalised score was produced with the comparison of a ClustalW alignment and a T-Coffee alignment. The normalised scores ranged from 0.61 to a maximum of 1, which corresponds to an identical alignment produced by both programs. The mean of all comparison scores was 0.995, which means that most MRCs running under the different alignment programs yielded extremely similar and in many cases identical alignments. This is due to the high similarity of the sequences, as seen in the mean average of all pairwise percent identity scores of 89% identity, and highlights the relative simplicity of aligning these sequences.

Inference of phylogenetic trees

For any given taxon tree of size n there are n – 3 possible resolved bipartitions, hence a 3 taxon tree offers no resolution. Given the number of MRCs corresponding to different taxon trees we can calculate the total possible resolved bipartitions. There were 127 size 4 MRCs; 132 size 5 MRCs; 423 size 6 MRCs; 87 size 7 MRCs; and 1701 size 8 MRCs hence a total of 10513 possible resolved bipartitions for all MRCs. For proteins a total of

3247 bipartitions had a posterior support of ≥ 0.95. Relaxing the thresholds to 0.90, 0.75 and 0.50 produced 3543, 3982 and 4595 resolved bipartitions respectively.

For nucleotides the total number of bipartitions resolved with a posterior support of ≥ 0.95 was 3065. Relaxing the posterior thresholds from 0.95 to 0.90 included 727 more bipartitions, yielding 3792 resolved partitions. At 0.75 and 0.50 thresholds, 5056 and 6518 partitions were resolved at the respective threshold. Bipartitions from nucleotide sequences are more affected by posterior probability support than are amino acid sequences.

PP threshold	Resolved protein bipartitions	Resolved nucleotide bipartitions
1.00	2162	2007
0.95	3247	3065
0.9	3543	3792
0.75	3982	5056
0.5	4595	6518

Table 1. Relationships of PP thresholds to the resolution of bipartitions

Reference topology

MrBayes generated 2470 individual nucleotide and protein consensus trees. Branches with PP support < 0.95% in the consensus trees were collapsed (see figure 6) into an unresolved branch. Trees with no resolution (i.e. do not define any relationships) were removed. This yielded 2061 nucleotide consensus trees and 2180 protein consensus trees with at least one strongly supported bipartition. 2019 trees used for the supertree construction were common to both sets. These were used as source trees by Clann for constructing a supertree.

The sensitivity of the PP threshold with consensus trees was also examined. Again nucleotide data are more sensitive to PP thresholds.

Protein			
			# MRC trees fully or
PP Threshold	# MRC trees	# Star topologies	partially resolved
1	2470	406	2064
≥0.95	2470	290	2180
≥0.90	2470	269	2201
≥0.50	2470	227	2243
Nucleotide			
			# MRC trees fully or
PP Threshold	# MRC trees	# Star topologies	partially resolved
1	2470	529	1941
≥0.95	2470	409	2061
≥0.90	2470	368	2102
≥0.50	2470	214	2256

Table 2. Resolution of source trees with at least one resolved bipartition at respective PP thresholds

Clann implements 4 supertree algorithms (Matrix representation with Parsimony (MRP), most-similar supertree (dfit), maximum quartet fit (qfit), and maximum split fit (sfit)), and they were used to construct supertrees of amino acid and nucleotide source trees.

The three algorithms of MRP, qfit, sfit produced the same unrooted supertree with amino acid and nucleotide data. The dfit algorithm produced a different supertree from the other 3 algorithms with both amino acid and nucleotide data. The dfit supertree grouped COL and MRSA252 together.

