TESI DOCTORAL UPF 2012

Analysis of multiple protein sequence alignments and phylogenetic trees in the context of phylogenomics studies

Salvador Jesús Capella Gutiérrez

Director

Dr. Toni Gabaldón

Comparative Genomics Group
Bioinformatics and Genomics Department
Centre for Genomic Regulation (CRG)

UNIVERSITAT POMPEU FABRA
Departament de Ciències Experimentals i de la Salut

CRG Centre for Genomic Regulation

A todos los que me han ayudado a llegar hasta aquí,
aunque en especial a mis padres.

Acknowledgments

Érase un 31 de Agosto de 2008 cuando llegué a Barcelona dispuesto a empezar una nueva etapa de mi vida con expectativas y temores por igual ante los retos que tenía por delante. 4 años después sólo puedo decir que la experiencia ha valido la pena. Durante este tiempo he aprendido que no sólo es importante las metas que conseguimos sino cómo las conseguimos. Y alcanzar nuestras metas depende en gran medida de las personas que nos rodean. Creo, profundamente, que todas las personas dejan su huella en nosotros, algunas son leves y otras durarán de por vida, aun así, todas son importantes. A todas las personas con las que he interactuado desde una simple sonrisa hasta con las que me veo a diario, mi más sincera y humilde muestra de gratitud: ¡GRACIAS!

Me gustaría empezar este apartado de la tesis, el último que escribo como muchos otros han hecho antes que yo, agradeciendo profundamente al gran artífice de esta experiencia: **Toni Gabaldón**. Aunque tengo muchos calificativos para ti (todos buenos), me gustaría empezar dándote las gracias por apostar por mí, un ingeniero un tanto despistado lleno de grandes sueños y sin experiencia. A través de mil y una experiencias, alguna que otra variopinta, has guiado mi crecimiento como científico y sobre todo, como persona. Gracias por todas las horas, esparcidas en miles de "¿tienes un minuto?", que me has dedicado, muchas veces para resolver pequeñas dudas. Hacer el doctorado contigo ha sido como aprender a ir en bicicleta. Bajo tu supervisión he ido evolucionando desde el principio cuando tenía terror a tomar cualquier decisión a ir, progresivamente, adquiriendo más confianza y conocimientos hasta hoy, después de tanto tiempo, que tengo la certeza de tener las herramientas necesarias para abordar cualquier problema por difícil que pueda parecer. Gracias por todas esas conversaciones, muchas con carácter científico, dónde

he podido expresar, por tontas que fuesen, todas mis dudas y preguntar todas mis "curiosidades". Aunque podría escribir otra tesis describiendo todas nuestras anécdotas, me quedo con algo muy importante: el trabajo bien hecho no está reñido con la diversión, todo tiene su momento y lugar, y ya sea en el lab o fuera de él, podemos dar lo máximo de nosotros en cada situación.

Mi querida **Marina**, amiga y compañera, gracias por compartir tantas y tantas experiencias, por ser mi gran confidente y, con el tiempo, mi compañera de piso. Resulta difícil definir en unas pocas palabras mi gratitud por tantos pequeños grandes detalles. Quizás una de las cosas que más agradezco de todo este tiempo es tu visión, mucho más realista del mundo, frente a mi optimismo desbordado. Me has salvado de muchos fiascos y, en cierta forma, me has ayudado a madurar moderando mi ímpetu. Sinceramente, gracias por soportarme durante este tiempo de forma tan paciente. **Jaime**, junto con Marina y Toni, miembros del grupo original. Gracias por 1) No enfadarte nunca a pesar de todas nuestras conversaciones/discusiones. 2) Por convencerme para dar el salto a python, 3) Por ser una fuente constante de pequeñas ideas para mis proyectos, 4) Por ETE (y todas las mejoras que me han facilitado la vida), 5) Por nuestras discusiones de política que no salvarán el mundo pero que, al menos, nos entretienen y sobre todo, por enseñarme a tomarme la vida con más calma y menos formalidad.

Ester (sin h, ya lo sé) aunque has sido la última en llegar a nuestro grupo, me encanta ese toque de alegría y energía positiva que transmites a todos los que estamos a tu alrededor. Gracias por hacer que el trabajo experimental me resulte atractivo aunque no hubiese tocado una pipeta en mi vida hasta hace unos meses.

My dear colleagues and friends, Leszek and Alexandros, thanks for all the time we have shared during our theses. **Leszek**, thanks for nice discussions, all the scripts I borrowed (although I ended up re-writing many of them) and, more important, for showing me it is possible to do a PhD without sacrificing all the fun and a life outside of the lab. **Alex**, ¿qué tal tu español? ¿estás practicando?, thanks for all those funny moments we shared along the day, they are very important for me. When I look at you, I can find

many parallelisms between you and me when I was starting, even in our chaotic way of waiting till the very last minute to do everything. And, of course, thanks for showing me that Greek is still a language. **Gabriela**, **Fran** and **Damien**, thanks for many inspiring conversations during this time. You have showed me how a postdoc is (hopefully my next step) and how we can start driving our scientific career.

Although it is easy to say "thank you" to any group leader, Id like to thank **Cedric** and **Fyodor** for showing me that being a geek is cool and that we always should go to prove our hypothesis and ideas without caring how strange they are.

Roderic, moltes gràcies (aquesta és la primera vegada que escric en català) per tot el temps que hem compartit en el departament des d'aquella llunyana entrevista per entrar al CRG. Encara que estàs sempre ocupat, m'ha sorprès molt gratament la teva proximitat i el tenir sempre 2 minuts per qui ho necessita (encara que sigui caminant pel passadís de camí d'una reunió a una altra). **Romina**, senzillament donar-te les gràcies per ser l'àngel de la guarda de tots nosaltres, gràcies per fer-nos la vida tan fàcil encara que no ho sapiguem apreciar sempre.

During this time I think I have made some great friends in the department: Michael, Marco and Joao. **Michael**, for instance, thanks my dear friend for helping me to improve my English, for sharing your time and thoughts about many things and for stimulating my curiosity to go beyond the evident. **Marco**, it has been a pleasure to work with you and, what is even better, to share time with you out of the lab. It has been very nice to discover there is an artist in every scientist (Im still looking for it). **Joao**, although I usually make jokes about your work, I know you work (a lot?). Thanks for being around and sharing plans and ideas of things we can do to enjoy our time. I cannot talk about my department, Bioinformatics & Genomics, without thinking in many nice people who have made this a very nice experience.

A **Carla**, un grato descubrimiento desde la tierra que me vio nacer, por todos esos momentos de desconexión durante todo el tiempo que he estado escribiendo la tesis. No sé si será la magia venezolana pero ha sido muy

divertido y desestresante compartir uno, dos y hasta tres cafés al día.

In the last four years we have met once a year to discuss how my thesis has been going, Id like to acknowledge to my thesis committee members: **Ben**, **Cedric** and **Eduardo** for very useful comments, ideas and suggestions for my thesis, and sometimes, beyond the thesis.

No sólo he hecho buenas migas con gente del departamento, también con gente de la 4ta planta del PRBB. Me gustaría aprovechar para agradecer a **Elena Carnero** por tantas y tantas experiencias compartidas dentro y (sobre todo) fuera del lab, sé que a veces no soy fácil de convencer para dejar el ordenador y salir del lab, y mucho menos, para dejar de lado mi lado más formal. Aprofito també per donar les gràcies a la **Macarena**, entre altres coses pel template de la tesi, però sobretot per aquelles tardes jugant a tennis de taula en els retreats (crec que m'he copiat això de la teva tesi).

During my thesis I haven't worked only in the dry lab but also in the wet lab, so I'd like to thank all the people who made my stay in Kaiserslautern, Germany, a very nice experience (Frank, Antje, Katha, Claudia, Anja, Kevin, Bart, Laura among others). It was a great time where I learnt a lot of Germany and Germans; guys, you are funny and laugh a lot. Thanks for accepting me as a member of your department, although I didn't have any idea about German (I'm still trying to remember the numbers). I'd like to especially thank **Antje** for helping me (a lot) with my wet lab work, for nice chats about everything and for sharing her time watching the Euro Cup (I'm sorry but I told you that Spain would win the cup). I'd like also to thank **Katha** for her time in the city, please, stay strange, and **Anja** for being always caring about me, for showing me nice places in K-town and sharing time out of the university.

Mis queridos y apreciados "niños" valencianos, gracias por hacer especial todo el tiempo que hemos pasado juntos en Valencia, en Barcelona y en muchos más sitios. Me he sentido muy arropado y valorado por todos vosotros siempre. Me gustaría mencionar con especial cariño a **Santi**, mi alumno tutorizado que ha terminado tutorizándome; a **Cris**, una amiga muy especial que no importa ni dónde (literalmente), ni cuándo pero siempre está ahí; a **Lola**, por darle un toque alegre (y rosa) a la vida; a **Don Carlos**

(para mí, Carlitos), porque hemos compartido un montón de cosas casi sin darnos cuenta. Tampoco quiero dejar de mencionar a **Dani**, ya van casi 10 años conociéndonos y aunque la distancia nos ha separado, la amistad es la misma. Cierro este capítulo agradeciendo a **Javi Moralejo** por estar siempre pendiente de mí. Y me dejo muchos más en el tintero, ¡buscaré la forma de compensarlo!.

Mi querida **Marta**, sólo puedo decirte ¡¡¡gracias!!!, desde lo más profundo de mi corazón, por compartir tantos buenos y malos momentos a lo largo de estos años. Me has ayudado mucho a crecer como persona, a valorar lo que realmente merece la pena y a luchar por lo que quiero. No sabría decir que % de esta tesis te corresponde pero ten por seguro que un trocito de ella es tuya.

Como he leído alguna vez, por último pero no menos importante, de hecho, diría que lo más importante, me gustaría agradecer profundamente a mi familia todo el apoyo que me ha brindado siempre. Terminar este doctorado no es más que un paso en el camino en el cuál **MIS PADRES** (en mayúsculas de forma intencionada) me han apoyado incondicionalmente con todo su amor y su cariño, dónde han depositado su total confianza en mis decisiones y me han animado siempre, sobre todo en las horas bajas, a seguir adelante, a perseverar en mis metas.

Soy consciente de que he dejado muchas personas fuera de estos agradecimientos, a ellas estas palabras como señal de gratitud y mis más sinceras disculpas por no detenerme un minuto a dedicarles unas palabras. Me gustaría haberlo hecho pero eso implicaría escribir una tesis entera.

No puedo cerrar este capítulo de agradecimientos sin mencionar al **Ministerio de Economía y Competitividad** de España, quien ha financiado mi doctorado (Beca: BES-2010-036260 - Proyecto: BFU2009-09168) a través de su programa de "Formación de Personal Investigador".

Salvador Jesús Capella Gutiérrez.

Barcelona, September 2012.

Abstract

Phylogenomics is a biological discipline which can be understood as the intersection of the fields of genomics and evolution. Its main focuses are the analyses of genomes through the evolutionary lens and the understanding of how different organisms relate to each other. Moreover, phylogenomics allows to make accurate functional annotations of newly sequenced genomes. This discipline has grown in response to the deluge of data coming from different genome projects. To achieve their objectives, phylogenomics heavily depends on the accuracy of different methods to generate precise phylogenetic trees. Phylogenetic trees are the basic tool of this field and serve to represent how sequences or species relate to each other through common ancestry. During my thesis, I have centered my efforts in improving an automated pipeline to generate accurate phylogenetic trees and its posterior publication through a public database. Among the efforts to improve the pipeline, I have specially focused on the problem of multiple sequence alignment post-processing, which has been shown to be central to the reliability of subsequent analyses. Subsequently I have applied this pipeline, and a battery of other phylogenomics tools, to the study of the phylogenetic position of Microsporidia, a group of fast-evolving intracellular parasites. Due to their special genomic features, Microsporidia evolution constitutes one of the classical examples of challenging problems for phylogenomics. Finally, I have also used the pipeline as a part of a newly designed method for selecting robust combinations of phylogenetic gene markers. I have used this method for selecting optimal gene sets to assess the phylogenetic relationships within fungi and cyanobacteria, showing that the potential of these genes as phylogenetic markers goes well beyond the species used for their selection.

Resumen

Filogenómica es una disciplina biológica que puede ser entendida como la intersección entre los campos de la genómica y la evolución. Su área de estudio es el análisis evolutivo de los genomas y como se relacionan las distintas especies entre sí. Además, la filogenómica tiene como objetivo anotar funcionalmente, con gran precisión, genomas recién secuenciados. De hecho, esta disciplina ha crecido rápidamente en los últimos años como respuesta a la avalancha de datos provenientes de distintos proyectos genómicos. Para alcanzar sus objetivos, la filogenómica depende, en gran medida, de los distintos métodos usados para generar árboles filogenéticos. Los árboles filogenéticos son las herramientas básicas de la filogenómica y sirven para representar como secuencias y especies se relacionan entre sí por ascendencia. Durante el desarrollo de mi tesis, he centrado mis esfuerzos en mejorar una *pipeline* (conjunto de programas ejecutados de forma controlada) automática que permite generar árboles filogenéticos con gran precisión, y como ofrecer estos datos a la comunidad científica a través de una base de datos. Entre los esfuerzos realizados para mejorar la *pipeline*, me he centrado especialmente en el post-procesamiento previo a cualquier análisis de alineamientos múltiples de secuencias, ya que la calidad del alineamiento determina la de los estudios posteriores. En un contexto más biológico, he usado esta pipeline junto con otras herramientas filogenómicas en el estudio de la posición filogenética de Microsporidia. Dadas sus características genómicas especiales, la evolución de Microsporidia constituye uno de los problemas clásicos y difíciles de resolver en filogenómica. Finalmente, he usado también la *pipeline* como parte de un nuevo método para seleccionar combinaciones óptimas de genes con potencial como marcadores filogenéticos. De hecho, he usado este método para identificar conjuntos de marcadores filogenéticos que permiten reconstruir con alto grado de precisión las relaciones evolutivas en Cyanobacterias y en Hongos. Lo más interesante de este método es que evalúa la fiabilidad de los marcadores en especies no usadas para su selección.

Contents

1 Introduction

1.1 Phylogenomics.

Phylogenomics (Eisen and Fraser, 2003) is a biological discipline that has arisen during the last decade, as an increasing number of completely sequenced genomes have become publicly available. This discipline can be conceived as the intersection of the fields of genomics and evolution, or in other words, as the study of how genomes evolve and how this information can be used to understand to what degree different species relate to each other. This latter point, the reconstruction of evolutionary relationships across species, has attracted many efforts in the last 20 years, and a number of approaches have been proposed that aim to exploit the information provided by entirely sequenced genomes (Delsuc et al., 2005; Whelan, 2011). Gaining insight into the evolutionary relationships of species offers a unique opportunity to study diverse biological phenomena, ranging from the dynamics of gain or loss of gene families (Eirín-López et al., 2010), the horizontal transfer of genetic material among species (Gogarten and Townsend, 2005), to the conservation of a limited number of genes across all domains of life (Ciccarelli et al., 2006).

In addition, phylogenomics has been proven very useful for the prediction of the function of uncharacterized genes at a large-scale (Eisen, 1998; Gabaldón, 2008a). This constitutes an alternative to other in-silico based methods for the prediction of functional associations, an approach that is gaining relevance, as experimental characterization cannot cope with the increasing flow at which new genes are sequenced. The functional annotation of new genes has traditionally been done based on the transfer of function from the most similar hits in public databases. However, this approach has been shown to often lead to wrong annotations and, what is even worse, to the propagation of such errors across databases (Galperin and Koonin, 1998). Functional annotations based on orthology, rather than just homology, have been shown to be more accurate than simple sequence similarity-based approaches and thus can help to alleviate the above-mentioned problems (Brown and Sjölander, 2006). Orthologous sequences, those derived from a common ancestor by a speciation event, are generally less prone to functional shifts and, therefore, to conserve a greater functional similarity than

paralogous sequences, related by gene duplication events (Gabaldón, 2008b; Altenhoff and Dessimoz, 2009). Hence, orthologs constitute nowadays the most accurate source for functional annotation in newly-sequenced genomes. Finally, it must be noted that in-silico functional annotation, even when using phylogenomics, constitutes only a prediction and, thus, final confirmation would always require experimental verification.

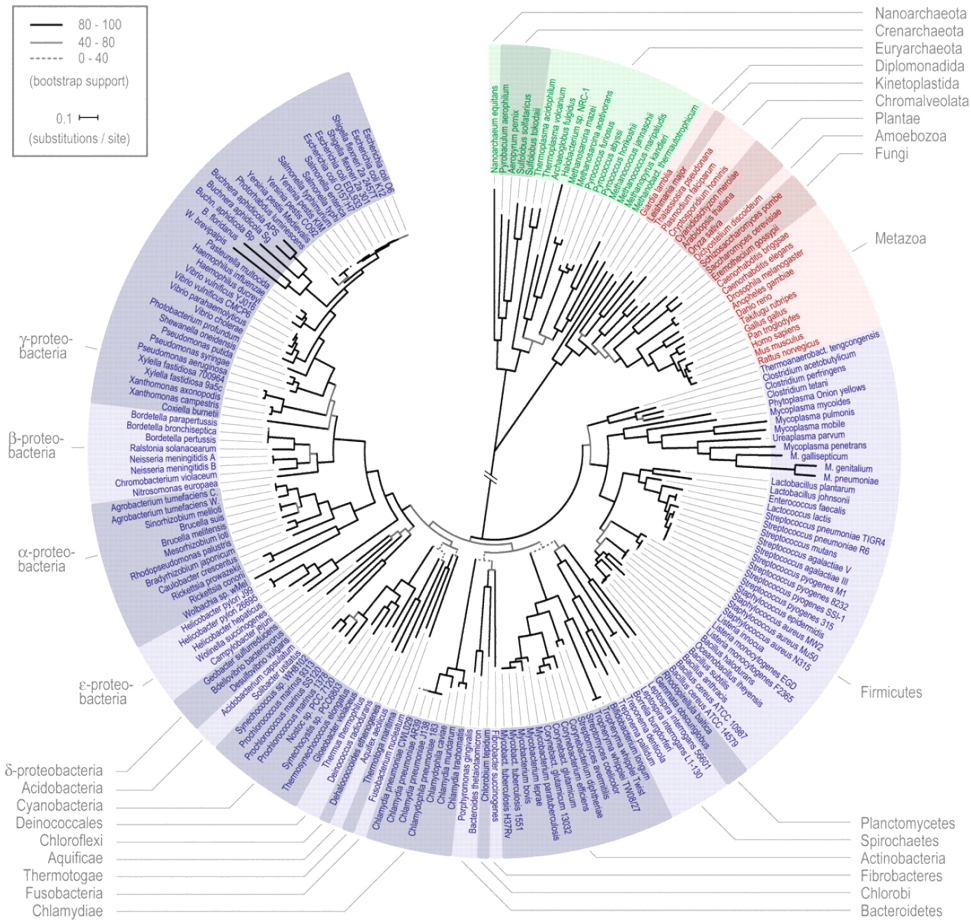

Figure 1.1: Tree of life reconstructed by Ciccarelli and colleagues using 31 proteins present in single copy across all domains of life. The tree includes 216 species, mostly prokaryotes. Image obtained from Ciccarelli et al. (2006).

Another field of research in which phylogenomics has been instrumental is that of the reconstruction of species relationships. This has been a focus of molecular phylogenetics since the times in which few molecular sequences

were available (Zuckerkandl and Pauling, 1965). A common limitation in classical phylogenetics, however, is the availability of few genes to resolve species relationships, which can ultimately lead to wrong conclusions. When a small number of positions is considered, random noise can affect the inference of different parts of the tree, resulting in artifacts and poor resolution (Delsuc et al., 2005). In this context, the advent of phylogenomics promised to overcome these limitations, since it enabled the use of hundreds of thousands or even millions of positions from several genes, thus resulting in highly resolved phylogenetic trees (Delsuc et al., 2005).

However, despite the fact that an increase in the amount of data clearly reduces sampling biases, it also has the drawback of reducing the homogeneity across sites, thus bringing the need for more complex models (Kumar et al., 2011). Thus, now when phylogenomics has opened the possibility to resolve long-standing questions about how species have evolved for large parts of the tree of life, the field faces enormous challenges regarding the development of new standards and methods that are able to deal with such amount of heterogeneous data. Model violations occur when assumptions made by the model are not met by the data. Under these circumstances, a phylogenetic inference can favor a wrong scenario with a strong support. Such cases have populated the literature since the early days, and prominent examples include the positioning of the fungal group Microsporidia at the base of eukaryotes (Corradi and Keeling, 2009) or the conflicting results regarding the existence of Coelomata or Ecdysozoa, two alternative hypotheses on how arthropods, nematodes, and chordates relate to each other (Holton and Pisani, 2010).

With the availability of tens, even thousands of genes, these model violations have become more evident leading to the realization that new methods are needed. Different strategies have been explored, such as mining the data to select the most-informative parts, or assessing the consistency of the results when using different kinds of phylogenomics approaches (Wolf et al., 2001). In addition, it would be desirable to have models that account for the heterogeneous and noisy nature of data. However, considering that we do not fully understand the data at hand, the design of new models is a challenging task.

Another area of intense development is the implementation of fully automated phylogenomic pipelines (Frickey and Lupas, 2004; Huerta-Cepas et al., 2007; Vilella et al., 2009). A phylogenomic pipeline is a set of programs working together with data flow monitorization and error control. In this way, the same workflow to reconstruct a single gene phylogeny can be extended to thousands of genes without human intervention. In addition, the deluge of data generated by these pipelines have lead to the development of different strategies to store, track and made it publicly available (Huerta-Cepas et al., 2011; Flicek et al., 2012).