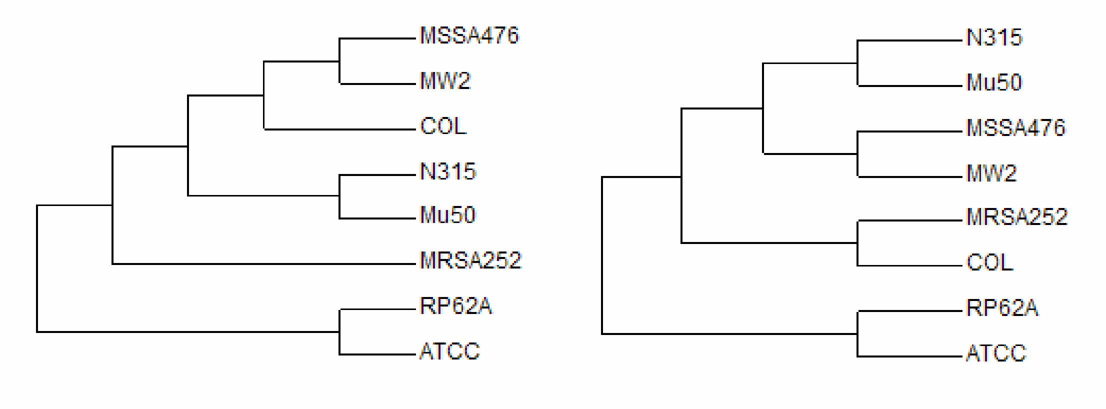

Figure 10. Supertrees constructed with different algorithms. The supertree on the left was produced with the MRP, qfit and sfit algorithms whereas the supertree on the right was produced with the dfit algorithm. Use of nucleotide and amino acid data yielded the same supertrees.

Topological comparisons of trees

Every bipartition with a posterior probability support of ≥ 0.95 from every amino acid and nucleotide tree was used to test for concordance/discordance in the 5 non-trivial bipartitions of the supertree. To analyse whether the bipartition sensitivity to PP threshold, shown in table 1, affected support of the supertree, bipartitions ≥ 0.90 were also tested against the supertree.

Nucleotide Bipartition number	Bipartitions tested	PP ≥ 0.90 Concordance	PP ≥ 0.90 Discordance
9	2386	745	9
10	2347	172	155
11	2388	873	9
12	1843	133	32
13	1773	1763	5

Bipartition number	Bipartitions tested	PP ≥ Concordance 0.95	PP ≥ 0.95 Discordance
9	2386	471	9
10	2347	113	108
11	2388	567	7
12	1843	95	20
13	1773	1762	5

Amino acid Bipartition number	Bipartitions tested	PP ≥ 0.90 Concordance	PP ≥ 0.90 Discordance
9	2386	530	20
10	2346	193	241
11	2388	694	11
12	1843	150	109
13	1773	1749	8

Bipartition number	Bipartitions tested	PP ≥ 0.95 Concordance	PP ≥ 0.95 Discordance
9	2386	469	15
10	2346	158	181
11	2388	612	11
12	1843	118	71
13	1773	1748	8

Table 3. Effect of varying threshold support of bipartitions to discordance and concordance

The overall discordance was calculated by the total discordance / (total discordance + total concordance). For PP ≥ 0.95 the overall nucleotides discordance was 149 / 3157 x 100 = 5% and for PP ≥ 0.95 the overall amino acids discordance was 286/3105 x 100 = 8%

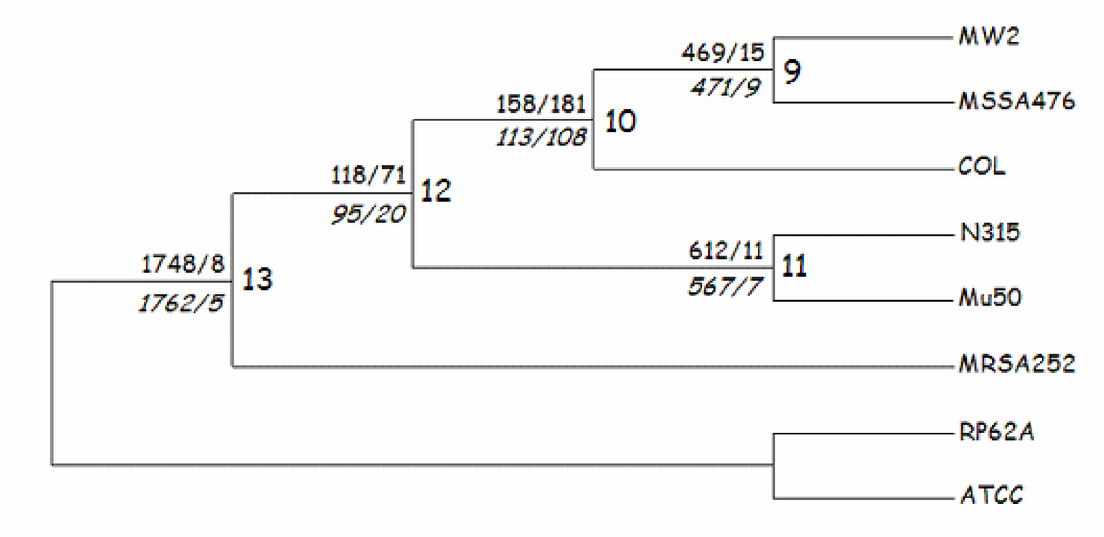

Figure 11. Supertree with ≥0.95 PP concordance and discordance bipartitions at each meaningful node. Values on the top of a branch are protein concordance/discordance bipartitions and those that are on the bottom are nucleotide concordance/discordance values. The supertree is arbitrarily rooted between *S. epidermidis* and *S. aureus* and the numbers of the right of the concordant/discordant are bipartitions number of the supertree used in table 3.

			Nucleotide	
		Concordant	Discordant	Unresolved
	Concordant	2491	5	512
Protein	Discordant	4	91	54
	Unresolved	610	190	6779

Table 4. Two-way table of consistent and conflicting nucleotide and amino acid bipartitions at ≥0.95 PP. Unresolved cases represented cases where PP < 0.95

Discordant protein bipartitions were analysed as to whether they were concordant nucleotide bipartitions and vice versa. This provides a view of how conflicting the signals are in nucleotide and protein sequences and results were summarized in Table 4.

Bootstrap analysis

Bootstrap analysis was done with the protein and nucleotide source trees with 0.95 PP branching support.

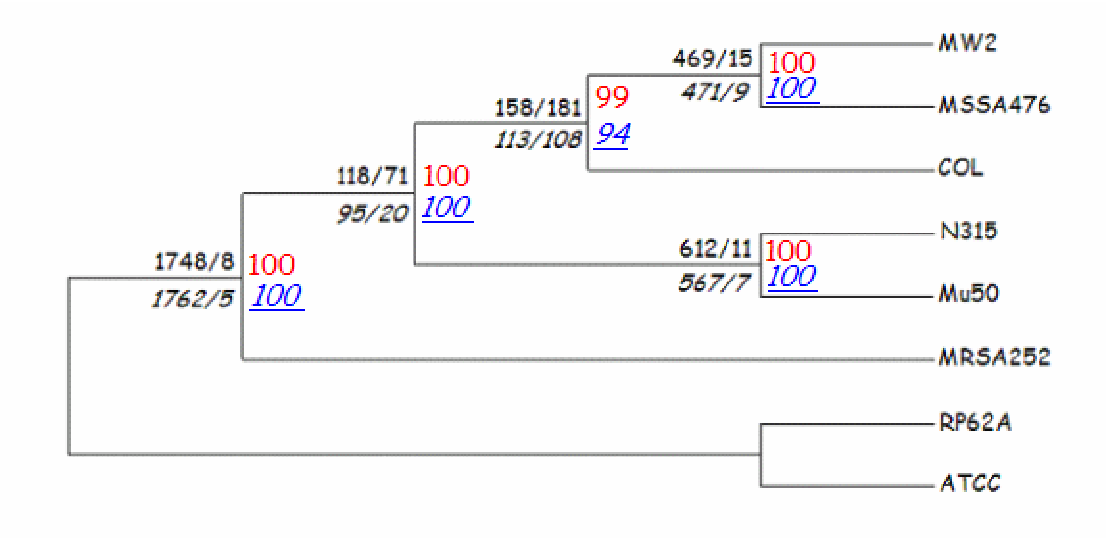

Figure 12. Supertree with measures of support by concordance/discordance and bootstrap percentages. The cases on the top correspond to protein cases.

The bootstrap shows extremely high support for all branchings in the supertree. Almost every single protein bootstrap replicate produced the same bipartitions expect once when a bipartition of COL N315 Mu50 MRSA252 | MW2 MSSA476 RP62A ATCC was constructed. The ancestral node leading to COL MW2 MSSA476 | N315 Mu50 MRSA252 RP62A ATCC is violated and hence the bootstrap value of 99.

The nucleotide bootstrap produced similar support for all the branchings except for the COL MW2 MSSA476 | N315 Mu50 MRSA252 RP62A ATCC. Out of the six times the node was not constructed, 5 times a bipartition of COL N315 Mu50 | MW2 MSSA476 MRSA252 RP62A ATCC was proposed and 1 time a bipartition of COL MRSA252 RP62A ATCC | N315 Mu50 MW2 MSSA476.