Apart from the mentioned challenges derived from the need of developing new approaches for different areas ranging from the reconstruction and posterior refinement of Multiple Sequence Alignments, to the phylogenetic inference, to the correct prediction of orthologous and paralogous sequences, phylogenomics faces the additional problem of a lack of established benchmarks, which makes difficult the validation and comparison of current and newly developed methods. This has triggered different initiatives that aim at establishing frameworks for comparisons of distinct methods (Thompson et al., 2005; Gabaldón et al., 2009). In the absence of a method that clearly outperforms others, a possible solution to assess the strength of any result is to compare the level of agreement or disagreement when different sources of data and/or different methodologies are used. Indeed, it is expected that different approaches are affected by different artifacts, although, there is still a common, strong and recognizable evolutionary signal consistent among them.

Despite the fact that future progress in technologies may be needed to solve some of the open questions, phylogenomics is already playing an important role for society. Phylogenomics is not only enabling us to better understand species evolution, which is fundamental in the context of the global biodiversity crisis, but also some of its findings are being translated into diverse areas of applied sciences, such as biotechnology or biomedicine, which have the ultimate goal of improving our lifestyle.

In the following sections, I will review different methodological aspects which are crucial to understand how large amounts of data are processed in phylogenomics, and how these analyses can lead to the generation of new

biological knowledge. In particular, I will not only briefly introduce key concepts, but I will also highlight the main handicaps and benefits of the different methodological alternatives. All these will draw the appropriate context for a better understanding of the body of my own research, which will be described in the next chapters.

1.2 Sequences and homology.

Biological sequences, ordered chains of nucleotides (in DNA or RNA) or amino-acids (in proteins), constitute the central object of analysis in phylogenomics. Since the first method to obtain the nucleotidic sequence of a DNA molecule was introduced in the late seventies (Sanger et al., 1977), several approaches have been developed to obtain DNA sequences faster, cheaper and more accurately, leading us to the current high-throughput sequencing techniques that have revolutionized many areas of biology (Shendure and Ji, 2008).

Phylogenetics, and by extension also phylogenomics, refers to the evolutionary analysis of sets of sequences that are related by common ancestry, i.e. homologs (Eisen and Fraser, 2003). Thus a common initial step consists of retrieving sequences which share a common evolutionary history, in other words, a set of homologs. Since sequences related by common ancestry, rather than emerged de novo (Durbin et al., 1998), are expected to bear similarities, then it is possible to use different similarity-based tools such as BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990) or HMMER (Eddy, 2011) to identify set of homologous sequences.

The general idea behind BLAST (Altschul et al., 1990) and similar platforms is to find similar sequences to the one used as a query based on the number and length of highly similar fragments rather than based on the global similarity over the whole sequence. This strategy was chosen with the aim to optimize speed over sensitivity in order to scan databases containing a continuously growing number of sequences. To achieve an increase in the accuracy of the predictions, posterior steps to ensure general similarity can be performed. In contrast, alternative platforms, such as HMMER (Eddy, 2011), implement Hidden Markov Models (HMMs) to search for

similar sequences in databases. HMMs define highly sensitive models for sequence conservation at residue level. This strategy offers an appropriate probabilistic framework where sequences are found with higher precision at the cost of larger computational requirements. The high computational demands of this approach has traditionally limited its usage. However, several optimizations have been introduced recently into HMMER (Eddy, 2011) that make it approximately as fast as BLAST with a higher accuracy in homology detection.

Once a set of homologous sequences has been identified, it is a common practice to describe in more detail, i.e at residue level, the homology relationships among them. This is often achieved by reconstructing Multiple Sequence alignments

1.3 Multiple sequence alignments.

In the context of evolutionary analyses, the aim of a Multiple Sequence Alignment (MSA) is to represent sequences in a way in which homologous residues are aligned on top of each other (Kemena and Notredame, 2009). To achieve this, gaps are introduced in the alignment to represent the lack of homology in residues that would result from the insertion or deletion events in some of the aligned sequences. Since the real chain of past events, and therefore the true homology relationships, is generally unknown, alignments are reconstructed by exploring different homology scenarios and scoring them according to the physico-chemical similarity of the different residues. In the case of amino-acids, these similarity scores have been previously derived using manually curated data and stored in different matrices. Such matrices reflect how often one residue is found replacing another. In the case of nucleotides, however, the matrices are not empirical and reflect simple assumptions on nucleotide variation. Additionally, the introduction and extension of gaps is penalized, using scores (or penalties) that are rather arbitrary. Ideally, the perfect MSA is the one with the best final score among all possible scenarios.

Despite the simplicity of the algorithm, exploring all possible alignments of a relatively small set of sequences is an intractable problem, even for

the most powerful computers. Indeed, this belongs to the Non-Polinomical (NP) category of computational problems, which means that the time required to get the solution grows faster with the number of sequences than an exponential distribution (Wang and Jiang, 1994). To address this challenge, different heuristic approaches have been proposed over the last 30 years and more than 100 methods have been published since then, including genetic algorithms (Notredame and Higgins, 1996), Hidden Markov Modeling (Eddy, 1995) and the progressive alignment algorithm (Hogeweg and Hesper, 1984).

Nowadays, the progressive alignment algorithm is implemented in almost every MSA program. This algorithm is based on the general idea of aligning the most similar sequences first, proceeding then to more distant ones, while always performing pair-wise alignments. To make this possible, the algorithm makes a comparison, in terms of sequence identity, of all possible pair of sequences, and using that information constructs a raw binary tree, known as the guide-tree. This tree is used to guide the alignment process, so that sequences are aligned from the leaves to the root. A pervasive problem in progressive approaches is that mistakes made at any stage are propagated to subsequent steps. To solve this, most programs implement one or more iterative phases to refine and correct, if possible, any mistake. Nevertheless, even using this iterative phase, noise is still present in the final alignment, specially at the most variable regions. More than decade ago, in an attempt to improve this situation, the so-called consistency based algorithms were introduced (Notredame et al., 1998), and are implemented in popular programs such as T-Coffee (Notredame et al., 2000) or Probcons (Do et al., 2005). The general idea behind this approach is to optimize different scoring schemes for different parts of the sequences in order to reflect their diverse nature. To achieve this, a primary library is built, prior to the multiple alignment reconstruction, based on the local and global pairwise alignments of all input sequences. This primary library is then used in conjugation with the residue scoring matrices and the gap penalties to reconstruct the final alignment. Although consistency-based algorithms have supposed an important step ahead in terms of accuracy for all applications depending on MSAs, there is still room for improvement. One of these additional

improvements is based on an extension of the idea of consistency applied to the case of several MSA programs. This approach, implemented in M-Coffee (Wallace et al., 2006) exploits the fact that different programs may align sequences following different strategies and, therefore, a strong signal close to the real alignment may emerge from the combination of all of them. In M-Coffee, the primary library is built using information from previously generated alignments, and residue pairs are weighted according to how often they appear in the individual alignments.

MSAs are often used to infer phylogenetic trees, however, during the reconstruction process most of the heuristics tend to ignore the biological meaning of gaps. Hence, although gaps are supposed to represent events of insertion or deletion occurred in evolution, they are regularly introduced in MSAs to maximize a scoring function. Recently, a new sort of algorithm being aware of the phylogenetic placement of gaps has been proposed, PRANK (Löytynoja and Goldman, 2008), where gaps are inserted aiming to reflect these evolutionary events rather than to maximize any mathematical function. However, the placement of these gaps are highly dependent on the accuracy of the guide-tree, which is initially inferred from the pairwise distances of the sequences involved and eventually refined at a later step.

1.4 Post-processing of multiple sequence alignments.

Multiple sequence alignments are central to many applications in Bioinformatics, and downstream analyses are highly dependant on their accuracy. As mentioned before, given the huge space of possible solutions to be explored, even the most sophisticated programs can fail aligning part of the sequences. Often, these misaligned regions correspond to the most variable parts of the sequences. The impact of misaligned regions in posterior analyses, mainly in phylogenetic inference, was noticed for the first time in the nineties by Lake (Lake, 1991). Since then, different approaches have been proposed to alleviate this situation. Apart from improving heuristics and introducing new ways to reconstruct more accurate alignments, several strategies have been proposed to tackle this problem. The simplest alternative has been the manual removal of these conflicting regions by researches. Al-

though possible in a context of alignments of few sequences, this approach faces the important problem of lack of reproducibility (Castresana, 2000).

Alternatively, two automated strategies have emerged during the last decade that involve a certain post-processing of an initially reconstructed alignment: i) refinement of the conflicting parts and ii) removal of misaligned positions. While the former is in itself another iteration of the alignment process applied only to certain parts, the later directly discards regions that are thought to contribute more noise than signal. Removal of misaligned positions, also known as alignment trimming, has received more attention and it has been the subject of more developments. Both approaches have in common the challenge of correctly identifying misaligned positions in the alignment. This can be done based on several criteria and at different levels of complexity. The simplest one, and certainly the most widely used, is based on the fraction of gaps present in the alignment column, since it is expected that many gaps are introduced to optimize scoring functions instead of representing true biological events and, therefore, many aligned residues in gappy regions may be wrong. A straightforward procedure, when alignments involve very closely related sequences and have few gaps, is to remove all columns with gaps. However, it is often the case that this strategy ends up being too aggressive, not leaving enough information. Thus other restrictions that may also involve considering the proximity of the columns with gaps (i.e. blocks longer or shorter than a given size) are generally introduced.

Another criterion to identify misaligned positions is based on the physico-chemical similarity of the aligned residues, because less conserved parts of the sequences are known to be more prone to misalignment. These low conserved regions can be identified using either similarity scores derived from substitution matrices or more complex methods based on the detection of entropy levels. Instead of considering just one criterion, some methods can identified potential misaligned columns based on a combination of different criteria such as the ones mentioned above, and also adjusting the parameters to certain characteristics of the alignment such as the fraction of gaps, physico-chemical similarities, number of sequences and/or alignment size. Finally, it is possible to identify potential misaligned regions through

the level of consistency of residues pairs aligned across different programs and/or settings such as gap penalties, substitution matrices, etc. The rationale behind this approach is that robust residue pairs that are consistent across alignment strategies are not method-dependent and thus less-likely to be the result of a misalignment.

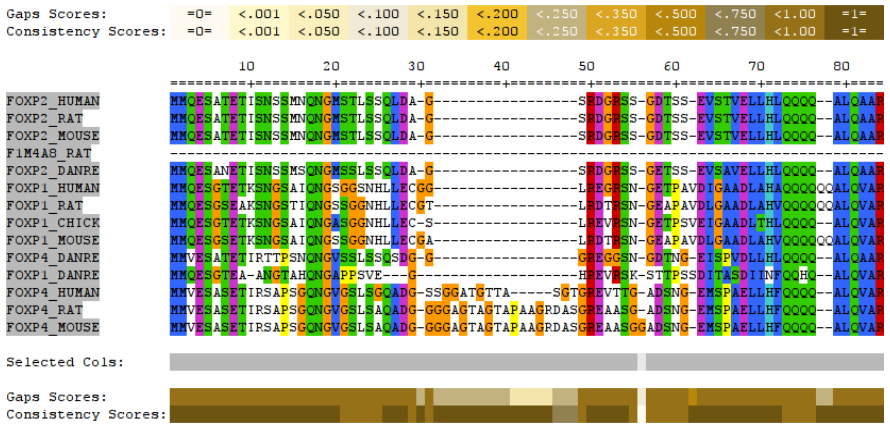

Figure 1.2: Partial alignment of the FOXP2 Human protein and its homologs in the context of vertebrate species. Bars below the alignment indicate whether the column is conserved, dark grey, for downstream analyses or not, light grey (just column 56), as well as different scores, in this case, the proportion of gaps and the consistency of residues in each column across several alignments. Image generated by trimAl v1.4 Capella-Gutiérrez et al. (2009).

Once potential problematic blocks have been identified in the alignment, one possible strategy to alleviate the situation is to realign again these blocks in an attempt to reduce their level of disagreement. The major drawback of this approach is that it uses similar optimizing functions as in the previous alignment phase and some misaligned regions may still remain after the optimization. Thus, when the inclusion of a region is more likely to mislead downstream analyses than providing true information it may still be necessary to remove it. It has been shown (Lake, 1991; Talavera and Castresana, 2007; Capella-Gutiérrez et al., 2009; Criscuolo and Gribaldo, 2010; Kück et al., 2010) that removing highly variable positions contribute significantly to improve phylogenetic reconstruction since the overall noise is minimized. Moreover, in some types of analyses when sequences are highly divergent, it is convenient to remove even some well-aligned regions to alleviate mutational saturation effect.

When applied in a phylogenomics context, these methods face the additional problem of the difficulty of dynamically adjusting the parameters in order to process a large number of alignments with a very heterogeneous range of sequence numbers, lengths and level of similarity. The appropriate selection of parameters is crucial to correctly identify those potential blocks of misaligned residues and maximizing the signal to noise ratio in the processed alignment.

Program	Approach	Reference
GBlocks	Trimming	Castresana (2000)
SOAP	Trimming	Löytynoja and Milinkovitch (2001)
Rascal	Refinement	Thompson (2003)
RF	Refinement	Wallace et al. (2005)
REFINER	Refinement	Chakrabarti et al. (2006)
Noisy	Trimming	Dress et al. (2008)
trimAl	Trimming	Capella-Gutiérrez et al. (2009)
GUIDANCE	Trimming	Penn et al. (2010)
BMGE	Trimming	Criscuolo and Gribaldo (2010)
ALISCORE	Trimming	Kück et al. (2010)
ZORRO	Trimming	Wu et al. (2012)
SeqFIRE	Trimming	Ajawatanawong et al. (2012)

Table 1.1: A survey of published programs to accurately identify and refine or remove conflicting regions in multiple sequence alignments

Several studies have shown the importance of removing conflicting regions in phylogenetics (Lake, 1991; Talavera and Castresana, 2007; Capella-Gutiérrez et al., 2009; Criscuolo and Gribaldo, 2010; Kück et al., 2010), and at least 10 different programs (see table 1.1 above) have been implemented to identify, and, subsequently, remove or refine poorly aligned regions. In contrast, some recent analyses have claimed that gaps carry phylogenetic signal that is systematically ignored (Dessimoz and Gil, 2010). This apparently conflicting results highlight the need of finding a proper balance between the removal of noise and true signal. In addition, it brings about the problem of identifying the phylogenetic signal carried by gaps and disentangling the effect that guide trees have in the placement of gaps. This is precisely the topic of the research presented in chapter 5.

1.5 Phylogenetic trees.

Tree structures have been used since the nineteenth century (Darwin, 1859) to represent relationships among kingdoms, species, morphological characters, and, in the last 50 years, sequences. It was in 1965 when Zuckerkandl and Pauling (Zuckerkandl and Pauling, 1965) proposed that either DNA, RNA or proteins can be used to establish evolutionary relationships among homologous sequences. To make such phylogenetic inference, it was clear from the beginning that homologous residues should be aligned on top of each other prior to any analysis. So, once an alignment is available, the next step is to represent the evolution of these homologous sequences, providing information of which of them diverged earlier or later. Because the origin of sequences through divergence from a common ancestor is a branching process, bifurcating trees provide an intuitive way of representing the evolution of a set of sequences, or taxa.

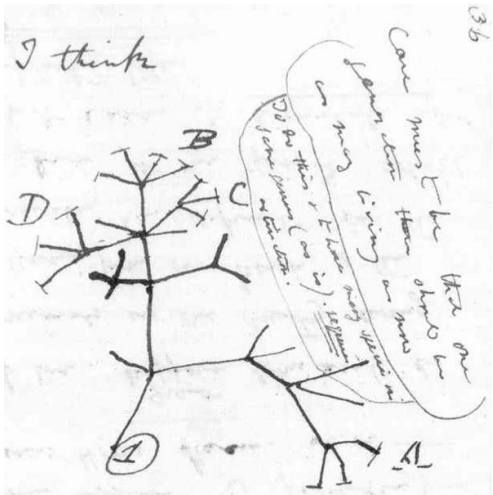

Figure 1.3: First use of a tree structure to explain the evolutionary relationships among species (as abstract entities) by Darwin in 1859 in his famous book "On the origin of species".

Nowadays, we count with different approximations to infer phylogenetic trees. Similarly to the reconstruction of multiple sequence alignments, the process of inferring such trees belongs to the category of Non-polynomial computational problems. In practice, this means that more or less sophis-

ticated heuristics have to be used to explore the space of tree solutions, to find the one that is optimal according to certain criteria. Based on the nature of the criteria used, the methods for phylogenetic inference are generally divided into three major categories: i) parsimony, ii) distance-based and iii) probability-based approaches, being the later the most popular in recent years.

1.5.1 Parsimony-based approaches.

Parsimony trees are reconstructed under the assumption that the real scenario is that which implies least changes along its branches. Thus, the closer the sequences, or the taxa, the more shared derived characters (synapomorphies) are expected, and this should lead to closer positions of these sequences in the tree. The aim of the process is to find the most parsimonious tree, in other words, the tree with the smallest number of changes among characters to explain the hierarchical relationships among them. The major drawback of this approach is the existence of multiple substitutions, since one observed change may actually involve many different past substitutions. This problem is particularly important when large evolutionary distances are considered. In addition the presence of unequal evolutionary rates may mislead parsimony. Indeed, sequences with faster evolutionary rates than others or long periods of divergence time can accumulate many changes and appear equally dissimilar to all the others. In such scenarios, parsimony will place divergence sequences next to each other, an artifact known as Long Branch Attraction (LBA) (Felsenstein, 1978) which especially affects this approach, albeit not exclusively. Given these drawbacks and the development of more suitable approaches, the usage of parsimony methods have decreased with time, and now are mostly restricted to the analysis of closely related sets of sequences, where the assumptions of the method are usually met.

1.5.2 Distance-based approaches.

In an attempt to avoid the use of heuristics for exploring the set of possible solutions, distance-based methods were introduced in the late eighties.

This category comprises methods such as neighbor-joining (NJ) (Saitou and Nei, 1987) or unweighted-pair-group (UPGMA) (Sokal and Michener, 1958; Murtagh, 1984). Trees are inferred based on pairwise distance comparisons among aligned sequences. Distances among all sequences are computed and stored as scores into matrices for its posterior usage. Scores do not only capture shared derived characters but also shared ancestral characters (symplesiomorphies) and unique derived ones (autapomorphies). The construction of such matrices leads to the direct inference of just one tree, thus these methods can easily deal with tens, hundreds or even thousands of sequences. However, they have the drawback of being very sensitive to the nature of the data. In other words, data assumptions should be fulfilled in order to get informative results. Methods such as UPGMA require the same divergence rates across all lineages in the tree to produce accurate phylogenies. Sampling is an important factor to have equal divergent rates across all sequences since slow or fast evolving sequences can be wrongly placed together even if they are distantly related. NJ can deal better with different evolutionary rates at different part of the tree, but the method can have problems to infer the expected phylogeny when different patterns of multiple residue substitutions are present in different parts of the alignment. Finally, but not less important, these methods are even more dependent than others on the alignment accuracy, since only accurate alignments would allow to compute precise distance scores.

1.5.3 Probabilistic-based approaches.

Probability-based reconstruction methods are those designed to find the best ranked phylogenetic tree according to its likelihood, as in Maximum Likelihood (ML) methods, or to its posterior probability, as in Bayesian approaches (Durbin et al., 1998). The general idea behind any probabilistic based method is to find the tree that better explains the observed data given an explicit model. The initial set of parameters depends on the selected approach but an input alignment and a predefined evolutionary model are at least necessary.

Evolutionary models describe the expected frequencies of residues and

probability that a residue changes into another. In the context of DNA, evolutionary models vary from the simplest one considering all possible substitutions equally probable (Jukes and Cantor model) (Jukes and Cantor, 1969) to more complex ones, which weigh differently distinct substitution patterns (Hasegawa, Kishino and Yano model) (Hasegawa et al., 1985). In contrast to mechanistic approaches used to construct DNA models, amino-acids ones are empirical, based on manually curated alignments that reflect different substitution patterns among amino-acids. The first model was introduced by Dayhoff and co-workers in the seventies (Dayhoff et al., 1978) and new ones are constantly being developed to be more general, for instance LG (Le and Gascuel, 2008), or more specific, e.g. MtArt (Abascal et al., 2007) or MOLLI60 (Lemaitre et al., 2011). An important step prior to any Maximum Likelihood or Bayesian inference is the selection of the appropriate evolutionary model. Depending on the nature of the data, the selection of evolutionary model that best fits the data varies slightly. On the case of DNA, trees are reconstructed using different models and then compared taking into account how many parameters (degrees of freedom) shows each model. On the case of proteins, trees are reconstructed under different evolutionary models, and scores are directly compared since models are empirically derived. In any case, given an equal amount of free parameters, it is assumed that the model with best score is the one which best explain the input data. Different approaches exist to select the best fitting method and properly penalizing for the complexity of the model (Posada, 2003).

Under a ML framework, tree inference starts by computing the likelihood of an initial tree, which may be a random topology. After this initial step, the algorithms explore the tree solution space making changes in the tree topology and recomputing its likelihood again until no further improvement is reached. Then, the tree with the best likelihood is returned as the one that best fits the input data under a predefined evolutionary model. Giving the statistical nature of the process, it is possible to compare different tree topologies, in term of their likelihoods, to see whether there are statistically significant differences among them. In this way, tests allow to discriminate among alternative scenarios, discarding those that

are not statistically significant. The usage of these tests are especially useful in complex scenarios such as when dealing with sequences with great evolutionary distances, alternative tree topologies with likelihood very similar, etc.

Although ML methods are a very powerful tool to infer phylogenetic trees under very complex scenarios, their accuracy can be compromised when i) data is too simple so that an overfitting effect can lead to infer wrong results because the different variables taken into account by the methods can measure minor fluctuations (random noise) rather than the real signal carried by the data, and ii) model assumptions are violated by the heterogeneous nature of the data such as different evolutionary rates, with huge differences, for different subsets of sequences with the effect of placing artificially sequences to the root. Recently, different strategies have been proposed to tackle the second condition such as developing models to account for different evolutionary rates at different part of the tree, the so-called co-varion model, or the usage of reduced alphabets in order to capture important changes, i.e. from hydrophobic to hydrophilic amino-acids, and reduce data complexity. However, there is still room to further improvement since these approaches are quite new and they have not been widely tested/used.