Phylogenetic profiling

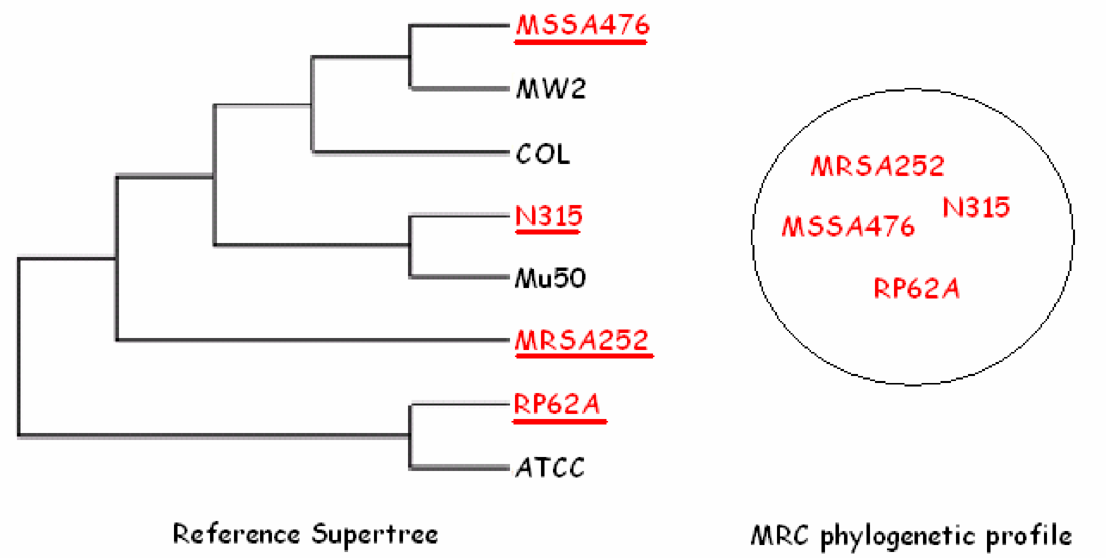

Figure 13. Representing a phylogenetic profile by gene loss events. The members of a MRC mapped onto the supertree represented by the red underlined *Staphylococcus*. 4 independent loss events are required to explain the MRC pattern.

The majority of the phylogenetic profiles are represented by MRCs of size 6 that contain all 6 *S. aureus* strains or MRCs of size 8 that contain all *Staphylococcus*, which contributed 397 of the 423 MRCs of size 6. Phylogenetic profiles were summarized by how many gene loss events were required by the supertree to the give rise to the phylogenetic profile.

Number of gene loss events	Number of MRCs
0	2147
1	247
2	59
3	15
4	2
Total	2470

Table 5. Number of gene loss events in MRCs necessary for explaining a phylogenetic pattern in the supertree

Functional characterization

	Protein			Nucleotide		
	Expected	Observed	Chi squared	Expected	Observed	Chi squared
Amino_acid_biosynthesis	10	8.71	0.19	7	5.33	0.52
Biosynthesis_of_cofactors	12	10.16	0.33	2	5.38	2.12
Cell_envelope	21	19.55	0.11	11	9.62	0.2
Cellular_processes	25	16.27	4.69	23	9.22	20.57
Central_intermediary_metabolism	4	4.81	0.14	2	2.71	0.19
DNA_metabolism	10	11.61	0.22	5	6.82	0.49
Energy_metabolism	19	18.33	0.02	10	11.15	0.12
Fatty_acid_and_phospholipid_metabolism	7	5.12	0.69	3	2.93	0
Mobile_and_extrachromosomal_element_functions	7	1.99	12.66	6	0.79	34.53
NULL	20	33.38	5.36	11	16.09	1.61
Protein_fate	9	11.76	0.65	4	6.12	0.73
Protein_synthesis	7	11.91	2.03	3	7.74	2.9
Purines	5	6.72	0.44	1	3.54	1.82
Regulatory_functions	18	11.76	3.31	10	5.86	2.93
Signal_transduction	2	3.21	0.45	2	1.92	0
Transcription	3	3.13	0.01	0	1.88	1.88
Transport_and_binding_proteins	42	33.91	1.93	21	19.11	0.19
Unknown_function	27	35.67	2.11	13	17.79	1.29
Sum ofChi-squared:	35.339546	17	DF	72.0955	17	DF

Table 6. Expected and observed functional roles assigned by TIGR and Chi squared values

Chi squared test for protein TIGR roles = 35.34

17 Degrees of freedom P value = 0.006

Chi squared test for nucleotide TIGR roles = 72.10

17 Degrees of freedom P value = 9.3 * 10e-9

The Chi squared test of TIGR role categories had statistically significant p values ≤ 0.05. Of the roles that were overrepresented in the data were roles involved in cellular processes and mobile and extrachromosomal element functions.