Rather than exploring the tree space while optimizing just one tree, bayesian inference samples several points of the tree space using a Markov chain Monte Carlo (MCMC) algorithm. Often more than one chain is used to reach the equilibrium state after sampling several times possibles phylogenetic trees inferred from the input alignment and the selected evolutionary model. The correctness of the phylogenetic tree is highly dependent on the convergence of the process. If convergence has been reached, then the phylogenetic tree reflects the most sampled topology in the tree distribution. If convergence is not achieved, less or nothing could be drawn from the phylogenetic tree obtained. Bayesian inference is one of the most accurate tool to reconstruct a phylogenetic tree but setting the convergence criteria and the enormous computational time needed to achieve convergence constitute its major drawbacks. Another drawback of Bayesian methods is the need to set the priors, that is the expectations, of observing certain

values of any given parameter before seeing the data. To avoid biases, uninformative priors, such as flat distributions, are usually chosen.

1.6 Single gene trees workflows.

Nowadays, the reconstruction of phylogenetic trees representing the evolution of gene families comprise three main steps which have been already described in this chapter: i) detection of homologous sequences, ii) reconstruction of the multiple sequence alignment, and iii) inference of the phylogenetic tree. Table 1.2 shows a survey of popular programs used at each of the different steps. The list includes two optional but important tasks: i) improvement of the alignment either optimizing again the conflicting parts or removing them, and ii) the model selection prior to infer a phylogenetic tree using any probabilistic based method.

Program	Step	Reference
Blast	Homology Search	Altschul et al. (1990)
Blat	Homology Search	Kent (2002)
HMMER	Homology Search	Eddy (2011)
ClustalW2	MSA:Progressive	Larkin et al. (2007)
Muscle	MSA:Iterative	Edgar (2004)
Mafft	MSA:Iterative/Consistency based	Katoh and Toh (2008)
T-Coffee	MSA:Consistency Based	Notredame et al. (2000)
M-Coffee	MSA:Meta-Aligner	Wallace et al. (2006)
Prank	MSA:Gaps aware placement	Löytynoja and Goldman (2008)
Gblocks	MSA:Trimming	Castresana (2000)
ProtTest	Phylogenetic trees: Model selection for proteins	(Abascal et al., 2005)
PhyML	Phylogenetic trees:ML	(Guindon et al., 2010)
RaxML	Phylogenetic trees:ML	(Stamatakis, 2006)
MrBayes	Phylogenetic trees:Bayesian	(Ronquist and Huelsenbeck, 2003)
PhyloBayes	Phylogenetic trees:Bayesian	(Lartillot et al., 2009)

Table 1.2: Popular programs used in the different steps needed to reconstruct a phylogenetic tree.

Once the phylogenetic tree is ready, depending on the biological question,

it can be used directly to make any inference or it can be the starting point for posterior analyses. Among possible analyses on a phylogenetic tree, I would highlight here the detection of speciation (orthology) and duplication (paralogy) events, the usage of sequences related by speciation events to make functional annotation of newly sequenced genes, the study of gene expansions in certain species, the detection of horizontal gene transfer cases, etc. All these analyses have been used during the present thesis.

1.7 Phylomes and their automated reconstruction.

A phylome has been defined as the complete set of phylogenetic trees for all proteins encoded in a genome (Sicheritz-Pontén and Andersson, 2001). Traditionally, homologous sequences from complete and incomplete sequenced organisms have been used to reconstruct these evolutionary relationships. However, it was realized that to fully understand the evolution of protein families, it is best to use only completely sequenced proteomes. Only complete genomes may reveal the complete dynamic history, i.e. gains and losses, for proteins in the considered species. A process similar to that used to reconstruct a single phylogenetic tree is used iteratively to reconstruct a phylome. In contrast to single gene tree workflows where the process can be done manually, for the reconstruction of this set of trees, the use of pipelines is needed. A pipeline, automated or not, is a system of computing programs that control, with minimal human intervention, the correct and ordered execution of the mentioned steps. Popular automated pipeline used in several projects comprises PhyloGenie (Frickey and Lupas, 2004), PhyOP (Goodstadt and Ponting, 2006) or EnsemblCompara GeneTrees (Vilella et al., 2009). Among the main differences between the use of a pipeline and the reconstruction of a single gene tree are i) the speed, necessarily a more important limiting factor in pipelines, and ii) the selection of parameters, since the pipeline should adjust automatically the parameters to accommodate the heterogeneity in the data. The first generation of automated pipelines tended to use default parameters for all trees resulting in generally poor performance. As a result, human intervention was needed to fix errors. With time the pipelines became

more sophisticated and incorporated programs that automatically adjust parameters (Huerta-Cepas et al., 2011) and/or directly used different sets of programs depending on the nature of the data (Vilella et al., 2009).

1.8 Downstream analyses: Inferring Species trees.

One of the major goals in Biology is to understand how species have evolved and how they related to each other. The first species tree based on sequence information was published in 1967 by Fitch and Margoliash (Fitch and Margoliash, 1967) in an attempt to relate animals and fungi using the mitochondrial protein cytochrome C. Although, the authors were successful on finding a common origin between these two groups, the major drawback of the approach was that cytochrome C was not present in all species. Ten years later, in 1977, Woese and Fox (Woese and Fox, 1977) published the first tree of life for the major living groups. They used ribosomal RNA, specifically the small subunit, since it was ubiquitous, showed high level of conservation and, what is even more important, it was already possible at the time to sequence it from diverse organisms. The most important finding of this first tree was the realization of the existence of a third domain of life: archaebacteria, lately renamed as archaea. Until then, only two major groups were accepted: eukaryotes and bacteria. With the time, ribosomal RNAs have become the classical markers in phylogeny to reconstruct species trees, especially in Bacteria where the transfer of genetic material is common, which leads to loss of phylogenetic signal. Ribosomal RNA has been proved to be useful to classify newly sequenced organisms and to establish relationships among different groups. However, its high level of conservation limits its power to accurately resolve deep phylogenies. This was noticed when trees drawn from newly sequenced individual genes were in conflict with the species tree inferred from ribosomal RNAs. It was in 2006, by Ciccarelli and colleagues (Ciccarelli et al., 2006), when a new attempt to draw a Tree of Life took place. In this case, widespread orthologous sequences from completely sequenced organisms were used. The general idea behind this approach was to use as much information as possible, taking advantage of the sequenced genomes at that moment.

Currently, there are different genome-wide approaches to infer species trees: i) alignment-free genome trees, ii) gene content trees, iii) gene order trees, iv) genome trees based on average sequence similarity, and v) phylogenomic trees (Snel et al., 2005). I will focus, therefore, on the phylogenomics approaches used to infer these species trees. Methods in phylogenomics for inferring species trees can be grouped into two categories (Figure 1.4): i) consensus trees from sets of single gene trees, and ii) concatenated alignment species trees. Both approaches make usage of genes rather than complete genomes. A genome is a much richer and complex source of evolutionary information than single genes and, therefore, a more difficult to align, a prior step before making any evolutionary inference, task. Any method for reconstructing species tree in a phylogenomics framework tries to use as much information as possible in order to infer real evolutionary events at species level. The idea behind a consensus tree is to find the tree that better explains the evolutionary history of the species considering the patterns observed in single gene trees. Most popular methods include the reconstruction of a consensus tree, which is the most compatible in terms of tree partitions with the total number of trees, e.g. CLANN (Creevey and McInerney, 2005), or finding the topology that is most parsimonious in terms of one or various events such as gene duplication, gene loss, or deep coalescence, GeneTree (Page, 1998) or iGTP (Chaudhary et al., 2010). In contrast, species trees based on concatenated alignments, also known as the super-matrix approach, are reconstructed after putting together alignments for single copy genes that are widespread in the species considered. Although there is much less data, it is expected a more consistent and recognizable phylogenetic signal since genes in this dataset are not under any evolutionary event such as gene duplications or losses that can affect the evolutionary signal. Accurately identification of genes belonging to this set is a crucial step to secure any phylogenetic inference. Counting as many genes as possible allow to minimize variation on phylogenetic signal due to methodological errors or different evolutionary rates and, therefore, increase the support for the species tree inferred. Another advantage of the concatenation approach is that it uses directly the information contained in the sequences rather than indirectly.

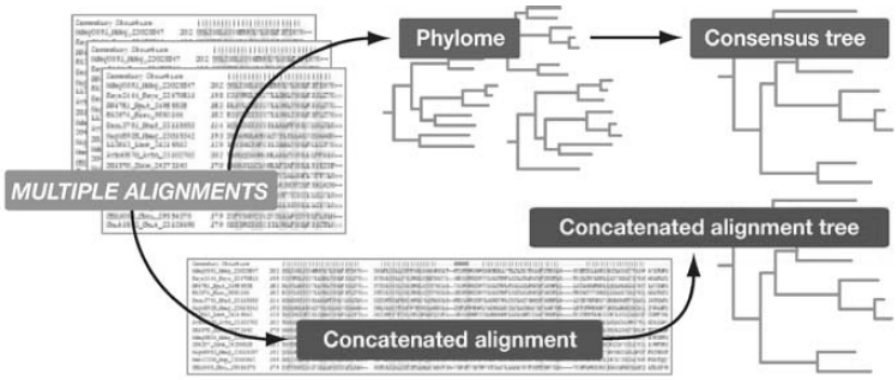

Figure 1.4: Different phylogenomics approaches for inferring species trees. Both methods try to use as much information as possible in two different ways: Consensus trees are reconstructed using all single gene trees available while trees inferred from concatenated alignments use direct information from widespread single copy gene among considered species. Image obtained from Snel et al. (2005)

1.9 Downstream analyses: Orthology and paralogy prediction.

In 1970, Walter Fitch coined the concepts of orthology and paralogy to distinguish two types of homology relationships among biological sequences (Fitch, 1970). Orthologous sequences are those derived from a speciation event while paralogous sequences are those that can be traced back to a duplication event. These two well defined concepts are sometimes misinterpreted by researchers because of the existence of complex scenarios of multiple duplications, speciations and gene losses or because of inaccurate methods. The accurate prediction of orthologs is important for many posterior analyses such as i) the inference of species trees based on concatenated alignment of undisputed orthologs (see above) or ii) the functional annotation of newly sequenced genes. In these two analyses, it is very important the correct identification of one-to-one orthologs since it is generally accepted than orthologs sequences conserved along evolution tend to have a great degree of conservation, making easier their recognition and posterior alignment, and, moreover, these genes tend to keep the same function making them the perfect candidates to annotate newly sequenced genes/genomes until further validations confirm such predictions.

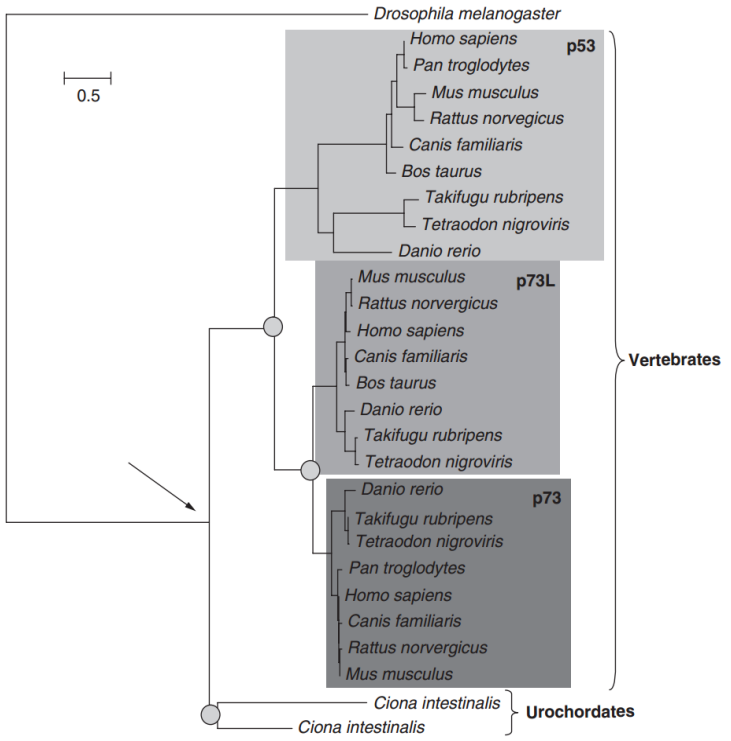

Figure 1.5: It shows how orthologs and paralogs, for gene p53, relate to each other by different evolutionary events. As it can be appreciated in the figure, making precise predictions can be a very difficult task highly dependant on the methods used to reconstruct the tree. In the picture, two consecutive ancestral duplications lead to the emergence of three gene families in Vertebrate that are orthologs, all of them, to the two copies in urochordates. Image obtained from Gabaldón (2008b)

1.10 Final remarks.

Phylogenomics offers us, for the first time, the opportunity to face, understand, and, in many cases, answer long-standing questions regarding the evolution of organisms. This opportunity implies many challenges about the suitability of current methods or/and the need of developing new ones in order to be able to use, learn and extract the most from an ever-growing amount of sequenced data.

2 Objectives

- Optimize an automated phylogenomics pipeline for the reconstruction of phylomes under the premises of high accuracy and speed.
- Implement several improvements on a public database to store and browse precomputed phylogenetic trees, alignments and homology relationships.
- Propose and implement different strategies to post-process multiple sequence alignments in the context of large-scale phylogenetics.
- Disentangle the phylogenetic signal carried by gaps in a multiple sequence alignment from other sources of signal such as guide trees.
- Use several phylogenomics approximations to answer the long-standing question of the phylogenetic position of Microsporidia.
- Propose and test a new methodology to identify optimal sets of phylogenetic marker genes that uses information from existing genome sequences.

3 Thesis overview

The present thesis addresses the dual objective of how to improve phylogenomic methods and how to use them efficiently to shed light onto diverse biological questions. The work is divided into different chapters, which I briefly introduce here.

In **Chapter 4**, an automated pipeline to reconstruct large collection of phylogenetic trees (i.e phylomes) is introduced. This pipeline had been initially developed in 2006, but from 2008 I was in charge of investigating ways for further optimizing it, particularly in the steps of alignment reconstruction, trimming, and evolutionary model selection. As a result, an improved pipeline was developed, which is used in current phylome reconstructions and constitutes one of the tools that I used in the following chapter. This chapter also describes a new release of phylomeDB, a public repository of phylogenetic trees, alignments and orthology and paralogy predictions. My main contribution to this team effort particularly consisted of the design and implementation of the database. Furthermore, I have been involved in the development of an Application Programming Interface (API) that enables programmatic access to phylomeDB, and in the incorporation of several quality controls to ensure the data integrity.

Chapter 5 focuses on multiple sequence alignment reconstruction, post-processing, and its impact in phylogenetic downstream analyses. On the one hand, I investigate the use of different alignment trimming strategies (gaps, similarity, consistency), specially focusing on finding heuristics that enable automated parameter selection in the context of large-scale phylogenetic pipelines. This work lead to the development of trimAl, an alignment trimming software that is central to the pipeline described in Chapter 2 and widely used by other groups (over 80 citations). On the other hand, I focus on the significance of gaps in the context of alignments used for phylogenetic reconstruction. For this, I develop a novel framework to disentangle the genuine phylogenetic signal carried by gaps from other sources of information such as that brought in by the guide tree.

In **Chapter 6**, I apply a battery of phylogenomic methods, including those developed in this thesis, to address a long-standing question in phylogeny: the position of microsporidia. Resolving the position of microsporidia has been described as one of the hardest tests for phylogeny, due to the

incredibly high rates of sequence evolution and the strong effect of biases such as long branch attraction. Using a taxon sampling of more than 100 completely sequenced fungi, and methods ranging from synteny analysis to selection of informative sites for statistical testing of tree topology, I could resolve the position of this elusive taxon as the most basal group within fungi.

Chapter 7 also deals with the use of complete genomes to resolve species phylogenies, but as a general case. In particular I propose a new methodology to prioritize sets of phylogenetically informative gene markers to resolve species trees at desired levels (e.g. genus or phylum). Contrary to existing approaches, our method explores the phylogenetic informativeness of genes using cross-validation, to ensure that markers are stable outside the set of species used for their selection. To demonstrate the validity of the method, we applied it to the reconstruction of species trees for a bacterial (Cyanobacteria) and a eukaryotic (Fungi) phylum. The Cyanobacterial marker set is currently being used in a collaborative project to elucidate the Cyanobacterial tree of life.

Chapter 8, I present a final discussion, summarizing the main implications of my research to current debates, and considering possible future directions in these topics.

Finally, the **Appendix** section compiles a series of studies in which I have contributed by applying the methods described here. These collaborations have not only provided me access to new and interesting data, but have also served to inspire new developments tailored to specific challenges. These studies include collaboration in the phylogenomic analyses of two newly sequenced genomes (the Melon and the red algae Chondrus), the study of the phylogenetic origin of peroxisomes, and the implementation of trimAl within the Phylemon phylogenetic webserver.

4 Reconstructing genome-wide collections of phylogenetic trees

4.1 Improving an automated phylogenomics pipeline

Salvador Capella-Gutiérrez & Toni Gabaldón

Improving an automated phylogenetic pipeline.

4.1.1 Abstract

The use of phylogenetic trees to describe the evolution of biological molecules was established in the 1960s and remains a fundamental approach to understand the evolution of genes and species. Performing phylogenetic studies at genomic levels, phylogenomics, offers nowadays a unique opportunity to address a huge range of biological questions. However, large-scale studies present many conceptual and methodological challenges. For instance, the automation of the whole process of tree reconstruction often involves the use of standard parameters and conditions for all protein families, inevitably resulting in poor or incorrect phylogenies in many cases. Here we present the different improvements done over an existent phylogenetic pipeline. To achieve this objective, we have centered our efforts in two steps of the pipeline: 1) the Multiple Sequence Alignment generation and 2) the phylogenetic tree reconstruction. In the Multiple Sequence Alignment phase, we have developed a novel program, trimAl, to remove those ambiguous regions from the alignment in an automated way. On this step, we have also discarded any possible bias to a certain method using several programs and sequences orientations (forward and reverse orientation) in order to keep only those columns that are less sensitive to the alignment process. In the phylogenetic tree reconstruction phase, we have been concentrated on the evolutionary model selection because it is the main bottleneck of our pipeline. In this point, we have worked to know whether an evolutionary model selection over Neighbour-Joining trees predicts the same model that the one predicted by the Maximum Likelihood trees. Addressing this question implies a substantial time-consumption optimization over the current pipeline.

4.1.2 Introduction

The use of phylogenetic trees to describe the evolution of biological molecules was established in the 1960s and remains a fundamental approach to understand the evolution of genes and species. Performing phylogenetic studies at genomic levels, *phylogenomics*, offers nowadays a unique opportu-

nity to address a huge range of biological questions (Eisen and Fraser, 2003). However, large-scale studies present many conceptual and methodological challenges. For instance, the automation of the whole process of tree reconstruction often involves the use of standard parameters and conditions for all protein families, inevitably resulting in poor or incorrect phylogenies in many cases. Moreover, interpreting such type of complex data poses many difficulties and does require the development of novel algorithms, tools, forms of representing the data and even new semantics and concepts (Gabaldón et al., 2008).

Many phylogenomic studies involve the reconstruction of large sets of phylogenetic trees, for which automated pipelines should be implemented. Even though there are several phylogenetic routines, most of them share three well defined stages: homology search, multiple sequence alignment generation and phylogenetic tree reconstruction. Homology search is the initial step and consists of a search for putative homologous sequences, inferred from their level of sequence similarity. This search can be performed by using local-alignment algorithms such as Smith-Waterman or BLAST (Altschul et al., 1990) to search in public or local databases. Sets of homologous sequences are subsequently aligned. Multiple sequence alignments are a central part of all phylogenomic pipelines, since the reliability and accuracy of subsequent analyses critically depend on their quality. Once the sequences are aligned, a phylogenetic tree can be reconstructed from the positional homology information contained in the alignment. There are three major approaches for phylogenetic estimation, namely distance methods, parsimony and statistical approaches such as Maximum Likelihood (ML) and Bayesian inference (BI) (Baldauf, 2003). In this context, our group has developed a pipeline to reconstruct phylomes, (i.e. the complete collection of phylogenies of encoded genes in a given genome). One of the first phylomes reconstructed with such pipeline was the human phylome in the context of 39 eukaryotic species (Huerta-Cepas et al., 2007). Since then, a similar pipeline has been applied to the generation of more than 25 phylomes, including those of yeast, pea aphid, *E. coli* and others that are deposited in PhylomeDB (www.phylomedb.org) (Huerta-Cepas et al., 2008).

In brief, this pipeline proceeds as follows: for every seed protein, homologs are searched against a database encompassing proteomes from the desired taxonomic scope. Subsequently, search hits are filtered based on criteria such as e-value, alignment coverage between the query and the hit proteins, and an upper limit in the number of sequences. Afterwards, significantly similar groups of proteins are aligned using the MUSCLE program (Edgar, 2004). Alignments are automatically trimmed to remove gap-rich regions and, finally, a phylogenetic reconstruction phase combines neighbour-joining (NJ) and maximum likelihood (ML) tree reconstruction approaches. Firstly, an NJ tree is reconstructed using BioNJ (Gascuel, 1997). Secondly, the precomputed NJ tree is used as a seed to search up to 4 different ML trees, based on different evolutionary models. Once the ML trees have been generated, the Akaike Information Criterion (AIC) (Akaike, 1974) is used to select the evolutionary model that best fits the data.

The main difference between our pipeline and similar routines such as PhyOP (Goodstadt and Ponting, 2006) or Emsembl Compara GeneTrees (Vilella et al., 2009), lies on the fact that ours reconstructs a phylogeny for each protein of a given genome, while others rely on an initial clustering step and then a single phylogeny is built for every cluster. Thus, our pipeline resembles more closely what a phylogeneticist will do when interested in the evolution of a given protein (i.e. genome-wide clustering, which is parameter-dependant and prone to other type of errors is not used in classical phylogenetic analyses). Other differences regard alternative taxonomic scopes or methodological details.

In the described implementation of the pipeline we relied on a single MSA program, namely MUSCLE. The choice of this program was based on its speed and the fact that it includes a final iterative phase that allows solving obvious mistakes generated during the progressive alignment phase. However, as it has been shown in previous studies (Golubchik et al., 2007; Landan and Graur, 2007), a particular method can be biased due to heuristic decisions taken during the program execution.

An important limiting factor in our pipeline is the phylogenetic tree reconstruction phase. More specifically, the selection of the best model in Maximum Likelihood (ML) analyses uses a considerable amount of

time. As explained, in our previous implementation, different ML trees are reconstructed using different evolutionary models. Subsequently the best fitting model is chosen based on a likelihood ranking. The high time-consumption of the ML computation limits the number of models that can be tested. In our previous pipeline this limit was set to 4 different evolutionary models.

In order to increase the accuracy and efficiency of our pipeline, we undertook an analysis of the current settings of our pipeline in order to identify possible areas of improvement. Here we present the results of such analysis, which led to the redesign of different parts of the pipeline. In particular we have focused on three important aspects of our pipeline: 1) alignment trimming, 2) alignment reconstruction, and 3) evolutionary model selection. To decide among alternative designs we have used different benchmarks trying to find an optimal balance between accuracy and speed. Based on these results, we propose a novel implementation of the pipeline which, as compared to our previous design, achieves higher accuracies, while reducing the overall computing time. Furthermore, this work has led us to the development of new software: trimAl (Capella-Gutiérrez et al., 2009), a tool for automated alignment trimming.

4.1.3 Methods

Multiple Sequence Alignment.

To reverse biological sequences and convert between different alignment formats readAl 1.2, included in the trimAl package (trimal.cgenomics.org), was used. Forward and Reverse sequence sets were aligned with different MSA programs. In particular we tested MUSCLE v3.7 (Edgar, 2004), MAFFT v6.712b (Katoh and Toh, 2008), KALIGN v2.03 (Lassmann et al., 2009), DIALIGN-TX (Subramanian et al., 2008), PROBCONS (Do et al., 2005) and T-COFFEE v8.06 (Notredame et al., 2000).

All programs were used with default parameters except for MAFFT, in which we set the *-auto* parameter to allow the program select the best algorithm depending on the input sequences features.

Multiple Sequence Meta-Alignment.

Selected combinations of MSA (always including forward and reverse replicates) were integrated to produce a consensus alignment with M-COFFEE (Wallace et al., 2006), implemented in the T-Coffee package.

Multiple Sequence Alignment trimming.

To remove ambiguous columns in the consensus alignment we used trimAl. trimAl is a tool for automated alignment trimming, which is especially suited for large-scale phylogenetic analyses. It was developed within the frame of this project. trimAl now incorporates many options, most of them are discussed in our publication (Capella-Gutiérrez et al., 2009).

In this project, we have implemented in trimAl a method to compute the column consistency score (S_c) for each column from the consensus alignment generated by M-COFFEE. Equivalent to the *sum-of-pairs score* (SPS) (Thompson et al., 1999), *the consistency* score measures the proportion of residue pairs that are paired identically in two alignments. In this case, we compared the consensus alignment with the rest of the alignments used to generate it. Then, those columns that do not achieve a certain consistency score are removed from the final alignment by trimAl.

Phylogenetic tree reconstruction.

Neighbour-Joining and Maximum Likelihood trees were reconstructed using phyml v3.0 (Guindon et al., 2010). Seven different evolutionary models were evaluated: JTT, LG, WAG, Blosum62, VT, MtREV and Dayhoff. In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was used, estimating the gamma parameters and the fraction of invariants sites from the data. In the case of the NJ computation, the branch length was optimized under the different evolutionary models to get likelihood values. The ranking of the likelihood values as well as the comparisons were generated using ad-hoc python scripts.

Phylogenetic tree accuracy benchmark I.

A benchmark composed by three different datasets has been used to test the general applicability of trimAl.

The first dataset has been used previously (Talavera and Castresana, 2007) to test the improvement in phylogenetic performance after an alignment trimming phase. This set comprises simulated sequences of various lengths (400 to 3200 positions). Simulations were performed with ROSE v1.3 (Stoye et al., 1998) along a phylogenetic tree with 16 tips. These trees have three different topologies varying in their level of symmetry, and whose branch lengths were multiplied by 0.5, 1.0 and 2.0.

Two additional sets of sequences were generated to expand the original dataset to the case of 32 and 64 tree tips, which we consider to be more realistic in phylogenomic analyses. In order to generate these additional sets we first took the reference trees, one tree per topology, and then twelve new reference trees were generated by using ETE (Huerta-Cepas et al., 2010). These new trees had the same level of symmetry as those in (Talavera and Castresana, 2007) study. Six of these twelve trees had 32 tips while the six left had 64 tips. The lengths of their branches were also multiplied to obtain the same three levels of divergence (0.5, 1.0 and 2.0, respectively) as in the previous study.

These reference trees were used to generate the sets of sequences as indicated in Talavera and Castresana (2007). For this purpose, the program ROSE was used with the same seed protein and parameters described in Talavera and Castresana (2007) to generate their benchmark sets. The simulations included insertions and deletions with a probability of 0.03. The other parameters for the simulation were the ones described in the original study. The same strategy was used to infer the patterns of rate heterogeneity of the seed protein. Finally, the sets generated by ROSE contain, similarly to the set of 16 sequences, simulated protein sequences of various lengths (400 to 3200 residues) and different topologies.

Phylogenetic tree accuracy benchmark II.

A partial dataset, composed by the hardest cases from the above benchmark has been used to measure the average time-consumption for each examined MSA program, as well as to quantify the general improvement achieved for the different combinations of programs. This dataset is composed by 300 sets of 64 sequences each generated along an asymmetric tree shape. It is

divided in three groups, 100 sets each, which vary on the level of sequence divergence.

A human phylome based on 12 model organisms benchmark.

A new human phylome (Huerta-Cepas et al., 2007) has been reconstructed in the context of 12 model organisms: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Escherichia coli*, *Gallus gallus*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Homo sapiens*. Homology searches were performed using an e-value cutoff of 1e-3, a minimum coverage of 50% over the query sequence and an upper limit of 150 sequences. Groups of homologous sequences were aligned using MUSCLE and trimmed with trimAl gappyout method. Finally, 7 different evolutionary models were tested using the Neighbour-Joining and Maximum Likelihood approaches as implemented in PhyML, in all cases rate variation across sites was approximated with a four rates gamma categories distribution, the gamma shape parameter and the proportion of invariable sites were estimated from the data.

20,624 different proteins were used and approximately 290,000 phylogenetic trees were reconstructed.

4.1.4 Results.

Multiple Sequence Alignment trimming.

Multiple sequence alignments (MSA) are central to many areas of bioinformatics, including phylogenetics, homology modeling, database searches and motif finding. Recently, such MSA-based techniques have been incorporated in high-throughput pipelines such as genome annotation and phylogenomics analyses. Accuracies of 80-90% have been reported for the best algorithms, but even the best scoring alignment algorithms may fail with certain protein families or at specific regions in the alignment. The situation worsens in large-scale analyses, where faster but less reliable algorithms and large numbers of automatically selected sequences are used. It is therefore generally assumed that trimming the alignment, so that poorly aligned regions are eliminated, increases the accuracy of the resulting MSA-based applications

(Talavera and Castresana, 2007).

Some programs such as GBlocks (Castresana, 2000) have been developed to assist in the MSA trimming phase by selecting blocks of conserved regions. They have become very popular and are extensively used, with good performance, in small-to-medium scale datasets, where several parameters can be tested manually (Talavera and Castresana, 2007). However, their use over larger datasets is hampered by the need of defining, prior to the analysis, the set of parameters that will be used for all sequences families. In particular, our own experience in phylomeDB was that any settings of GBlocks will produce a significant number of alignments in which the number of columns removed was either excessive or too reduced producing largely unexpected tree topologies. In the previous implementation of our pipeline a python script was being used in which the threshold of gaps allowed in a column could be set. This limited our possibilities to implement alternative trimming strategies. Thus, driven by our own needs we developed trimAl, a tool for automated alignment trimming. The speed of trimAl, and the possibility for automatically adjusting the parameters to improve the phylogenetic signal-to-noise ratio for a given alignment, makes trimAl especially suited for large-scale phylogenomic analyses, involving thousands of large alignments.

trimAl reads and renders protein or nucleotide alignments in several standard formats. trimAl starts by reading all columns in an alignment and computes a score (S_x) for each of them. This score can be either a gap score (S_g), a similarity score (S_s) or/and a consistency score (S_c). The gap score (S_g) for a column is the fraction of sequences without a gap in that position. The residue similarity score (S_s) consists of mean distance (MD) scores as described in Thompson et al. (1999). This score uses the MD between pairs of residues, as defined by a given scoring matrix. Finally, the consistency score (S_c), can only be computed when more than one alignment for the same set of sequences is provided. Details on how these scores are computed are provided in the Supplementary Material of our publication available at trimal.cgenomics.org.

Alternatively, if the automatic selection of parameters options is selected, trimAl will compute specific score thresholds depending on the inherent

characteristics of each alignment. So far, trimAl incorporates three modes for the automated selection of parameters, namely, gappyout, strict and strictplus, which are based on the different use of gap and similarity scores. Moreover, the option *-automated1* implements a heuristic to decide the most appropriate mode depending on the alignment characteristics. The heuristics to define such parameters have been designed based on the results of a benchmark. Details on the heuristics and the benchmark can be found in the on-line documentation of the program. In brief, the automatic selection of parameters approximate optimal cut-offs by plotting, internally, the cumulative graphs of gap and similarity scores of the columns in the alignment (see trimal.cgenomics.org).

The whole dataset from the phylogenetic accuracy benchmark I (see methods) has been used to measure the improvement achieved using trimAl over other alternatives. This dataset simulates several evolutionary scenarios varying in the number and length of the sequences, the topology of the underlying tree and the level of sequence divergence considered. We compared the results obtained from MUSCLE alignments before and after trimming with trimAl using automated selection of parameters. The accuracy of the resulting trees was measured by comparing them with the original trees used to generate the sequence sets, and measuring the Robinson and Foulds distance (Robinson and Foulds, 1981). We observed an overall improvement of the phylogenetic accuracy after trimming. Using *-automated1* option of trimAl, the trimmed alignment always produced Maximum Likelihood trees that were of equal (36%) or significantly better (64%) quality as compared with the tree derived from the complete alignment. For Neighbor Joining reconstruction the *-strictplus* option of trimAl worked best, improving the phylogenetic accuracy in 89% of the scenarios. In most scenarios (90%), trimAl outperformed Gblocks v0.91b with default parameters. Most importantly, the use of Gblocks default parameters diminished the accuracy of the subsequent tree reconstruction in half of the scenarios considered. In contrast, the use of trimAl automated methods rarely (1.5%) undermined the topological accuracy of the resulting phylogenetic tree (see Supplementary Material for more details in trimal. cgenomics.org). An example figure of a particular benchmark is presented

in Figure 4.1.

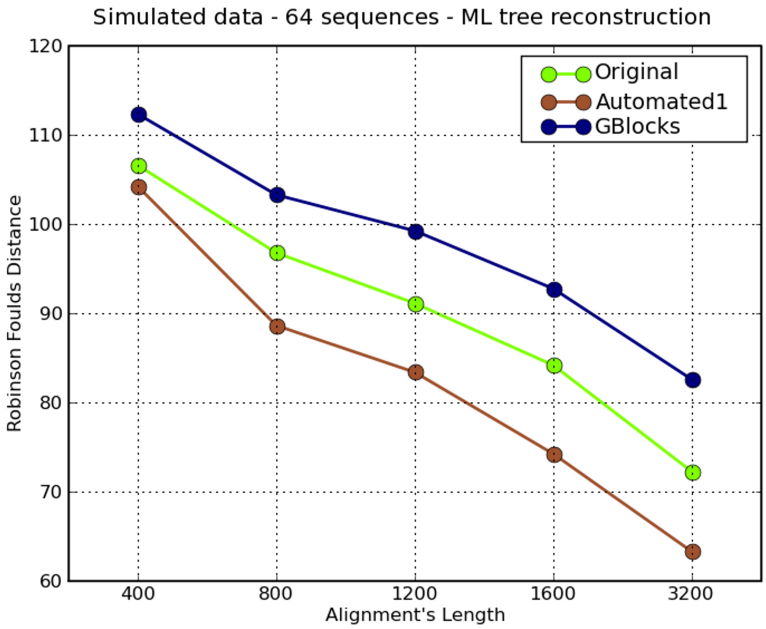

Figure 4.1: Benchmark results for phylogenetic tree accuracy. Y-axis represents the Robinson and Foulds (RF) distance to the real tree for the different alternatives (lines) having into account the effect of the sequences length. The lower RF values, the more accurate. 100 sets of sequences were considered for each possible sequence length.

Multiple Sequence Alignment.

As mentioned above, Multiple Sequence Alignment (MSA) constitutes the basis of our pipeline and subsequent analyses depend on their accuracy. A plethora of computer programs and algorithms for MSA are currently available (Notredame, 2007), which implement different heuristics to find mathematically optimal solutions to the MSA problem. In the context of large-scale studies, we have relied on a single MSA program (MUSCLE) chosen for its speed and the final MSA iterative refinement. However, since different studies (Golubchik et al., 2007; Landan and Graur, 2007) have shown that a given program can be biased by heuristic decisions taken during the program execution, we wanted to test the possibility of combining different methods rather than relying on a single one. Moreover, the HoT method has been shown to effectively detect those columns that are

more variable to heuristic decisions.

We wanted to evaluate an alternative design that could use the information on the variability of 1) the specific alignment method used and 2) the specific orientation of the input sequences. Since time-consumption imposes a strong limitation to our pipeline, we decided to first benchmark the speed of the different methods Table 4.1. The phylogenetic accuracy benchmark II (see methods) has been used to carry out that time evaluation.

Programs	divergence x 0.5	divergence x 1.0	divergence x 2.0
KALIGN2	00.04 min	00.05 min	00.06 min
MUSCLE	00.50 min	00.59 min	00.68 min
MAFFT	02.66 min	03.36 min	07.54 min
DIALIGN-TX	06.62 min	06.59 min	07.01 min
T-COFFEE	28.51 min	31.08 min	34.65 min
PROBCONS	64.33 min	57.63 min	63.63 min
M-COFFEE	01.27 min	00.77 min	00.74 min

Table 4.1: Benchmark results of time-consumption for the different programs. These average execution times have been computed over 100 alignments. As expected, the more divergence between the sequences, the more time is needed to generate the MSA.

Considering these results, we selected two alternative combinations of methods that tried to 1) diversify the alignment approaches and 2) maximize the total speed. The 2 chosen combinations are *KMM* combination, comprising by KALIGN2, MUSCLE and MAFFT programs, and *DMM* combination, composed by DIALIGN-TX, MAFFT and MUSCLE programs. The selected methods were used to align the forward and reverse sequences and the resulting 6 MSAs were combined with M-COFFEE. The combined MSA was then trimmed based on the level of consistency, using a threshold of 0.1667, and the level of gaps, with a threshold of 0.1.

The improvement in the accuracy of both combinations was benchmarked and compared with previous results from the phylogenetic accuracy benchmark II. The sequences were aligned using the above described methods as well as using either just MUSCLE, or MUSCLE with an automated trimming phase (trimAl *-gappyout*). The maximum likelihood trees for all MSAs were

reconstructed using PhyML. The topological differences, namely the Robinson and Foulds distance, were calculated for each reconstructed tree against the tree used during the simulations. A summary can be found in Figure 4.2.

These results show a clear improvement in accuracy over the previous implementation (MUSCLE + trimAl *-gappyout*). This is specially true for the most divergent dataset. Considering these results, we decided to implement the DMM implementation, despite a somewhat higher time-consumption, since it produced a desirable increase in accuracy.

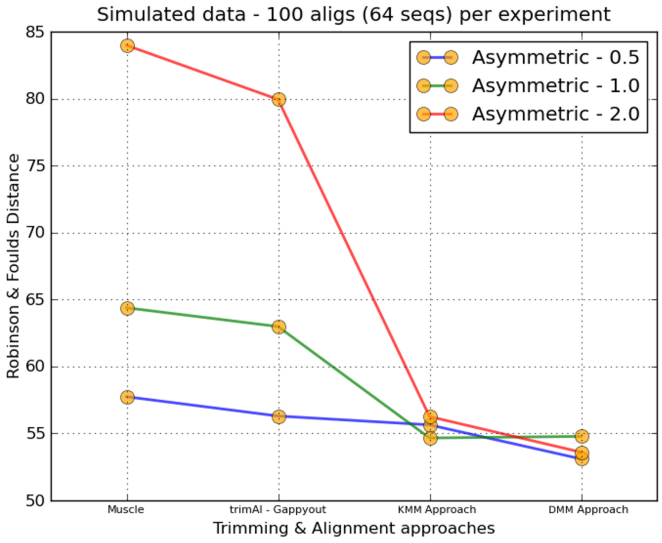

Figure 4.2: Benchmark results for phylogenetic accuracy. Y-axis represents the Robinson and Foulds (RF) distance to the real tree for the different alternatives (x-axis). The lower RF values, the more accurate.

Evolutionary Model Selection.

An important parameter for the Maximum Likelihood (ML) approach is the evolutionary model used to produce the tree. It has been shown that model misspecification can lead to wrong topologies with high support (Bruno and Halpern, 1999). In our previous implementation of the pipeline different ML trees were reconstructed with different evolutionary model, to subsequently select the best tree based on the Akaike Information Criterion (AIC). Since the ML approach requires large amounts of time, the model selection step

should be applied to a limited number of models.

In our current implementation, up to 4 different evolutionary models are tested, and this step constituted the most important bottleneck in our pipeline. It would be desirable to test more than 4 evolutionary models to avoid selecting a sub-optimal model. Considering that, we decided to investigate whether a tree topology obtained by the NJ method would serve to predict the best-fitting model for an ML analyses. This reasoning is based in that NJ topologies are largely similar to those obtained by ML. To investigate these points, we have used a new version of the Human Phylome (see methods). 7 evolutionary models: JTT, LG, WAG, Blosum62, MtREV, Dayhoff and VT, were used during this test in the Neighbour-Joining (NJ) and Maximum Likelihood (ML) tree reconstruction. Once all the trees were generated, we computed how many times the best evolutionary model for the ML trees was among the *i* best models for NJ (Figure 4.3).

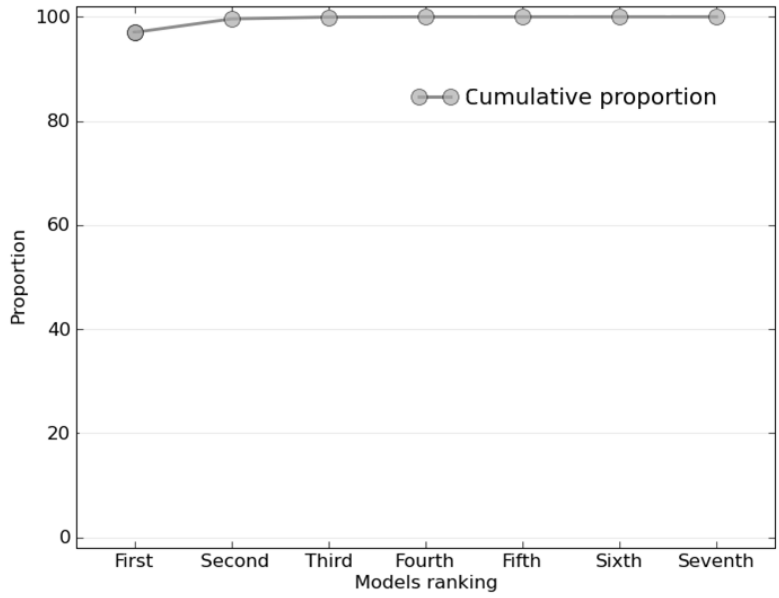

Figure 4.3: Benchmark results for the evolutionary model selection benchmark. Y-axis represents the proportion of how many times the best evolutionary model for the ML tree is in the best *ith* model for the NJ trees.

These results clearly show that model-assessment on the NJ topology could be used to accurately predict the best model in an ML reconstruction.

Since this approach is significantly faster than the original one, we decided to implement it in the new implementation of the pipeline. With that implementation, we are able to test more evolutionary models without increasing the pipeline time-consumption. In fact, we are speeding-up the whole process because we would reconstruct one or two ML trees instead of the four ML trees in the previous pipeline.

4.1.5 Discussion.

Since phylogenomics studies offer a unique opportunity to address a huge range of biological questions, it is important to have tools to give an as accurate as possible answer to those questions. In the curse of this project, we have been focused on different aspects of our previous pipeline in order to improve its accuracy as well as to increase its speed.

In the multiple sequence alignment phase, we have developed a novel program, trimAl, to maximize the signal-to-noise ratio removing the ambiguous regions in the alignment. Since its publication, we have continued its development based on the feedback of trimAl's users. The improvements include the incorporation of new methods to improve the alignment, not only removing columns but also removing sequences. In this stage of the pipeline, the use of different alignment programs and orientations in the sequences has allowed us to discard possible biases towards a specific approximation, increasing at the same time the accuracy of the generated alignment. Thanks to Dr. Cedric Notredame's feedback, we have decided to use instead of a combination of 3 different programs, the 4 fastest programs used during our benchmarks (MUSCLE, MAFFT, KALIGN2, DIALIGN-TX). The use of these 4 programs does not represent a significant increase of the time-consumption while the number of samples is increased in order to identify those columns less sensitive to the way that we align our homologous sequences.