Discussion

The use of a phylogenetic approach for detecting LGT has clear benefits (Ragan, 2001) as well major disadvantages that have led to the development and use of other, non-tree-based methods (Eisen, 2000). However other methods, which employ similarity measures, are not based on any evolutionary model and hence can only provide an approximate estimate of the relationships of genes (Jain et al., 2002). Therefore phylogenetic conclusions, such as LGT, shouldn't be made on observations of similarity due to its inappropriateness to phylogenetics (Doolittle, 1999). In addition most claims of LGT in *Staphylococcus*, either among *Staphylococcus* or between *Staphylococcus* and other prokaryotes, have been based on similarity methods (Kuroda et al., 2001, Baba et al., 2002). Hence a phylogenetic analysis was performed to detect LGT within *Staphylococcus*.

The single most important goal of this investigation is to determine and find orthologous genes or proteins that are incongruent to a reference, which is the assumed natural history of all *Staphylococcus*. This discussion highlights the confidence that can be assigned to these incongruent cases and hence confidence to infer LGT events within *Staphylococcus*.

The accuracy of the supertree

The supertree of all 8 *Staphylococcus* used in this analysis shows relationships that are consistent with those proposed in the literature. Isolates N315 and Mu50 are both methicillin resistant *S. aureus* hospital isolates (Kuroda et al., 2001); MW2 and MSSA are two similar community strains (Baba et al., 2001, Holden et al., 2004); and ATCC and RP62A belong to a different species, *S. epidermidis.*

The three relationships (see above) implied by the literature are strongly supported by the ratios of concordant to discordant bipartitions. However, higher-order relationships within *S. aureus* are not strongly supported by protein data, with one of these two nodes not strongly supported by nucleotide data as well. The node ancestral to *S. aureus* COL, MW2 and MSSA476 is not strongly supported by either type of sequence data with concordance versus discordance values of 158 / 181 and 113 / 108 for amino acid and nucleotide data respectively. Relaxing the PP threshold to ≥ 0.90 recovered just as much discordance as concordances for both data (see Table 3). Thus there is evidence from nucleotide and amino acid data that this node is not well supported.

However the ancestral node leading to Mu50, N315, COL, MSSA476 and MW2, has much less discordance in proportion to concordance in nucleotide data (95 / 20) than in amino acid data (150 / 109). Relaxing the PP threshold to ≥ 0.90 had the same effect of increasing both concordant and discordant cases of this node (see Table 3). However what is interesting is the sensitivity of nucleotide bipartitions to PP thresholds (see Table 1). Much more bipartitions are recovered when relaxing the threshold in nucleotides for example relaxing the PP threshold to ≥0.90 from ≥0.95 yields around 700 bipartitions in nucleotide and 300 in amino acid sequences. Relaxing the threshold further to ≥0.75 yields a total of 5056 bipartitions in nucleotide sequences compared to the 3982 seen in amino acid sequences (see Table 1). Thus what is postulated, though not tested thoroughly, is that there is less discordance in nucleotide sequences due to the fact that a ≥0.95 threshold removed many cases potential cases of discordances. This can be shown by how much consistency is gained from adjusting the PP threshold. At ≥0.95 threshold there is a lot of unresolved conclusions between nucleotide and amino acid sequences (see Table 4). Thus by looking at consistencies between lower thresholds the hypothesis that the ≥0.95 threshold removes a lot of nucleotide discordances.

The phylogenetic profiles of the MRCs of size 5 were either a profile containing all 6 *S. aureus* except COL (49 out of 132 MRCs of size 5), hence showing that COL is less related to the other five *S. aureus* or a profile containing all 6 *S. aureus* except MRSA252 (54 out of 132 MRCs of size 5) showing that MRSA252 is less related to the other five *S. aureus*. The results from these phylogenetic profiles also highlighted the inconsistency between the phylogenetic relationships of MRSA and COL to the other isolates.

Bootstrapping of the MRP/sfit/qfit supertree with amino acid and nucleotide trees was also used as an accuracy measure of the supertree. However, results of the bootstrap are displeasing due to the high support for nodes that are not highly supported by the bipartitions in the MRCs. It should be noted that the theory of bootstrap rests on assumptions that are not natural to molecular sequences (Brocchieri, 2001). The ancestral node leading to MW2, MSSA476, COL, N315 and Mu50 was given a bootstrap of 100% for both sets of sequences and the ancestral node leading to MW2, MSSA476 and COL have bootstrap values of 99% and 94% for amino acid and nucleotide sequences respectively. However in a simulation study (Alfaro et al., 2003) bootstrap and PP proportions were quite different and the conclusion was that these support values measure different features of the data and generally cannot be compared together.