All the improvements achieved in the multiple sequence alignment phase have been tested over an extensive but simulated dataset. For this reason, we are interested in testing our methods either with real or new simulated data. The underlying problem is how to obtain a dataset to benchmark our

methods now that a recent study (Edgar, 2010) has shown the ambiguities in one of the most popular databases, BALiBASE (Thompson et al., 2005), for real data. Additionally, simulated data is not able to capture the complexity of the real world and can therefore not be fully trusted. A new study (Dessimoz and Gil, 2010) tries to offer a new approach to evaluate the accuracy of the different alternatives based on how many times the correct tree topology is reconstructed from the alignments generated by the different alternatives. While their conclusions can only be based on limited data, we incorporated this kind of test in our benchmarks to evaluate the accuracy of the different alternatives as well as to measure the improvement achieved when the ambiguous regions are deleted using trimAl.

Since the evolutionary model selection step in the phylogenetic tree reconstruction phase constitutes the main bottleneck of our pipeline, we are interested in optimizing this process. After addressing that we are able to perform the model selection over Neighbour-Joining trees instead of Maximum Likelihood trees, we would like to find the automated way to decide how many evolutionary models should be considered during the Maximum Likelihood reconstruction. Actually, this automated way would avoid us to take arbitrary decisions about the number of evolutionary models to be considered, even when we have clear results that support these arbitrary decisions. To test this point, we would like to evaluate the performance of CONSEL (Shimodaira and Hasegawa, 2001), which incorporates different statistical tests that allow us to decide how many evolutionary models should be considered. One important point to investigate here is which test will be used to take these automated decisions. On the other hand, we are following the latest QUDA developments in the phylogenetic field in order to incorporate this technology as soon as possible.

Acknowledgements.

The author (SCG) is grateful to Marta Catala Ivañez de Lara for carefully reading and revising the manuscript. We also acknowledge to members of the Gabaldón group for fruitful discussions, suggestions and paper revision

4.2 PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions

Jaime Huerta-Cepas, Salvador Capella-Gutiérrez, Leszek P. Pryszcz,
Ivan Denisov, Diego Kormes, Marina Marcet-Houben
& Toni Gabaldón

Originally published in: *Nucleic Acids Research*. January, 2011.

Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, et al. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res. 2011 Jan;39: D556-60.

5 Multiple sequence alignment trimming

5.1 trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses

Salvador Capella-Gutiérrez, José M. Silla-Martínez & Toni Gabaldón

Originally published in: *Bioinformatics*. August, 2009.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1;25(15):1972-1973.

5.2 trimAl 1.4: Recent developments in automated multiple sequence alignment post-processing in large-scale phylogenetic analyses.

Salvador Capella-Gutiérrez & Toni Gabaldón

In preparation

trimAl 1.4: Recent developments in automated multiple sequence alignment post-processing in large-scale phylogenetic analyses.

5.2.1 Introduction.

Since its first implementation and publication in 2009 (Capella-Gutiérrez et al., 2009), trimAl has gained broad recognition and is now extensively used by several labs around the world. trimAl publication has so far been cited in over 70 publications, and we constantly receive requests from users and suggestions on further developments. As a result, trimAl has evolved to incorporate new methods to accommodate our own needs and those from other users. Here I briefly list the main novel implementations that are part of the current version of trimAl (v1.4).

5.2.2 New implementations.

Newly supported formats.

New multiple sequence alignments formats have been included, leading to a total of 12 different supported formats. Special attention has been put in developing specific formats for popular programs such as PAML (Yang, 2007) or those part of the popular phylogenetic package PHYLIP.

Alignment editing functionality.

Using the trimAl package, specifically the readAl program, it is now possible to edit input multiple sequence alignments in order to get the reverse alignment, that means, the first column is now the last one, the second one is the second last one and so on. This method enables using the Head or Tails (HoT) approach for assessing alignment variability and consistency (Landan and Graur, 2007). Moreover, it is possible to reshuffle randomly the input sequences as an alternative way to assess alignment variability and consistency. Additionally, it is now straightforward to get the unaligned sequences from the input alignment as well as an HTML file colouring the input alignment with standard colour schemes.

Back-translating protein MSA into its corresponding codons MSA.

In the current version, it is possible to back-translate an input protein sequence alignment into its corresponding nucleotide codons, just providing to trimAl the unaligned nucleotide sequences. Moreover, the program can apply any trimming strategy to the input alignment and then give back only those codons corresponding to the untrimmed columns in the protein alignment. trimAl checks input coding sequences to look for universal stop codons, and there is an option to split sequences just by stop codons, in order to check for correspondence, in terms of length, between proteins and coding DNA sequences. Furthermore, trimAl can complete missing residues using N symbols. These procedures have been implemented to accommodate the needs from studies relying on incomplete genome assemblies, and transcriptomic data such as those from RNAseq or Expressed Sequence Tag (EST) experiments.

Improved HTML summary output.

The HTML-based summary output has been redesigned and improved. On this new implementation, columns are colored according to its nature: DNA, RNA or proteins, following the colour scheme used by Clustal, Jalview and the PFAM-Squared Server (see web). Additionally, there is a bar to indicate which columns are kept, dark grey, or removed, light grey. Moreover, there as many bars as scores used to trim the alignment with the different scores for each columns. In these cases, bars contains a gradient of 12 different colours covering scores from 0, the lightest colour, to 1, the darkest one.

Manual selection of sequences to be removed.

The functionality for removing user-defined columns has been complemented with the possibility of removing specific sequences from the input alignment.

Refining existing trimming methods: asking for a minimum block size.

In order to avoid using single (or very small blocks of) columns, it is possible to ask the program to keep only those blocks of a minimum size predefined by the user. This option is very useful when input alignment is large enough

so that independently of the used method, large continuous portions of the alignment are still recognizable.

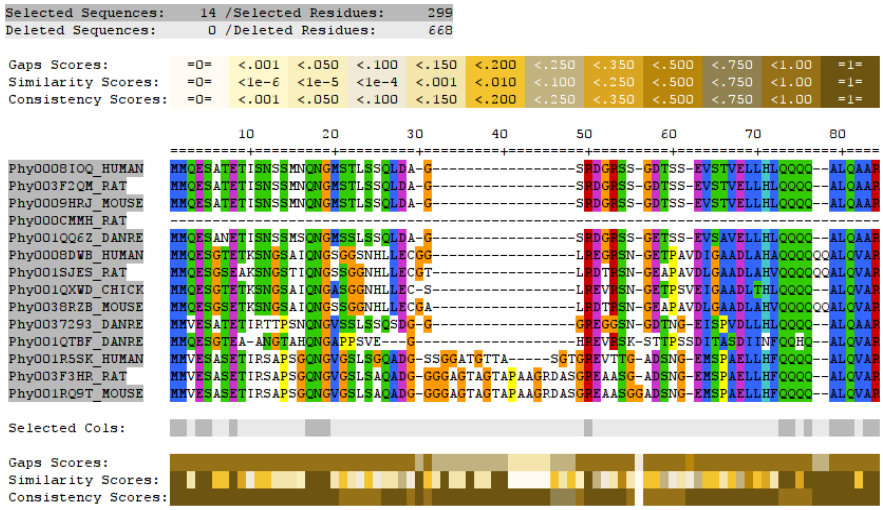

Figure 5.1: An example of an HTML-based summary output from trimAl. Different filters applied by trimAl to remove columns in the input alignment are shown. Dark grey colour indicates which columns have been kept in the final alignment and light grey which ones have been removed. Additional bars show the score values used to trim the alignment.

Refining existing trimming methods: keeping the alignment core.

The most variable parts in multiple sequence alignments are often at the beginning and at the end. In order to get rid of these regions, it is possible to apply any trimming method only to these parts after identifying the boundaries between the core alignment and the outside regions. Boundaries are defined as the first and the last column composed only by any kind of residue, figure 2 shows an example of such definition, highlighting these boundaries with red lines.

Figure 5.2: Example about the identification of alignment boundaries to differentiate the core alignment from outside regions. Once boundaries are defined, any trimmed method can be applied to outter columns.

Refining existing trimming methods: Selecting the alignment to be trimmed using other alignments as reference.

One of the most useful methods to detect the potential misaligned columns in an alignment is based on the comparison of residue consistency across several alignments. In the published version, given a set of alignment, the program selects the one which is more consistent among all of them and proceeds to trim it according to the selected method/s. In the new version, it is possible also to select which alignment is used to compute the consistency scores for each column. Using this approximation, it is possible to use the same set-up in a context with several sets of sequences, where consistency across set of programs/configurations can experience a great variation leading to select, for individual cases, alignments produced by different strategies.

New trimming methods: Removing automatically sequence redundancy in the input alignment.

In a context where thousands of sequences are generated everyday, it is possible to face the situation of alignments with high sequence redundancy. In order to accurately identify and remove such redundancy, two new methods have been implemented. In the first one, sequences with identity levels lower than a threshold set by the user are kept for the final alignment. That implies sequences are collapsed into clusters according to their identity values. Subsequently, the longest sequence, in terms of number of residues, of each group is kept as a representative for the cluster. In the second case, the user selects the number of final sequences for the alignment. trimAl iteratively finds the optimal identity cut-off to get such number of clusters.

Extending benchmarks: Using real data.

In order to measure the performance of the different trimming methods implemented in trimAl, three new datasets containing real cases have been added to the sets of benchmarks. One of the dataset comprises the original data from Dessimoz and Gil (2010), which was used to

show phylogenetic information in gaps, was accessed from their website (cbrg.ethz.ch/research/msa/). This dataset contains groups of orthologous proteins for three different taxonomic clades eukaryotes (609 orthologous groups), fungi (844), and bacteria (1,999). The second dataset corresponds to 1,502 single copy proteins detected in 7 yeast genomes by Wong et al. (2008). In addition, we downloaded the original data from (Marcet-Houben and Gabaldón, 2009), accessed through the public database phylomedb. org (Huerta-Cepas et al., 2011). This dataset (phylome ID = 7) contains trees for all *Saccharomyces cerevisiae* proteins across a phylogeny of 12 Saccharomycotina species. The data was filtered out to keep only 857 sets of 1-to-1 orthologous proteins.

Designing new benchmarks: Measuring accuracy of gap insertion.

Using simulated data (Capella-Gutiérrez et al., 2009) as well as real data coming from BAliBASE v3.0 (Thompson et al., 2005), where gaps placements are manually curated, we have measured how precise are gaps inserted by different programs and how precise the remaining ones are placed after applying different trimming strategies.

5.3 Are gaps phylogenetically informative?: disentangling the signal carried by alignment gaps and guide trees.

Salvador Capella-Gutiérrez & Toni Gabaldón

manuscript submitted

Are gaps phylogenetically informative?: disentangling the signal carried by alignment gaps and guide trees.

5.3.1 Abstract.

Motivation: Multiple sequence alignments are generally reconstructed using a progressive approach that follows a guidetree. During this process gaps are introduced at a cost to maximize residue pairing, but it is unclear whether they reflect actual past events of sequence insertions or deletions. It has been found that patterns of gaps in alignments can be used to reconstruct the true phylogeny, but it is as yet unknow whether gaps are simply reflecting information that was already present in the guide-tree.

Results: We here develop a framework to disentangle the phylogenetic signal carried by gaps from that which is already present in the guide-tree. Our results indicate that most gaps are incorrectly inserted in patterns that, nevertheless, follow the guidetree. Thus, most gap patterns in current alignments are not informative *per se*. This affects different programs to various degrees, being PRANK the most sensitive to the guide-tree.

5.3.2 Introduction.

Multiple sequence alignments (MSA) play a central role in modern molecular biology, and are used in a broad set of applications, ranging from phylogenetic analyses to the identification of functional motifs (Notredame, 2007). Since the quality of an alignment will inevitably affect the quality of downstream analyses, different strategies have been proposed to improve the quality of MSA. In the context of the reconstruction of phylogenetic trees to establish the evolutionary relationship among a given set of sequences, a major problem is the interpretation of gaps. Theoretically, gaps in an alignment may serve to represent past events of sequence insertions or deletions. In practice, however, they are generally introduced to maximize residue pairing scores. Most alignment reconstruction programs use a progressive approach in which most similar sequences are aligned first, following a guide-tree.

During the alignment reconstruction, optimization is based on two main components: residue pairing and gap penalties. In contrast to residue pairings, where empirical models exist, gap penalties are rather arbitrary. As a result highly gapped regions are generally considered unreliable (Golubchik et al., 2007), and it is common practice to ignore them prior to phylogenetic analyses (Talavera and Castresana, 2007; Capella-Gutiérrez et al., 2009).

A recent study has reported an unexpected accuracy of maximum parsimony trees reconstructed solely from the information contained in presence/absence patterns of gaps in protein alignments (Dessimoz and Gil, 2010). This result has been attributed to phylogenetic information contained directly in gaps introduced by alignment programs, and would imply that current phylogenetic methods could be improved by exploiting such information. However, for this to be true, gaps should carry independent phylogenetic information, truly reflecting past evolutionary events such as insertions and deletions. Alternatively, due to the progressive nature of the alignment reconstruction, gap patterns may simply reflect information already present in the guide-tree, which is usually reconstructed from pair-wise sequence distance information. If this would be the case, usage of the gap patterns in phylogenetic reconstruction would be biased towards the guide-tree, which is prone to contain errors. Disentangling the two scenarios is of central importance in order to design proper strategies to exploit the potential information contained in gaps. At the same time, this task is challenging, given the lack of a proper framework to measure the effect that guide-trees have in the introduction of gaps. Here we develop a novel approach to assess whether the information contained in gap patterns reflect true evolutionary events, and whether this is different from the phylogenetic signal already present in the guide-tree. We apply such framework to several synthetic and real datasets and using five different alignment strategies that represent the main alignment approaches (Notredame, 2007). Our results show that most gaps are incorrectly inserted in patterns that, nevertheless, tend to follow the guide-tree. Hence, gaps carry little additional information, distinct from that already present in the guide-tree. Although, the impact of this effect varies across datasets, some alignment algorithms are consistently more affected than others.

5.3.3 Methods.

Simulated and benchmark sequence datasets.

As a synthetic scenario in which the real history of insertion and deletion events is known, we worked with one of the simulated dataset previously used for the benchmarking of trimming methods (Capella-Gutiérrez et al., 2009). This consists of 600 sets of 32 simulated protein sequences each divided into 2 categories, asymmetric and symmetric, depending on the original tree topology used to simulate the alignments. In addition, we worked with a commonly used alignment benchmark dataset from BAliBASE v3.0 (Thompson et al., 2005). This consists of 386 set of protein sequences divided into 6 major datasets that are subdivided into 2 categories: Complete, containing all residues for all sequences, and Core regions, containing only manually curated homologous regions for all sequences.

Real sequence datasets.

We used two different sets of real sequences. First, the original data from Dessimoz and Gil (2010), which was used to show phylogenetic information in gaps, was accessed from their website cbrg.ethz.ch/research/msa. This dataset contains groups of orthologous proteins for three different taxonomic clades eukaryotes (609 orthologous groups), fungi (844), and bacteria (1,999). In addition, we dowloaded the original data from Marcet-Houben and Gabaldón (2009), accessed through the public database phylomedb.org (Huerta-Cepas et al., 2011). This dataset (phylome ID = 7), which we will refer to as yeast, contains trees for all *Saccharomyces cerevisiae* proteins across a phylogeny of 12 Saccharomycotina species. The data was filtered out to keep only 857 sets of 1-to-1 orthologous proteins.

Alignment programs.

We reconstruct MSAs using 5 different approaches, which could be classified depending on the scoring strategies into scoring-matrix-based Mafft FFT-NS-2 v6.712b (Katoh and Toh, 2008) and ClustalW v2.0.12 (Larkin et al., 2007), consistency-based Mafft L-INS-i v6.712b and TCoffee v9.01 (Notredame et al., 2000); and tree-aware-gap-placing Prank v.100701 (Löytynoja and

Goldman, 2008). All programs were used with default parameters. Additionally, SATé II (Liu et al., 2012), a program which combines the estimation of the MSA and the Maximum Likelihood phylogenetic tree, was used to evaluate its performance as an alternative to the rest of aligners used in this work.

Accuracy and precision of gap placement.

Using the true alignments from the simulated datasets and the reference alignments in BAliBase, we compared the opening positions of gaps in reconstructed alignments. Gap positions were recoded using the corresponding surrounding residues in reference and reconstructed alignments (see supplementary figure S1). Gaps opened between the same residues in the reference and the test alignment were considered true positives (TP), whereas those present only in the reference or in the test alignment were considered as false negatives (FN), and false positives (FP), respectively. Finally, true negative (TN) represent residues well-placed regarding to the number of gap-blocks opened prior to each residue. Precision was computed as P(aligner) = TP/(TP + FP) and accuracy was computed as A(aligner) = (TP + TN) / (TP + FP + TN + FN).

Tree discordance tests.

Reconstructed trees were compared in terms of their normalized split distance (Robinson and Foulds, 1981) with a canonical or a wrong tree. The canonical tree was the real tree in the simulated dataset and the canonical species tree for the Dessimoz and Marcet-Houben datasets (these trees are represented in supplementary figure S2). The ”wrong tree” is an alternative topology, which has the highest distance in terms of wrong splits (100%) to the canonical species tree. Since there are many possible wrong trees with the maximal distance to the canonical tree, for one of the datasets: simulated data - symmetric topology, we repeated the same procedure using 100 alternative possible wrong trees, the results obtained were similar (see supplementary figure 3), and thus a single wrong tree was used in subsequent analyses. The wrong trees used for the different datasets are provided in Supplementary figure S4). The ETE package (Huerta-Cepas et al., 2010) was used to perform all operations related to phylogenetic trees.

Gap parsimony reconstruction.

To assess the amount of phylogenetic information contained in gap patterns, we used the procedure proposed by Dessimoz and Gil (2010). That is, alignments are re-coded in presence/absence patterns of gaps (2-state character: for a given alignment, each column containing at least one gap was considered a character and the presence/absence of a gap its state); Subsequently, a maximum parsimony tree is reconstructed using the gap patterns from the recoded alignment (GP), using Wagner parsimony as implemented in Darwin v2.0 (Gonnet et al., 2000), and as described in Dessimoz and Gil (2010).

Maximum Likelihood phylogenetic reconstruction.

Maximum Likelihood (ML) phylogenetic trees were reconstructed using PhyML v3.0 (Guindon et al., 2010) with a discrete gamma-distribution model with four rate categories plus invariant positions, estimating the gamma parameter and the fraction of invariant positions from the data. LG was used as evolutionary model and branch and topology were optimized.

5.3.4 Results.

Most gaps in sequence alignments are incorrectly inserted.

Accuracy of sequence alignments is generally assessed on the basis of residue pairings, but only recently developed distance measures that also include similarities in terms of gap placement have been developed (Blackburne and Whelan, 2012). However, these distances do include information from residue pairing differences, making it difficult to assess what is the relative distance in terms of gap positioning and residue pairings. To assess to what degree gaps were inserted at correct positions we used reference alignments in BAliBASE (Thompson et al., 2005) and one set of simulated sequences (Capella-Gutiérrez et al., 2009). Sequences in these sets were re-aligned and the positions of the newly inserted gaps were compared with those in the reference alignments. Our results (Table 5.1 and Table 5.1) show that, in any given alignment, a significant fraction of the inserted gaps (30-90%) is placed at incorrect positions. This was true for all aligners,

and for both simulated and benchmark datasets. Surprisingly, ClustalW, a program that is usually outperformed by other aligners in terms of residue pairing (Kemena and Notredame, 2009), showed the best performance in terms of correctly placed gaps in the BAliBASE benchmark.

Program	Dataset	Precision	std	Accuracy	std
ClustalW2	*Asymmetric*	0.5082	0.0472	0.2178	0.0619
Mafft FFT-NS-2	*Asymmetric*	0.5358	0.0923	0.3925	0.0941
Mafft L-INS-i	*Asymmetric*	0.5895	0.0777	0.4288	0.1137
Prank+F	*Asymmetric*	0.4593	0.0606	0.5390	0.0710
T-Coffee	*Asymmetric*	0.5190	0.0433	0.5288	0.0388
ClustalW2	*Symmetric*	0.5083	0.0532	0.3604	0.0607
Mafft FFT-NS-2	*Symmetric*	0.6766	0.0548	0.6074	0.0726
Mafft L-INS-i	*Symmetric*	0.6808	0.0505	0.6403	0.0690
Prank+F	*Symmetric*	0.5898	0.0570	0.6781	0.0523
T-Coffee	*Symmetric*	0.5408	0.0600	0.5746	0.0690

Table 5.1: Accuracy and precision, in terms of gap placements, for the different strategies used to reconstruct MSAs divided according to the nature of the simulated data

Program	Dataset	Precision	std	Accuracy	std
ClustalW2	*BAliBASE*	0.2626	0.1308	0.2876	0.1299
Mafft FFT-NS-2	*BAliBASE*	0.1403	0.0838	0.2069	0.1024
Mafft L-INS-I	*BAliBASE*	0.1704	0.0908	0.2384	0.1130
Prank+F	*BAliBASE*	0.1128	0.0533	0.2028	0.0779
T-Coffee	*BAliBASE*	0.1189	0.0801	0.2017	0.1047

Table 5.2: Accuracy and precision, in terms of gap placements, for all alignments present in BaliBASE, a commonly used benchmark containing real cases.

Gap patterns follow the guide-tree, and carry little additional phylogenetic information.