However the most illuminating evidence that this supertree is the best representation of all *Staphylococcus* was done by analysing 14 other topologies. Nodes with high support such as the ones proposed by the literature were collapsed, thus yielding a 5-taxon tree and 14 other topological relationships. The concordance versus discordance values of each tree showed that the supertree proposed by MRP/sfit/qfit best maximizes the concordant versus discordant bipartitions (data shown in laboratory book 1 page 106 – 109).

Phylogenetic artefacts

Phylogenetic artefacts are non-phylogenetic signals and thus need to be removed so that any signal is reflective of a phylogenetic signal. Phylogenetic artefacts can be caused by poor alignments, misuse of phylogenetic reconstruction methods, and poor data (Eisen, 2000). Other phylogenetic artefacts can be due to sensitivity to unequal evolutionary rates and bias in species sampling (Brocchieri, 2001). The use of a gamma distribution in the Bayesian analysis aims to sample different evolutionary rates. The lack of *Staphylococcus* species included in this study may cause artefacts due to a small sample size. All forms of artefacts must be identified in an analysis before making conclusions that LGT is the contributor of incongruent phylogenetic trees.

Relationship of discordant bipartitions to cluster support

The cluster support of a MRC can be related to as being the relative strength of correctly identifying orthologues. Hence it is important to know if weakly supported clusters represent discordant bipartitions more often. The distribution of cluster support was skewed towards higher cluster support values (see Figure 9), of which showed a median of the distribution of 0.67. Consistent discordant MRCs among nucleotide and amino acid sequences were used as the representation of discordant cases. The median of the distribution of 0.56 was observed for cluster support of these discordant MRCs, which shows a slightly lower distribution of cluster support. A statistical test was not done to assess the significance of this difference but however it is noted that discordant MRCs represent MRCs that have a lower cluster support.

Relationship between difficult alignments to discordant MRCs

The mean average of all mean pairwise percent identify scores, which is an estimation of the similarity of sequences, was around 89%. The difficulty to align sequences is directly proportional to how diverged the sequences are, so therefore aligning sequences of *Staphylococcus* shouldn't represent a problem. A different alignment produced by one program to another is also indication of a difficult alignment problem, as there is more than one way to align the set of sequences. This assumption was used and T-Coffee alignments were compared to ClustalW alignments. A normalised score from 0 - 1 was produced to assess how different the alignments were, where a score of 1 is interpreted as an exact alignment produced by both methods. A mean average of all normalised score was 0.995 showing that both methods produced extremely similar alignments. Since the majority of alignments will be similar, the mean average normalised score for discordant MRCs would probably be very near 0.995. The mean of all normalised score for consistent discordant MRCs among nucleotide and amino acid was 0.988. Since most MRCs were aligned the same way, discordances are not related to difficult alignment problems.

Confidence of discordant cases caused primarily by LGT

If all orthologous relationships have been correctly identified and inference of trees were supported by significant statistical values, and all cases of phylogenetic artefacts removed, then we can confidently say that LGT is the sole contributor to the incongruent phylogenetic trees. However it is all but impossible to find all phylogenetic artefacts as they can be caused by a wide variety of biases. Though if the artefacts considered in this project (i.e. poor alignments and missed paralogs) and the PP offers a good statistical support then preliminary conclusions can be made that these discordant bipartitions are indeed caused by LGT.

Functional properties of putatively transferred cases

In an analysis of 116 prokaryotic genomes, a study (Nakamura et al., 2004) investigated whether genes implicated in LGT have a biased biological function. The detection method they used was built primarily on differential nucleotide patterns in genomes; hence as they noted detecting recent transfer events. They showed that particular TIGR functional roles are over represented in their putative LGT genes. Comparisons of the functional roles over represented by discordant MRCs in this project are consistent with some of their results even though the detection methods differ. Statistically significant (p value ≤ 0.05) Chi squared results were obtained with the TIGR functional roles for discordant protein and nucleotide MRCs. Mobile and extra chromosomal elements were shown by far to be the most over represented functional role. The subclasses of this role are plasmid functions, prophage functions, transposons functions and "others". The subclass that represents most of the mobile and extra chromosomal elements was prophage functions. This is highly consistent with the idea that LGT in *S. aureus* are believed to be transferred by transduction, since conjugative plasmids are relatively uncommon and S. aureus are not naturally competent (4). Bacteriophages are also believed to have shaped much of the *S. aureus* since most strains have also shown to carry at least one bacteriophage (4). The next overly represented functional role is cellular processes, which encompasses a wide range of functions from toxin production to detoxification. The subclass of cellular processes most represented was pathogenesis, which is fairly consistent with much literature of pathogenicity genes being transferred (Lindsay and Holden, 2004).