If most gaps are incorrectly placed, how can gap-patterns carry phylogenetic information as suggested by recent reports (Dessimoz and Gil, 2010)? One possible explanation to this apparent conundrum is that gaps are placed following a pattern that is consistent with the phylogeny. Multiple

sequence aligners use evolutionary information that is provided by the guide-tree, a cladogram that dictates in which order the sequences are initially aligned to each other. This guide-tree is generally built from the pairwise distances of the sequences involved, and thus inherently carries phylogenetic information. In order to test the extent to which gap patterns follow the guide-tree, we measured the effect of altering the guide-tree. We tested this in the previously mentioned simulation dataset and in two real datasets: that used in Dessimoz and Gil (2010), which comprises alignments from bacteria, fungi, and vertebrate sequences, and one taken from Marcet-Houben and Gabaldón (2009) comprising sequences from yeast species. More specifically, we repeated each alignment in the previously mentioned datasets by using i) the normal procedure -enabling the program to build its own guide-tree-, ii) forcing the use of the correct tree (or a canonical species tree) as a guide-tree, and iii) forcing the use as a guide-tree of a synthetic "wrong" tree having the maximum split-distance to the correct tree.

If gap patterns are mostly dictated by the guide-tree, then the use of a very distinct guide-tree should have a large impact on the ability of gap patterns to reconstruct the correct tree. Indeed, under such conditions one would expect that information contained in gaps is biased towards the guide-tree to a degree that would reflect the strength of the guide-tree dependency of the aligner. Maximum parsimony reconstruction from patterns of gap presence/absence has been used to show that gaps contain unexploited phylogenetic information (Dessimoz and Gil, 2010). We thus applied the same approach using the three different strategies mentioned above. Since our procedure requires the program to enable using a userdefined guide-tree without altering it, we limited our analyses to ClustalW, T-Coffee, PRANK, and MAFFT, using the latter in two different modes: the consistency based L-INS-i and the progressive FFT-NS-2 (Larkin et al., 2007; Notredame et al., 2000; Löytynoja and Goldman, 2008; Katoh and Toh, 2008). Thus, although our choice of programs is limited, it covers a range of alignment strategies from progressive to iterative, going through consistencybased and phylogeny-aware strategies (Kemena and Notredame, 2009).

Figure 5.3 shows the distance to the correct tree, of Parsimony trees reconstructed from gap patterns (Gap Parsimony) in alignments using the

alternative three guide-trees mentioned above. In most cases, the use of the wrong tree as a guide-tree destroyed most of the signal towards the true tree, indicating that wrong guide-trees mislead gap placement. Conversely, the use of the correct tree as a guide tends to improve the phylogenetic information contained in gaps. These results indicate, as expected, that guide-tree accuracy is an important factor determining the phylogenetic information contained in gaps. However, this does not solve the issue of whether gaps harbor additional information as compared to the guide-tree. Some additional lines of evidence suggest that gaps mostly carry information dictated by the guide-tree. Firstly, alignments reconstructed from wrong guide-trees carry phylogenetic information pointing towards that wrong topology (supplementary figure S5). Secondly, the guide-tree reconstructed by the alignment program is generally a better estimator of the true topology than the tree reconstructed from gap patterns (supplementary figure S6), indicating that the use of gap parsimony actually erodes, rather than increases, phylogenetic information contained in the guide-tree. Finally, Gap Parsimony trees were reconstructed for the simulated alignments without realigning them to evaluate whether these perfectly placed gaps are able to resemble the trees used to generate them or not. As it can be seen (Figure 5.3 yellow dashed lines) simulated gaps cannot properly reconstruct the simulated phylogeny. Of note, the normal process of alignment reconstruction (blue dots) significantly erases the signal in gaps, and only in some cases, and always using the canonical tree as a guide (green dots), the recovered signal is similar to the one present in real gaps.

A measure for guide-tree dependency.

We have shown that most gaps are inserted incorrectly, but following a pattern mostly dictated by the guide-tree. These effects seem to be present in all programs but to different degrees. A measure that would allow us to comparatively assess the guide-tree dependency of the different aligners in terms of their gap placement would be useful to make informed choices of methodologies or parameters. We here propose the following methodology to derive a simple measure that captures the effect of guide-tree: Given a two-dimensional space where the coordinates are, respectively, the split

distances to i) a canonical tree (the true tree) and to ii) a wrong tree with maximum split-distance to the canonical tree, a given tree topology could be represented by its respective coordinates. If two alternative trees, each one derived from a different alignment using either the canonical tree or the wrong tree as a guide, are projected into this space. Then, the euclidean distance between these points will effectively measure the effect on the topology of altering the guide-tree. Such a plot and the derived distance is shown for the bacterial dataset and ClustalW2 (Figure 5.4). In this framework a high level of guide-tree dependency will produce trees that are close to the guide-tree thus maximizing the distance in the mentioned space. We computed this value, which we will refer to as *guidescore*, for other combinations of aligners and datasets (Figure 5.5). Our results indicate that the phylogeny-aware method PRANK is generally the most dependent on the guide-tree. This distance measure can be applied to assess the effects of guide-trees on other reconstruction methods, and we here assessed the impact of guide-tree on Maximum Likelihood reconstruction, using the same framework (Figure 5.6). Our results indicate that guide-tree determination affects ML phylogenetic reconstruction to a much lower degree than gap parsimony, suggesting that gap patterns are more affected by guide-tree determination than residue pairings.

Strategies to overcome guide-tree dependency of gap placements.

We finally set out to explore potential strategies that would serve to overcome the shown effect of gap tree dependency on gap placement. In particular we explored two possible strategies i) minimize the effect of guide tree dependency and ii) select gaps that are more likely to contain true phylogenetic information (use of consistency-based aligment trimming). We want to note that our intention is not to explore the full range of possibilities but rather to show that the observed effect can be tackled. Intuitively, methods that iteratively reconstruct trees and alignments, such as that implemented in SATé (Liu et al., 2012), should be less prone to the effect of an initially-set guide tree. Similarly, averaging over different aligners by means of consistency-based methods such as M-Coffee (Wallace et al., 2006), would be expected to minimize the effect. Indeed, as shown in Supplementary

figure S7, both strategies were found to be among the least affected by the guidetree in most of the datasets. Finally, besides minimizing the effect, one may wish to select those gaps that are less likely to be the result of guide-tree guidance and thus expected to contain independent phylogenetic signal. To do so, we investigated whether consistency-based trimming, as the one implemented in trimAl v1.4 (Capella-Gutiérrez et al., 2009), served to select gaps that are more likely to contain true phylogenetic information. To do so we aligned each set of sequences in forward and reverse orientation (i.e. Head or Tails approach (Landan and Graur, 2007)) and then trimmed the alignment using trimAl with a cut-off of 0.05 consistency score. Our results indicate that the precision of gaps present in trimmed alignments was significantly higher than in nontrimmed ones (Supplementary tables 1 and 2).

5.3.5 Discussion.

Altogether our results indicate that most of the apparent phylogenetic signal carried by gaps in this analysis is actually a result of the preferential inclusion of shared gaps in sequences that are closer in the guide-tree. In other words, under these circumstances, many gaps do not contain additional phylogenetic information per se but rather reflect information already present in the guide-tree. Several lines of evidence support this. First, the initial guide-tree produced by the alignment program is highly similar to the canonical tree (Supplementary Figure S6) indicating that it carries a strong phylogenetic signal. Importantly, this guidetree is usually more similar to the canonical tree than the parsimony tree, solely based on gap information, indicating that the use of gaps in a parsimony framework actually erases part of the signal contained in the guide-tree. Secondly, the use of a clearly wrong guide-tree to guide the process erodes the phylogenetic signal contained in gaps and biases it towards the wrong tree topology (Figure 5.5, and Supplementary figure S5). Blackburne and Whelan (2012) already noted that the different placement of gaps by different aligners rarely altered the inferred evolutionary histories of insertions and deletions events, but failed to propose a possible source for such apparent contradiction. Our results provide an answer to this conundrum by showing that all aligners follow a

similar guide-tree in different ways, thus resulting in disparate gap patterns that are nevertheless compatible with the same guide-tree.

We consider that these results are not in contradiction with the idea that insertions and deletions are rare evolutionary events that can be used for phylogenetic reconstruction. Indeed we share the opinion of Dessimoz and Gil (2010) and others that an effort should be made in finding new ways of exploiting this information. We consider, however, that a necessary step is to disentangle what fraction of the apparent signal results from the guide-tree, and identify those informative gaps that are carrying truly new phylogenetic signal in order to avoid biases. As we have shown, in current algorithms, the guide-tree and arbitrary gap parameters seem to dominate the nature and strength of the signal carried by gaps. This effect may be even stronger in alignments with more sequences and higher divergence. Finally, we have shown possible solutions to alleviate this effect, which include iterative reconstruction and the use of consistency across different alignments.

5.3.6 Acknowledgements

The authors want to thank Cedric Notredame for discussions on this topic.

Funding: TG group research is funded in part by a grant from the Spanish Ministry of Science and Innovation (BFU2009-09168).

5.3.7 Supplementary material

Supplementary material can be found online at: SupplementaryMaterial.Capella-Gutierrez&Gabaldon.AreGapsInformative.pdf

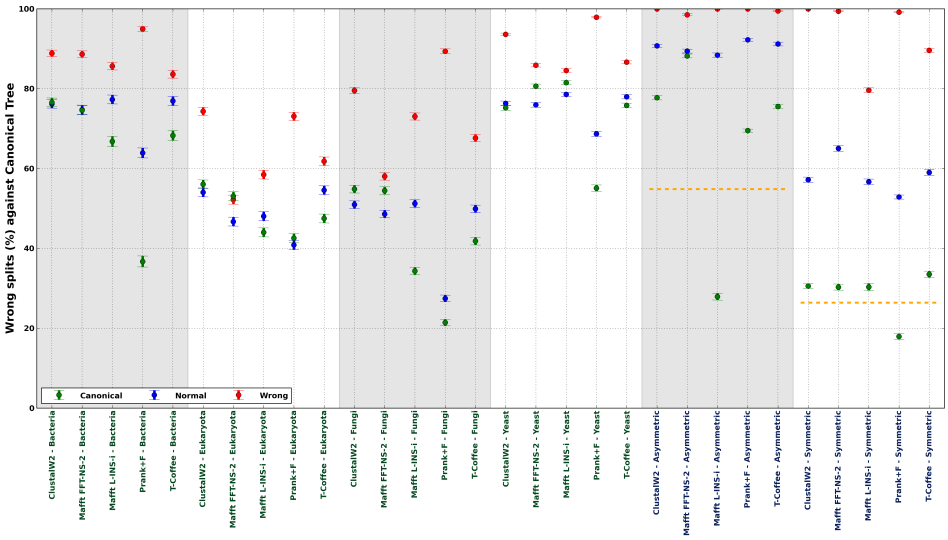

Figure 5.3: Mean distance, in term of wrong splits, to the Canonical trees of the different Gap parsimony trees reconstructed after allowing to the programs to build its own guide-tree (blue dots) or forcing them to use either the canonical tree (green dots) or an alternative topology (red dots), with maxium splitdistance to the canonical tree. Wrong splits measure the number of topological differences between two given trees. Yellow dashed lines in the simulated datasets indicate the signal retrieved from the real gaps using the same gap-parsimony approach.

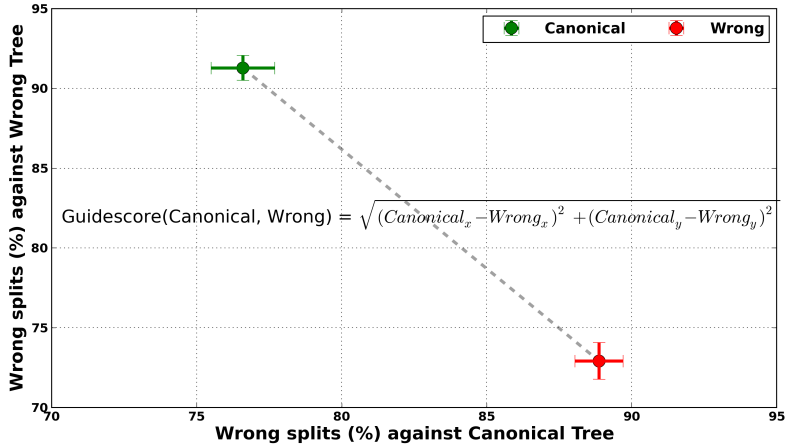

Figure 5.4: Example showing how to compute the *guidescore* for two alternative (sets of) trees computed using different approaches. In this case, the score is computed considering the Gap parsimony trees inferred after a normal execution of ClustalW2 and those after forcing to use an alternative topology with the maximum split-distance to the canonical tree

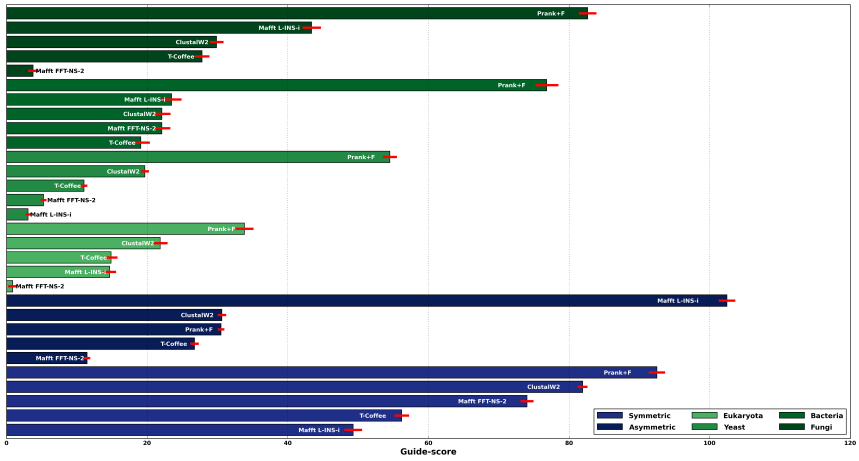

Figure 5.5: *Guidescores* computed for all available datasets, simulated data in blue and real data in green, for all approaches mentioned in the study. Gap parsimony trees for normal execution and forcing programs to use a maximum split-distance tree to the Canonical tree were used to compute the score.

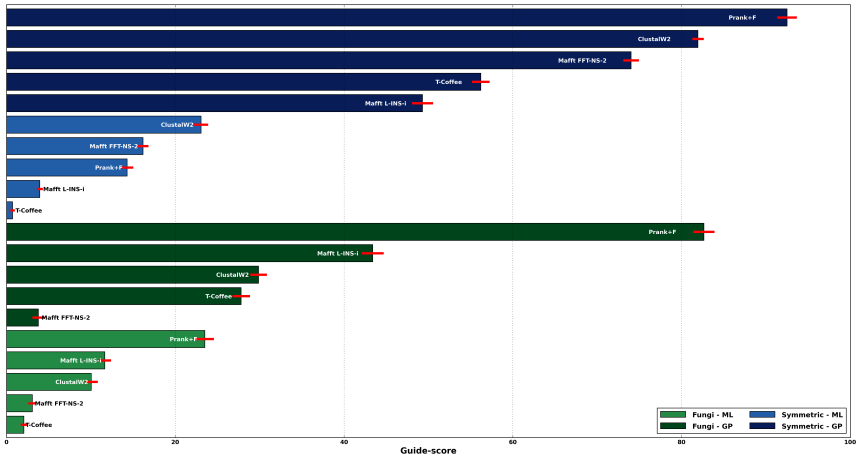

Figure 5.6: *Guidescores* for two datasets, one simulated (blue) and another one real (green) using all available methods and considering in this case two alternative approaches for reconstructing phylogenetic trees: Gap parsimony (darker colors) and Maximum likelihood (lighter colors). Guide scores were computed between trees inferred after forcing programs to use either the canonical reference trees or trees with maximum split distance to the reference one.

6 Resolving the phylogenetic position of an elusive taxon: Microsporidia

6.1 Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi

Salvador Capella-Gutiérrez, Marina Marcet-Houben & Toni Gabaldón

Originally published in: *BMC Biology. June, 2012.*

Capella-Gutierrez S, Marcet-Houben M, Gabaldon T. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. BMC Biol. 2012 May 31;10:47.

7 Quest for phylogenetically stable gene markers.

7.1 A phylogenomics approach for selecting robust sets of phylogenetic markers.

Salvador Capella-Gutiérrez, Frank Kauff & Toni Gabaldón

In preparation

A phylogenomics approach for selecting robust sets of phylogenetic markers.

7.1.1 Abstract.

Reconstructing the evolutionary relationships of species is a major goal in biology. Despite the increasing number of completely sequenced genomes, a large number of phylogenetic projects rely on the targeted sequencing and analysis of a relatively small sample of marker genes. The selection of these phylogenetic markers should ideally be based on accurate predictions of their combined, rather than individual, potential to accurately resolve the phylogeny of interest. Here we present and validate a new phylogenomics strategy to efficiently select a minimal set of stable markers able to accurately reconstruct the underlying species phylogeny. In contrast to previous approaches, our methodology does not only rely on the ability of individual genes to reconstruct a known phylogeny, but it also explores the combined power of sets of concatenated genes to accurately reconstruct trees of species not previously analyzed. We applied our approach to two broad sets of cyanobacterial and fungal species, and provide two minimal sets of seven and four genes, respectively, necessary to fully resolve the target phylogenies. This approach paves the way for the informed selection of phylogenetic markers in the effort of reconstructing the Tree of Life.

7.1.2 Introduction.

Evolutionary relationships among species have been traditionally inferred using ribosomal genes (Woese and Fox, 1977), especially 16S, given their ubiquity and high degree of conservation. With the increasing availability of completely sequenced genomes, however, we have now a whole range of genes at our disposal. Several phylogenomics approaches aim at using most of the information available on sets of complete genome sequences to derive a species phylogeny (Delsuc et al., 2005), however there is still the need to select phylogenetic marker genes to target unsequenced species. This poses the important question of which combination of genes is the

most informative to establish the phylogenetic relationships of a given group of organisms. Earlier work has focused on ranking phylogenetically informative genes based on their availability of reconstructing a known species phylogeny (Aguileta et al., 2008; Walker et al., 2012). The assumption is that genes which carry sufficient information to reconstruct the known part of the phylogeny are expected to do similarly well in so far unsampled regions of the tree. However, this assumption is usually not proven within the framework of phylogenetic marker selection. Additional limitations of current marker selection procedures is that individual genes, rather than combinations of genes, are ranked. Ideally, an informative set of genes should be present in the studied species and remain informative when more taxa are added to the study. In addition, to limit costs of targeted sequencing, this set should be of a minimal possible size, but of sufficient size to carry enough information to reconstruct a phylogeny that goes beyond the one used in the selection phase.

To address these limitations, we here present a method to automatically identify, from whole genome sequences, small subsets of widespread genes that can accurately reconstruct the target phylogeny. In contrast to previous methods our approach ranks combinations of genes, rather than individual genes. In addition our approach uses a cross-validation technique to ensure high accuracy when using sequences not previously seen in the marker selection step, thus better reflecting real scenarios. To validate our method we applied it to the selection of phylogenetic marker genes in a prokaryotic group -cyanobacteria- and a eukaryotic group -fungi. Our results indicate that small sets of 7 and 4 genes, respectively, are able to precisely recover the target phylogenies, even when including species not used for the selection of markers.

7.1.3 Material and Methods

Sequence data.

Proteins encoded in 63 and 83 completely sequenced genomes from Cyanobacteria and Fungi, respectively were downloaded from different sources.

Phylogenetic tree reconstruction for individual genes.

Once sets of widespread single copy proteins were identified (see below), the pipeline described in Huerta-Cepas et al. (2011) was used to infer single gene tree phylogenies. In brief, sequences were aligned using three different programs: MUSCLE v3.8 (Edgar, 2004), MAFFT v6.712b (Katoh and Toh, 2008), and DiAlign-TX (Subramanian et al., 2008). Alignments were performed in forward and reverse direction (i.e using the Head or Tail approach (Landan and Graur, 2007)), and the six resulting alignments were combined into a consensus alignment using M-Coffee (Wallace et al., 2006). The resulting combined alignment was subsequently trimmed with trimAl v1.4 (Capella-Gutiérrez et al., 2009), using a consistency score cutoff of 0.1667 and a gap score cutoff of 0.1, to remove poorly aligned regions. Then, phylogenetic trees based on Maximum Likelihood (ML) approach were inferred from these alignments. ML trees were reconstructed using the best-fitting evolutionary model, which as selected as follows: A phylogenetic tree was reconstructed using a Neighbour Joining (NJ) approach as implemented in BioNJ (Gascuel, 1997); The likelihood of this topology was computed, allowing branch-length optimisation, using seven different models (JTT, LG, WAG, Blosum62, MtREV, VT and Dayhoff), as implemented in PhyML v3.0 (Guindon et al., 2010); The two evolutionary models best fitting the data were determined by comparing the likelihood of the used models according to the AIC criterion (Akaike, 1974). Then, ML trees were derived using these two models, using the default tree topology search method NNI (Nearest Neighbor Interchange), and the one with the best likelihood was used for further analyses. A similar approach based on NJ topologies to select the best-fitting model for a subsequent ML analysis has been previously shown to be highly accurate (Huerta-Cepas et al. (2011)). Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution, as implemented in PhyML. In all cases, a discrete gamma-distribution with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data.

Reconstruction of reference species trees.

Alignments from selected sets were concatenated to reconstruct a single reference tree using as evolutionary model the one which best fits the data for most of the cases. Then, a ML tree was derived using as tree topology search method SPR (Subtree Pruning and Regrafting), a discrete gamma-distribution with four rate categories plus invariant positions and estimating from the data the gamma parameter and the fraction of invariant positions. Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution, as implemented in PhyML v3.0 (Guindon et al., 2010).

Construction of training and testing sets for cross-validation.

Available genomes were split into two sets: i) the training set, accounting for around 2/3 of the available genomes, which was used to identify potential gene markers, and ii) the training dataset, comprising the remaining 1/3 of the genomes, which was used to evaluate whether marker genes were widespread and phylogenetically informative when species not included in the selection of markers are included. Composition of both sets are made randomly but a manual inspection phase ensured that representatives for the different taxonomic groups were present in both training and testing sets.