The only over represented COG letters were cases where no COG letter was assigned to the MRC, represented by NULL cases. The other over represented case was the COG letter S that represents unknown functions. Since TIGR role categories were assigned to a much larger proportion of MRCs they provided a much larger sample. Also it should be noted that the COG letters were based on MRC clustering of 144 prokaryotes genomes (Beiko et al. 2005a), although the implications of this have not been considered since the results were uninformative, a much more systematic methods of annotation is preferred such as the one provided by TIGR.

Conclusion and further directions

The supertree of all *Staphylococcus* provides a general view of the overall relatedness of all the 6 *S. aureus* species. Some nodes have been found to be problematic and this could be due to the forcing a relationship of the 6 *S. aureus* when such a small species sample is used. However of all representations considered this was considered the best representation based on concordant and discordance values. Though bootstrap values provided extremely high support for this reference, the understanding of bootstrap values to PP to this dataset is not well understood as they can represent different support for the dataset. Taking all information into account the supertree offered a good general representation of the *Staphylococcus* to which individual orthologous trees could be compared.

All orthologous trees were based on a ≥0.95 PP threshold, thus this provided a good support to the creditability of the tree (see *reference tree construction* section for justification of using this threshold). As discordances could be associated to other factors other than LGT, preliminary investigation was conducted. Some of the considered artefacts were not seen to contribute much to the discordances on a general level. Thus discordances were considered to be a consequence of LGT.

Discordances are evidence for putatively transferred genes among *Staphylococcus*. Their over representation of functions implicated in Mobile and extra chromosomal elements is an exciting result though much work is needed to further analyse the relationships of discordance and other factors. Phylogenetic artefacts are vast and further work relating the methods used in this study need to be done. This could include the choice of orthology definition (i.e. by a hybrid clustering approach) or the use of a Bayesian inference of trees. Although 2470 orthologous relationships were made, the representation by 8 species is still small. Further studies would include more genomes into the study such as the two newly sequenced *Staphylococcus* genomes. Since this is within a subset of Staphylococcus, detecting LGT from other organisms would be interesting, thus further analyses done could include genomes of *Bacillus* since literature points (Kuroda et al., 2001, Gill et al., 2005 and Holden et al., 2004) out that there has been a lot of transfer among these genuses.

LGT in *Staphylococcus* is associated with certain islands (Novick et al., 2001), which lie in physical positions in the genome. Analysing the positions of putative laterally transferred genes can yield important information about whether a certain part of the genome is more likely to be involved with transfer. Another interesting question is whether laterally transferred genes are transferred as a cluster. If genes are transferred in clusters, the positions of LGT cases should be in close proximity to each other.

Virulence and antibiotic resistant genes are known to be transferred frequently and genes considered to be putatively transferred can be investigated at much more depth i.e. at a gene level rather than a functional category. Genes known to be implicated with LGT such as MecA (Ito et al., 2003), which confers resistance to methicillin, can be compared with results in this study. An agreement between other studies adds confidence to the results in this project and adds further questions which can be investigated.

Bibliography

Alfaro, M. E., Zoller, S. & Lutzoni, F. (2003). Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol Biol Evol* **20**, 255-266.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.

Baba, T., Takeuchi, F., Kuroda, M. & other authors (2002). Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet* **359**, 1819-1827.

Beiko, R. G., Chan, C. X. & Ragan, M. A. (2005b). A word-oriented approach to alignment validation. *Bioinformatics*.

Beiko, R. G., Harlow, T. J. & Ragan, M. A. (2005a). Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A*.

Bishop, M. J., & A. E. Friday. (1987). Tetrapod relationships: the molecular evidence. In *Patterson, C (ed.) Molecules and Morphology in Evolution: Conflict or Compromise?*: 123-139. Cambridge: Cambridge University Press.

Brocchieri, L. (2001). Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol* **59**, 27-40.

Creevey, C. J. & McInerney, J. O. (2005). Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* **21**, 390-392.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, pp. 345-352. Edited by M. O. Dayhoff. Washington, D.C.: National Biomedical Research Foundation.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129.

Eisen, J. A. (2000). Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr Opin Genet Dev* **10**, 606-611.

Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3**, 722-732.

Gill, S. R., Fouts, D. E., Archer, G. L. & other authors (2005). Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant Staphylococcus aureus strain and a biofilm-producing methicillin-resistant Staphylococcus epidermidis strain. *J Bacteriol* **187**, 2426-2438.

Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**, 2226-2238.

Gribaldo, S. & Philippe, H. (2002). Ancient phylogenetic relationships. *Theor Popul Biol* **61**, 391-408.

Harlow, T. J., Gogarten, J. P. & Ragan, M. A. (2004). A hybrid clustering approach to recognition of protein families in 114 microbial genomes. *Bmc Bioinformatics* **5**.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919.

Hiramatsu, K., Watanabe, S., Takeuchi, F., Ito, T. & Baba, T. (2004). Genetic characterization of methicillin-resistant Staphylococcus aureus. *Vaccine* **22 Suppl 1**, S5-8.

Holden, M. T. G., Feil, E. J., Lindsay, J. A. & other authors (2004). Complete genomes of two clinical Staphylococcus aureus strains: Evidence for the rapid evolution of virulence and drug resistance. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9786-9791.

Holder, M. & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet* **4**, 275-284.

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310-2314.

Ito, T., Okuma, K., Ma, X. X., Yuzawa, H. & Hiramatsu, K. (2003). Insights on antibiotic resistance of Staphylococcus aureus from its whole genome: genomic island SCC. *Drug Resistance Updates* **6**, 41-52.

Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. (2002). Horizontal gene transfer in microbial genome evolution. *Theor Popul Biol* **61**, 489-495.

Jones, D. & Sneath, P. H. (1970). Genetic transfer and bacterial taxonomy. *Bacteriol Rev* **34**, 40-81.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-282.

Jukes, T.H. and C.R. Cantor. (1969). Evolution of protein molecules. In Mammalian protein metabolism (ed. H.N. Munro), pp. 21-132. Academic Press, New York, NY

Kuroda, M., Ohta, T., Uchiyama, I. & other authors (2001). Whole genome sequencing of meticillin-resistant Staphylococcus aureus. *Lancet* **357**, 1225-1240.

Lindsay, J. A. & Holden, M. T. G. (2004). Staphylococcus aureus: superbug, super genome? *Trends in Microbiology* **12**, 378-385.

Lio, P. & Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome Res* **8**, 1233-1244.

Morschhauser, J., Kohler, G., Ziebuhr, W., Blum-Oehler, G., Dobrindt, U. & Hacker, J. (2000). Evolution of microbial pathogens. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **355**, 695-704.

Muller, T. & Vingron, M. (2000). Modeling amino acid replacement. *J Comput Biol* **7**, 761-776.

Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**, 760-766.

Novick, R. P., Schlievert, P. & Ruzin, A. (2001). Pathogenicity and resistance islands of staphylococci. *Microbes and Infection* **3**, 585-594.

Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217.

Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-4288.

Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K. & White, O. (2001). The Comprehensive Microbial Resource. *Nucleic Acids Res* **29**, 123-125.

Ragan, M. A. (2001a). On surrogate methods for detecting lateral gene transfer. *Fems Microbiology Letters* **201**, 187-191.

Ragan, M. A. (2001b). Detection of lateral gene transfer among microbial genomes. *Current Opinion in Genetics & Development* **11**, 620-626.

Raghava, G. P., Searle, S. M., Audley, P. C., Barber, J. D. & Barton, G. J. (2003). OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics* **4**, 47.

Suzuki, Y., Glazko, G. V. & Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc Natl Acad Sci U S A* **99**, 16138-16143.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D. & other authors (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278**, 631-637.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680.

Thompson, J. D., Koehl, P., Ripp, R. & Poch, O. (2005). BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins* **61**, 127-136.

Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691-699.

Whittam, T. S. & Bumbaugh, A. C. (2002). Inferences from whole-genome sequences of bacterial pathogens. *Current Opinion in Genetics & Development* **12**, 719-725.

Woese, C. R. & Fox, G. E. (1977). The concept of cellular evolution. *J Mol Evol* **10**, 1-6.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39**, 306-314.

Zhang, Y. Q., Ren, S. X., Li, H. L. & other authors (2003). Genome-based analysis of virulence genes in a non-biofilm-forming Staphylococcus epidermidis strain (ATCC 12228). *Molecular Microbiology* **49**, 1577-1593.

Zuckerkandl, E. & Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theor Biol* **8**, 357-366.