Ranking of individual phylogenetic marker genes.

A first step in the selection process identifies widespread proteins present in single-copy in all genomes of the training set. This is done by performing a BLAST (Altschul et al., 1990) search from a seed species into all other genomes, and selecting those proteins with a single hit (e-value cut-off 10-5 and coverage ¿50%) in every other genome. The cut-off in terms of number of species in which the marker should be present could be relaxed if a very limited number of genes fulfill this criterium. The selection of the seed species is arbitrary and more than one seed can be used in order to increase the number of detectable single-copy proteins. Here, we selected multiple seed species, one from each of the four and five major phylogenetic groups in Cyanobacteria and Fungi, respectively (species used as a seed are marked

figures 2 and 3, respectively, with dark grey boxes). Each seed species defines a set of widespread proteins, which may overlap significantly with those obtained from the other species. The union of all sets is used as the initial set of widespread groups of homologous proteins present in single copy. Then, proteins in each homologous group is aligned and used to reconstruct a gene tree (see above). In addition, a reference species tree is reconstructed from the concatenation of all alignments. Then, each marker is ranked according to its ability to reconstruct the reference phylogeny. For this the distance of the reconstructed individual gene tree to the reference phylogeny is measured using the Robinson and Foulds distance (Robinson and Foulds, 1981).

Selection of combined minimal sets of phylogenetic marker genes.

In order to get a first estimation of how many phylogenetic markers are needed to fully recover the reference species tree, a progressive concatenation of individual gene sets is performed as follows: Sets of genes are concatenated, progressively, according to its score against the reference tree. That is the n top-scoring markers are concatenated and used to reconstruct a species phylogeny (see below). This is repeated from n = 2 to n = m, being m the minimal number of concatenated marker genes that reach a cut-off Robinson and Foulds distance to the reference tree (we here used a distance cut-off of 0). This set of m marker genes is referred to as the *initial marker set*. Then, another iterative phase is started to find subsets of size smaller than m, which nevertheless reach the same cut-off distance. To do so sets of size ranging from 2 to m-1 are formed by randomly subsampling genes from the *initial marker set*. Each subset is scored according to the Robinson and Fould distance to the reference tree. This iterative process finishes when either i) all possible combinations have been explored, ii) at least one smaller combination with distance lower than the cut-off has been found, or iii) a number of predefined iterations has been reached (we here explored a minimum of 100 combinations). When one smaller combination is found, the iterative process can be re-started, setting that combination as the *initial marker set*. On the case of exploring all possible combination without finding a smaller sets of genes, the *initial marker set* is returned as

the minimum possible concatenation of individual gene sets that recover the reference tree.

Validation.

After the selection of one or more potential combinatorial sets of marker genes, two different tests are carried out to verify the ability of these sets to properly reconstruct the phylogeny when including species not present in the selection phase. Firstly, for each genome in the testing set, a new reference tree is reconstructed as described above that includes all species in the training set plus the new species. Then, the ability of the phylogenetic marker set to recover that topology is measured by reconstructing a phylogeny of the same set of species using only the set of phylogenetic marker genes. In a second test, only the new set of genomes (the testing set) is used to derive the reference topology and the tree based on the set of marker genes. Marker genes and the set of widespread genes are found in the new genomes using BLAST (see above) searches from the ones identified in the training phase. Then, both topologies are compared in terms of the Robinson and Foulds distance to evaluate the ability of the set of marker genes to recover the reference topology. Results on the cross validation tests provide the means to choose among different marker gene sets derived in the first phase, and to estimate the ability of the selected markers to go beyond the species used in their selection.

7.1.4 Results.

From individual gene markers to combined sets.

The rationale behind the proposed methodology is that phylogenetic marker genes are generally used in combination, rather than in isolation, and that their performance to reconstruct accurate phylogenies should be evaluated beyond the set of species used for their prioritization. Like other recently developed genome-wide methods (Aguileta et al., 2008), our procedure starts by evaluating the ability of single gene trees to recover a reference species phylogeny. This produces a ranked list of marker genes. While other procedures stop there, ours goes one step further and evaluates

combinatorial subsets of marker genes. This is done using a multi-attribute optimization of two conflicting criteria: a minimal gene size, and a maximal information content. A final cross validation step evaluates the performance of such selected gene marker subsets to reconstruct accurate phylogenies including species not previously seen.

In brief, our proposed pipeline proceeds as follows (see figure 1), additional details are provided in the material and methods section. The set of available genomes for a given taxonomic phylum is divided into two sets. One set, comprising two thirds of the available genomes, is used as the training set to prioritize sets of phylogenetic markers. This is done by creating a reference phylogeny using all available widespread, single-copy genes and testing the ability of each individual gene marker to recover this phylogeny. This produces a ranked list of phylogenetic gene markers. A first iteration will define the minimal set of marker genes to be used in combination by sequentially adding marker genes following their order in the ranked list, until a phylogenetic analyses of the concatenated alignments of the marker genes reaches a predefined distance to the reference tree. Here we used a Robinson and Foulds (Robinson and Foulds, 1981) distance of 0 as a cut-off, but other thresholds and distance measures could be used. The genes included in this combination, referred to as *initial marker set*, constitute the entry point for a second iteration aiming at finding smaller subsets of marker genes, which nevertheless have the same potential to recover the reference phylogeny. This second iteration finds one or several combinations of phylogenetic markers that are evaluated for their potential to recover a reference phylogeny that includes species not present in the training set. For this the remaining one third of available genomes not used in the training set, i.e. the testing set, are used in two different ways: 1) each one of the genomes of the test set is added to the training set, and the selected sets of marker genes are tested for their ability to correctly place the newly added species; and 2) the selected sets of gene markers are tested using only the genomes in the test set. Altogether, the results of the two iterations of phylogenetic marker set selection and the validation analyses constitute a valuable source of information on the ability of selected combinations of marker genes to assess the phylogenetic relationships of species beyond those used in the

marker prioritization. To assess the potential of this pipeline for the selection of marker genes in real data, we chose one prokaryotic and one eukaryotic phyla for which the reconstruction of the tree of life is an active field of research: Cyanobacteria and Fungi.

Seven gene markers for cyanobacteria phylogeny.

Cyanobacteria are prokaryotes capable of oxygenic photosynthesis, and the origin of the chloroplasts of today's green plants. Being about 3.5 billion years old (Schopf, 2002), they now inhabit all ecosystems and continents on earth, including the Antarctic. Taxonomy and phylogeny were always challenging in the cyanobacteria. For prokaryotes, they are comparatively feature-rich in their morphology, but still the number of morphological traits is insufficient to provide enough information for a phylogenetic analysis. Already the first molecular analyses based on 16S rDNA only suggested that the traditional classification of cyanobacteria is highly artificial. e. g. Giovannoni et al. (1988). Although the 16S rDNA ist still the most common phylogenetic marker in cyanobacteria, other genes have been used to generate phylogenies at various taxonomic levels, e.g. gyrB, rpoC1, rpoD1 (Seo and Yokota, 2003), nifD (Henson et al., 2004), and others (see (Kauff and Büdel, 2011) for an overview). However, the availability of specific single locus data varies tremendously across taxa and species, and the number of taxa for which sequence data is available decreases quickly as the number of loci increases. As a result, data sets with larger numbers of loci often include only few cyanobacterial taxa.

We applied our proposed method to identify reliable sets of markers genes using 62 species with completely sequenced genomes. In the training phase, a reference tree for 43 species was inferred using a concatenated alignment of 287 single-copy genes present in all species (Figure 7.2 panel A). This tree is fully congruent, for the shared species, with a recently phylogeny based on 340 genes (Swingley et al., 2008), except for the relative positions of *Acaryochloris marina* and *Thermosynechococcus elongatus*. Our pipeline defined an initial marker set of 34 genes able to fully recover the reference phylogeny. The iterative search for smaller sets with an equal potential yielded a subset

of seven genes (see table 1). Cross-validation tests on the remaining 19 genomes showed that in 17 (89%) of the cases the seven marker genes were found and they were able, when used in combination, to correctly place the test species. In the remaining two cases, only six of the seven marker genes were found, which yielded a topology that correctly placed the test species but which showed small differences with the reference trees (2.5% of different splits). Finally, when the seven marker genes were used to reconstruct the 63-species phylogeny including all genomes in the training and testing sets, they resulted in a topology largely similar to the reference tree (Figure 7.2 panel C) except for two conflicting nodes, of which one is due to a change in the arrangement of some strains of the same species.

Four gene markers for the fungal tree of life.

With estimated 1.5 million species (Hawksworth, 2001), fungi constitute one of the most diverse eukaryotic groups. In addition, their generally unicellular organization and their broad phenotypic and metabolic plasticity makes genetic approaches the best suited for establishing fungal diversity and phylogenetic relationships. Previous studies to establish phylogenetic relationships in fungi have used widespread gene markers such as subunits 1 and 2 of RNA polymerase II, elongation factor 1, -tubulin, and mitochondrial ATP synthase (James et al., 2006; Walker et al., 2012). In addition, as a result of the growing availability of fully-sequenced fungi, genome-wide approaches are increasingly being used (Marcet-Houben and Gabaldón (2009) and others). Despite large international initiatives to sequence thousands of fungal genomes (e.g. http://1000.fungalgenomes.org), the need for phylogenetic markers to target a broader diversity as well as unculturable species will still exist for the coming years. We thus applied our approach to select stable phylogenetic markers using 83 available fungal genomes belonging to the Ascomycetes taxonomic group. A reference phylogeny based on 169 widespread, single copy genes of the 55 species in the training set is largely congruent (Figure 7.3 panel A), for the shared species, with earlier reconstructed trees (Wang et al., 2009; Capella-Gutiérrez et al., 2012). The sequential concatenation of markers in decreasing order of their phylogenetic informativeness, defines an initial marker set of seven

genes to accurately recover the reference topology. The subsequent sampling and testing of subsets reduces the number of necessary markers to only 4 genes (table 2). Of note this number is smaller than the six-gene marker set used in previous large-scale phylogenetic surveys of fungi (James et al., 2006; Schoch et al., 2009). A validation of this set of marker genes in the testing set, comprising 28 species, showed that in most cases (25 genomes) the four genes could be found in single copy, while in three genomes one of the marker genes was missing or in multiple copies. In all cases the gene marker set was able to reconstruct an expanded phylogeny with less than 6% of wrong-splits, being fully congruent in 12 (43%) of the species. In the second test performed using only the genomes from the testing set, full agreement was found between the trees derived using either the complete sets of single-copy genes or just the set of marker genes (Figure 7.3 panel B). Altogether our results show that the four selected gene markers, used in combination, have a strong potential to reconstruct accurate phylogenies of fungal species (Figure 7.3 panel C) and that they will be valuable in the expansion of the fungal tree of life.

7.1.5 Concluding remarks.

Reconstructing the Tree of Life is a daunting task that will require the combination of diverse efforts and methodologies. It is most likely that the expansion and increase in resolution of the Tree of Life will proceed through the agglutination of several studies. Some, based on complete genomes will establish a backbone of the main lineages, while some more focused studies will resolve internal diversity within a specific clade based on targeted markers. In addition the expansion of the Tree of Life towards less explored clades will likely proceed in a two steps manner. First, an overview of phylogenetic relationships within the new clades will be sketched through targeted amplification and analysis of selected phylogenetic markers. Then, based on these results, several species will be selected to be completely sequenced and thus provide a first backbone of the new clade, from which to build on in order to increase resolution. In all these contexts, the informed selection of phylogenetic marker genes constitute a necessary step. Here, we have developed a new approach that is based on the selection of sets

marker genes from completely-sequenced based on their combined power to resolve a reference phylogeny. The assessment of combination of genes, rather than in isolation, constitutes one of the major novelties of the proposed approach. This, in our view better reflects current scenarios in which several, rather than a single phylogenetic marker is obtained from a set of selected species. In addition, as we have shown here for Cyanobacteria and Fungi, the exploration of combinatorial effects of the combination of good phylogenetic markers is able to reduce the number of selected markers while keeping a similar potential for phylogenetic reconstruction. Furthermore, our procedure comprises a cross validation test to assess the performance of the selected markers outside the genomes used in the selection of marker genes. This is, to the best of our knowledge, the first time that such a validation is built-in in the marker selection pipeline. As shown here, the validation test provides information on how the gene markers will behave when used on additional species, as well as an indication of how the resolving power may diminish when expanding the tree to include other species within the clade. These are important considerations for the selection of phylogenetic marker genes, and for which tools were so far lacking. Thus our proposed approach fills in an important gap in the field of phylogenetic marker selection. Additional criteria, such as the suitability of markers for primer design and experimental amplification are not specifically tackled here and should be considered in downstream analysis. Altogether our results show that our approach is a valuable tool for the informed selection of phylogenetic markers.

7.1.6 Acknowledgements.

We thank Ben Lehner and members of the Gabaldons group for discussions on this project.

Funding: TG group research is funded in part by a grant from the Spanish Ministry of Science and Innovation (BFU2009-09168).

Uniprot Id	Length	Evidence	Description
B2IVU1	246 AA	Inferred from homology	Probable 2-phosphosulfolactate phosphatase.
B2J427	979 AA	Inferred from homology	Glycine dehydrogenase [decarboxylating].
B2IT89	480 AA	Inferred from homology	Trigger factor.
B2IW68	816 AA	Inferred from homology	Phenylalanyl-tRNA synthetase, beta subunit.
B2J5B7	719 AA	Predicted	RNA binding S1 domain protein.
B2J6R0	312 AA	Predicted	Cytochrome oxidase assembly.
B2J980	1087 AA	Inferred from homology	Carbamoyl-phosphate synthase, large subunit.

Table 7.1: List of selected phylogenetic marker genes in Cyanobacteria. Protein information has been taken from *Nostoc punctiforme*.

Uniprot Id	Length	Evidence	Description
YHR186C	1557 AA	Evidence at protein level	Target of rapamycin complex 1 subunit KOG1
YMR012W	1277 AA	Evidence at protein level	Clustered mitochondria protein 1
YJL029C	822 AA	Evidence at protein level	Vacuolar protein sorting-associated protein 53
YAR007C	621 AA	Evidence at protein level	Replication factor A protein 1

Table 7.2: List of selected phylogenetic marker genes in Fungi. Information about protein is related to *Saccharomyces cerevisiae*.

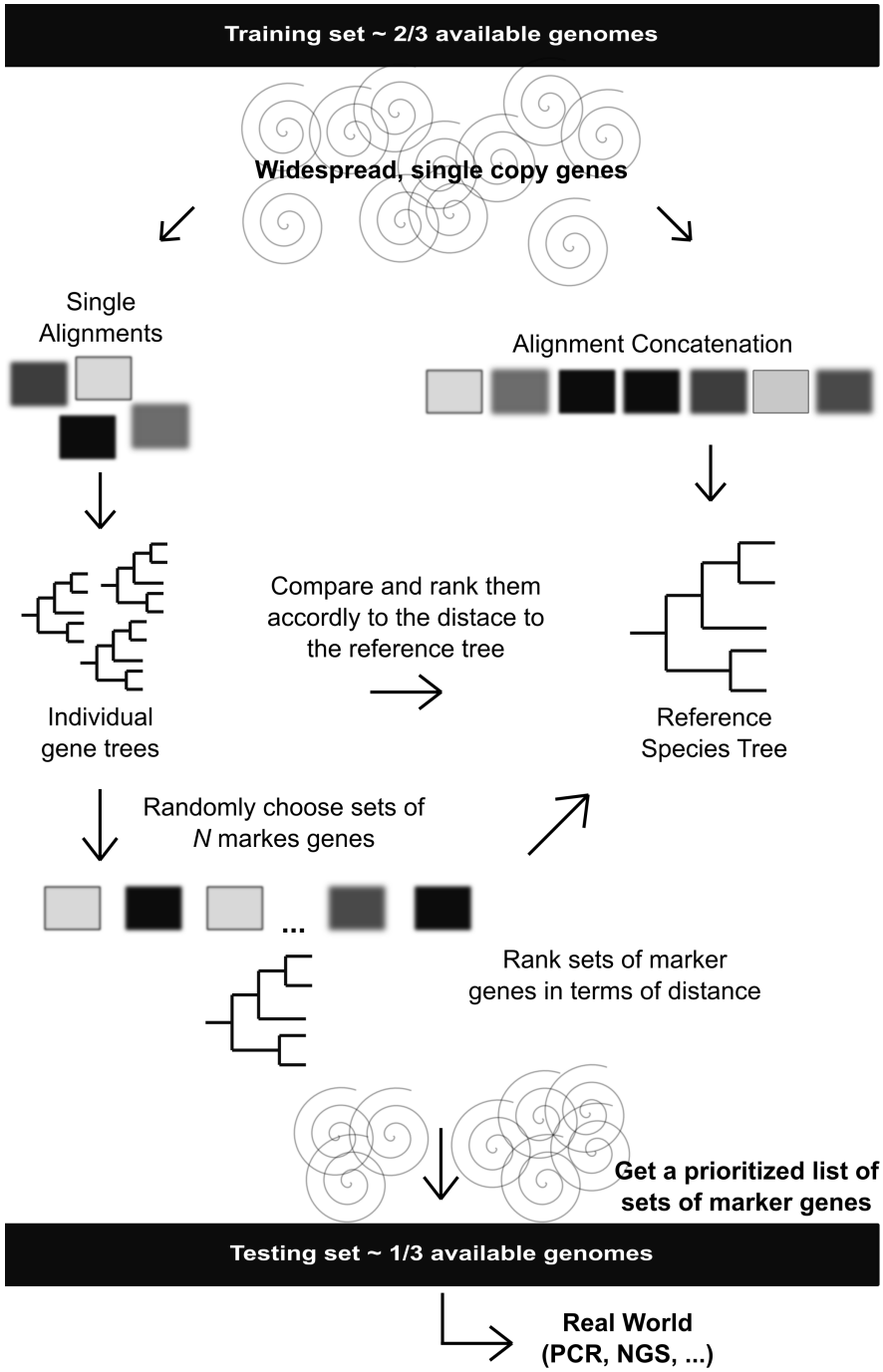

Figure 7.1: Schematic representation of the marker selection pipeline including the training and the testing phases. The iterative process of searching potential sets of marker genes finishes when either a group of genes, smaller than *initial marker set*, is found or all possible combinations have been explored.

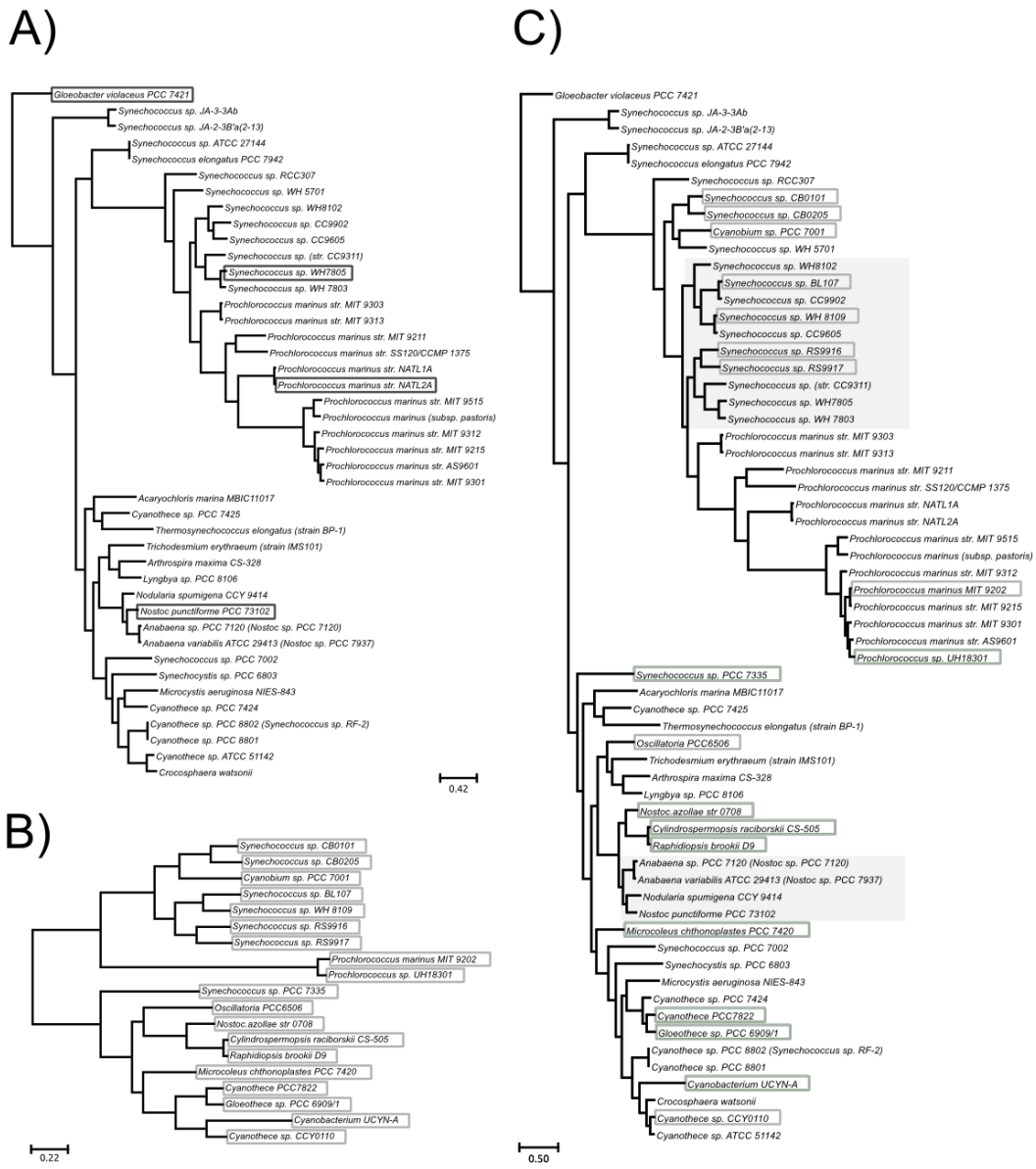

Figure 7.2: Cyanobacterial phylogenetic trees comprising different sets of species. A) Reference tree for the training phase, comprising 43 species, used to look for potential groups of marker gene. Dark gray boxes indicate which species were used as seed to perform the BLAST search. B) Tree derived from the concatenation of the group of 7 gene markers. Species have been marked with light gray boxes to highlight their positions in the bigger tree. C) Tree inferred using the concatenation of the set of 7 gene markers for the 62 species used in this study. Shadow boxes indicate conflicting nodes from this tree and the previously ones reconstructed.

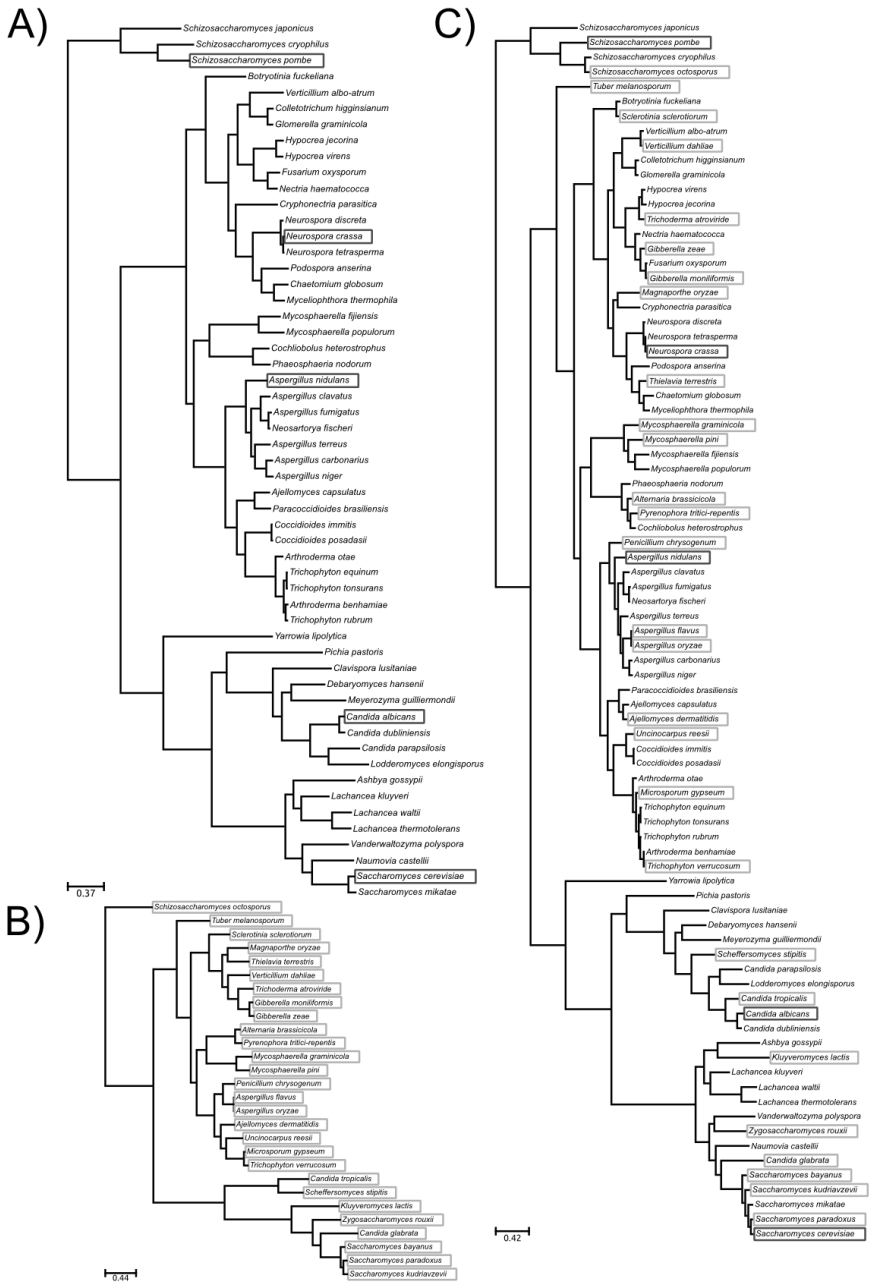

Figure 7.3: Fungal reference trees for different set of species. A) Species tree used as reference in the training phase. This tree has been inferred using the concatenation of 169 single-copy proteins present in 55 species. Dark grey boxes indicate which species have been used as seed to perform the BLAST search. B) Phylogenetic tree inferred using the set of 4 marker genes for the 28 species used in the testing phase. C) Phylogenetic tree inferred from the set of 4 marker genes for all species in the study. Full agreement between this tree and the trees inferred in the training and testing phase when species are superposed.

8 General discussion

During this thesis I have worked on different methodological aspects related to phylogenomics. In addition, to evaluate the applicability of such work to real data, I have used these methods to resolve long-standing biological questions. Each chapter of this thesis has its specific discussion section. Here, I will summarize the general implications of my research at the same time I will offer my particular view about the future of phylogenomics.

8.1 Reconstructing genome-wide collections of phylogenetic trees.

Since the first phylogenetic tree was constructed from molecular sequences by Zuckerkandl and Pauling (Zuckerkandl and Pauling, 1965), many and diverse methodological improvements have been achieved. Nowadays, we are able to construct thousands and thousands of single gene trees in an automated manner using powerful computers and sophisticated methods. In this context, I have contributed to the development of a very efficient pipeline that enables automated evolutionary analyses of newly-sequenced genomes. Moreover, this pipeline constitutes the basis of the largest repository of pre-computed phylogenies (phylomeDB). Although current approaches can deal with the heterogeneous nature of the data, there is still room for further improvements.

In a phylogenomics context, the first step for reconstructing phylogenetic trees is the search of homologous sequences. It is a common practice to use only the longest gene transcript in the process, since this is the one carrying the most information. However, the use of the longest transcripts, also known as isoforms, is not always the most appropriate strategy due to extreme variations, in terms of sequence length. Such variations can make difficult the reconstruction of alignments and, therefore, impact negatively in the final results. It is expected that the situations get worse with the deluge of data coming from different genome projects, especially those generating transcriptomic data. So, a further improvement of our pipeline would ideally consider more clever ways to select upon several alternative transcripts.

Selecting the most appropriate combination of transcripts is not a trivial task. All possible combinations have to be evaluated and scored to select the one which maximizes the information in terms of alignable residues. However, the availability of different isoforms per gene offers a new opportunity to evaluate the robustness of methods to infer phylogenetic trees. Using, for instance, the best combination of isoforms, according to any scoring system, and the longest isoforms it is possible to compare their agreement in terms of the reconstructed trees. On a more biological ground, studying the evolution of isoforms can offer a new opportunity to understand how different mechanisms have evolved and whether different gene parts are under different evolutionary constraints. For instance, the longest isoform of gene p53 is involved in apoptosis, whereas a shorter variant is preventing the occurrence of such mechanism of cellular death (Jänicke et al., 2008).

After identifying sets of homologs, multiple sequence alignments are reconstructed prior to inferring phylogenetic trees. With the years, many improvements and alternatives have been proposed to build alignments, so, nowadays, in a context of a continuously growing number of sequences, the challenge is how to align them. Independently of how complex the methods are, the progressive nature of the alignment process makes it difficult to escape from the effect of algorithmic decisions taken in earlier stages. Such decisions are, to a more or lesser extent, propagated along the aligning process so that the larger the number of sequences, the more likely it is to have poor alignments. An alternative for constructing an accurate alignment is the one explored, for instance, in Clustal Omega (Sievers et al., 2011). In this case, sets of sequences are split into small clusters, then, clusters are either split again or aligned depending on how many sequences they contain. Once all clusters are aligned, they are merged to produce the final alignment. Although it is not the perfect strategy to align homologous residues on top of each other, it allows the reconstruction of alignments with a minimal loss of information.

Once alignments are ready, it is possible to post-process them prior to any analyses. During my thesis, I have deeply worked in different alternatives to post-process alignments in order to increase the signal-to-noise ratio.

I have paid special attention to the significance of gaps in alignments. Ideally, gaps should represent only biological events of past insertions and deletions. However, in practice, gaps are generally introduced to maximize scoring functions during alignment reconstruction without a real biological meaning. Such decisions lead to gappy regions where making any inference is very difficult. It was in 1991 when Lake (Lake, 1991) realized, for the first time, the impact of gaps in downstream phylogenetic analyses so, therefore, a common practice is to remove gappy positions. Removing such positions became a manual task that does not allow reproducibility and, what is even worse, it is not feasible in a phylogenomics context (Castresana, 2000). As a result different programs have been published during the last decade to deal with this situation. It has been shown that the removal of gap-rich positions contributes significantly to improvements in the accuracy of downstream phylogenetic analyses (Talavera and Castresana, 2007; Capella-Gutiérrez et al., 2009). The biggest challenge for all these methods is the accurate identification of such noisy positions, since the removal of all positions containing gaps is often too aggressive.

It is generally accepted that most conflicting positions are gap-rich but it has recently been noticed that not only gaps are affecting downstream analyses. Other biases in alignments such as the presence of heterogeneous positions (Philippe and Roure, 2011) or the highly sensitivity of residues to the way they are being aligned (Landan and Graur, 2007) can have a strong impact in any posterior analysis, especially, in phylogenetic studies. Again, methods should be able to identify accurately such regions and, therefore, remove them before making any inference. In order to identify heterogeneous regions programs such as BMGE (Criscuolo and Gribaldo, 2010) uses entropy values for their identification and posterior removal. To identify residues pairs highly sensitive to the algorithm and/or parameters used to make the alignment, it is necessary to use more than one alignment in order to score each pair and remove those that are badly ranked (Capella-Gutiérrez et al., 2009). In my thesis, one of the main lines has been the development of methods able to automatically adjust trimming parameters in a phylogenomics framework with thousands of alignments of diverse nature, in terms of the number of sequences, number of residues and

divergence rates.

Despite the publication, in recent years, of different programs designed to accurately identify conflicting positions and their posterior removal. Further work is still needed to create a golden benchmark datasets, similar to BAliBASE (Thompson et al., 2005) or OXBENCH (Raghava et al., 2003) for testing multiple sequence alignments, containing simulated and real data. Such dataset would allow to settle the appropriate background for comparing different approximations and effectively measure the impact of individual trimming strategies in downstream analyses. The use of simulated data has been questioned because they do not appropriately reflect the complexity found in real situations. However, using them is the only way to precisely control different parameters and, thus evaluate the performance of diverse methods. The use of real data will allow to capture all complexity of authentic sequences and, therefore, see whether proposed methods are able to discriminate real signal from noise. Moreover, scoring systems to evaluate gaps placements are needed, Blackburne and Whelan (Blackburne and Whelan, 2012) have reviewed current scores for comparing residues pairing and proposed a new measure for gaps placement but we are far of explaining how well a gap is opened or kept in alignments.

In contrast to the efforts of identifying conflicting positions, specifically the gappy ones, for their posterior removal, it has been proposed that gaps carry phylogenetic information that is being systematically ignored in downstream analyses (Dessimoz and Gil, 2010). Although the study highlights that gaps should be consider as valid information, representing biological events, and, thus, used in evolutionary inferences, there is not a clear distinction between the signal carried by gaps due to true past evolutionary events and the signal reflecting already inferred relationships due the reconstruction of the guide tree. Without making such distinction using gaps for posterior analyses should be taken with caution because it is possible to erode the phylogenetic signal present in the residues just by adding random noise derived from algorithmic decisions. Furthermore, there is the real danger of biasing the results towards the error-prone guide-tree.

Regarding the phylogenetic inference, there are two main issues to deal with. The first of them is the incredible amount of computational power needed for phylogenetic inferences, especially, when more complex models are considered. To tackle this problem, there are some initiatives in the community aiming to implement phylogenetic methods in GPUs (Graphic Processor Unit), since these processors can execute parallel computations and, therefore, speed-up the general process. The main problem with these re-implementations is the complexity of the current algorithms since not all mathematical operations can be executed in such processors. The second issue relates to the phylogenetic inference itself, methods assume that residues evolved independently at the same rate but it has been shown that this is not true. To deal with different evolutionary rates, programs can have different categories but so far there are not general models which account for the interdepency of residues. If one day such models are available, they will surely need to take into account large computational requirements.

I have previously mentioned the need for having golden datasets for benchmarking purposes regarding to the alignment trimming step. There is an urgent need as well to design general benchmarks to evaluate different phylogenomics pipelines. In this direction, the quest for orthologs project (Gabaldón et al., 2009) looks like a reasonable strategy to evaluate the final result of strategies that differ in their conception: the reconstruction of phylogenetic trees for either all genes encoded by a genome, i.e. phylomes, or just for gene families. Each strategy have their own advantages and disadvantages. In similar terms, it would be desirable to account with initiatives to design appropriate benchmarks for alignments, in an evolutionary sense, since the correctness of alignments have a great impact in the general performance of any approximation.

In an interconnected world, it is equally important to generate accurate data as well as to have standards for sharing and tracking all the information. Hence, public databases face the enormous challenge of making compatible the results of sophisticate pipelines with other sources of information across several databases and versions. In return to this effort, the availability of a maximized amount of readily usable information will facilitate the generation of new knowledge in a variety of biological fields.

8.2 Applying phylogenomics methods to address relevant biological questions.

Having an accurate phylogenomics pipeline is not a purpose *per se*, but because it can be of great help for answering relevant biological questions. Besides working on several methodological aspects in the first part of my thesis, I applied this knowledge to the long-standing question of the phylogenetic position of the fungal group Microsporidia. Resolving this question poses many challenges, not only methodological but also conceptual since it has been very difficult to demonstrate that their position, the earlier branch in the fungal tree of life, is stable and not the result of the long branch attraction artifact. Long branch attraction is one of the most problematic biases in phylogenetics since it may create robust clustering of highly diverging groups despite their real phylogenetic positions.

In order to prove that our results shows the real placement of this fungal group, rather than being the product of an artifact, we have combined several phylogenomics approximations to discriminate real phylogenetic signal from noise. Although our result is strong enough and consistent across the different used strategies, the discussion about the real placement of this group will continue until more species around the base of fungi will be available. New species will allow to shorter the branches in the species tree, and, therefore, alleviate the possible long branch attraction artifact. Moreover, these species will provide, presumably, better insights about some intriguing microsporidian characteristics such as the lost of eukaryotic organelles such as the mitochondria, the peroxisomes, or the golgi apparatus. In a more biomedical aspect, elucidating the phylogenetic position of microsporidia is a first step to understand their evolution and, therefore, look for developing vaccines against these opportunistic human pathogens which seriously compromise the health of immuno-compromised patients (Keeling and Fast, 2002).

One of the biggest challenges in phylogenomics is the accurate reconstruction of trees establishing the evolutionary relationships among of a set of species. Although current methods allow to infer precise phylogenies from

a large enough sample of informative genes, the key issue is how to precisely select this sample when whole-genome approaches are not applicable. I have contributed to this problem by developing a new method for finding phylogenetically stable gene markers, which enables, for the first time, to exploit available genomic information to design marker sets that would work beyond the set of species used for their design. In fact, we have predicted sets of informative gene markers in order to elucidate the Cyanobacterial tree of life. Such predictions are being validated in the frame of a collaboration project with the molecular phylogenetics group in Kaiserslautern, Germany. One of the biggest problems until now has been the lack of gene markers that are sufficiently conserved across evolutionarily distant set of species for being identified at the same time that such markers are sufficiently different to accurately resolve deep phylogenies. Apart of identifying such genes, our method offers the possibility of increasing rapidly the taxonomic sampling of species tree using any of available sequencing technologies from the traditional PCRs to the new NGS techniques.

In a more practical manner, we have used phylogenomics techniques for a better understanding of different biological process. I have participated in several plant genome projects: *Chondrus crispus* -a multicellular red algae-, *Cucumis melo* -melon-, *Beta vulgaris* -sugar beet- and *Phaseolus vulgaris* -common bean-, where we have elucidated different mechanisms such as the evolution the resistance gene families, the transposon activity specific for each species, the possibility of Whole Genome Duplications events or how certain genes families -mitochondrial genes- can be traced to their bacterial origin through symbiosis. Given the benefits of a comprehensive evolutionary approach already on the annotation phase of a genome, I think these approaches will be more and more common in future genomics projects.

8.3 Final remarks.

To gain a full understanding of how species relate to each other and how different mechanism have been gained or lost across the evolution of species,

phylogenomics is highly depending on two complementary aspects. So, the advance of phylogenomics as science depends upon a full interaction of the development of methods for answering biological questions and the lessons that can be learnt from such answers to improve current methods.

I foresee an exciting future where we will be able to understand known and unknown mechanisms. Such understanding will ultimately lead us to a significant improvement of our lives in two ways: increase our quality of life *per se* and to a better preservation of our planet because as Theodosius Dobzhansky stated in 1973 ”Nothing in Biology Makes Sense Except in the Light of Evolution”.

9 Conclusions

- A newer pipeline for automated phylome reconstruction has been developed which improves in speed and accuracy over the preceding one. As compared to existing pipelines, this one is able to test more evolutionary models and efficiently exploit alignment variability to better select informative residues.

- A new version of PhylomeDB has been implemented with an improved database design that enables appropriate scaling without compromising performance. Currently, 11,5 million proteins of 870 species and over 2,6 million trees and alignments are stored, being phylomeDB the largest existing phylogenetic repository.

- Removing conflicting positions from multiple sequence alignments leads to improved signal-to-noise ratios and more accurate downstream phylogenetic analyses.

- Dynamic selection of trimming parameters is crucial in the context of large-scale phylogenetic analyses comprising large and heterogeneous datasets of sequence alignments. Several automated procedures that achieve efficient parameter selection have been implemented in trimAl.

- Gaps carry phylogenetic information, but current methods are unable to distinguish the true phylogenetic signal inherent to gaps from that inherited from the guide tree inferred to reconstruct the alignment.

- The use of consistent signals and diverse phylogenomics methods has enabled to overcome Long Branch Attraction artifacts and resolve, with strong support, the phylogenetic position of Microsporidia as the most basal group among sequenced fungi.

- A new method is proposed for the selection of stable phylogenetic markers that exploits information available in sequenced genomes. This method has been successfully applied to the selection of markers to chart the fungal and cyanobacterial trees of life.

10 Appendix: List of publications

1. **Capella-Gutiérrez S**, Marcet-Houben M, Gabaldón T. (2012). Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biology* 10:47.

2. **Capella-Gutiérrez S**, Silla-Martínez JM, Gabaldón T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972-3.

3. **Capella-Gutiérrez S**, Gabaldón T. Are gaps phylogenetically informative?: disentangling the signal carried by alignment gaps and guide trees. *submitted*.

4. **Capella-Gutiérrez S**, Kauff F, Gabaldón T. A phylogenomics approach for selecting robust sets of phylogenetic markers. *In preparation*.

5. **Capella-Gutiérrez S**, Gabaldón T. trimAl 1.4: Recent developments in automated multiple sequence alignment post-processing in large-scale phylogenetic analyses. *In preparation*.

6. Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González V, Hénaff E, Câmara F, Cozzuto L, Lowy E, Alioto T, **Capella-Gutiérrez S**, Blanca J, Cañizares J, Ziarsolo P, Gonzalez-Ibeas D, Rodríguez-Moreno L, Droege M, Du L, Alvarez-Tejado M, Lorente-Galdos B, Melé M, Yang L, Weng Y, Navarro A, Marques-Bonet T, Aranda MA, Nuez F, Picó B, Gabaldón T, Roma G, Guigó R, Casacuberta JM, Arús P, Puigdomenech, P. (2012) The genome of melon (*Cucumis melo L.*). *Proceedings of the National Academy of Sciences of the United States of America* 109:11872-11877.

7. Sánchez R, Serra F, Tárraga J, Medina I, Carbonell J, Pulido L, de María A, **Capella-Gutíerrez S**, Huerta-Cepas J, Gabaldón T, Dopazo J, Dopazo H. (2011). Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic acids research* 39 Suppl.2:W470-4.

8. Huerta-Cepas J, **Capella-Gutiérrez S**, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, Gabaldón T. (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments

and phylogeny-based orthology and paralogy predictions. *Nucleic acids research* 39:D556-60.

9. Gabaldón T, **Capella-Gutiérrez S**. (2010). Lack of phylogenetic support for a supposed actinobacterial origin of peroxisomes. *Gene* 465:61-5.

10. Collén J, Porcel B, Carre W, Ball S, Chaparro C, Tonon T, Barbeyron T, Michel G, Noel B, Valentin K, Elias M, Artiguenave F, Arun A, Aury JM, Barbosa-Neto JF, Bothwell JH, Bouget FY, Brillet L, Cabello-Hurtado F, **Capella-Gutíerrez S**, Charrier B, Cladiere L, Cock JM, Coelho SM, Colleoni C, Czjzek M, Da Silva C, Delage L, Denoeud F, Deschamps P, Dittami SM, Gabaldón T, Gachon CM, Groisillier A, Hervé C, Jabbari K, Katinka M, Kloareg B, Kowalczyk N, Labadie K, Leblanc C, Lopez PJ, McLachlan D, Meslet-Cladiere L, Moustafa A, Nehr Z, Collén PN, Panaud O, Partensky F, Poulain J, Rensing SA, Rousvoal S, Samson G, Symeonidi A, Weissenbach J, Zambounis A, Wincker P, Boyen C. The genome of the red alga Chondrus crispus unravels eukaryotic gene and genome evolution. *Submitted*.

References

Abascal, F., Posada, D., and Zardoya, R. (2007). MtArt: a new model of amino acid replacement for Arthropoda. *Molecular biology and evolution*, 24(1):1–5.

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)*, 21(9):2104–5.

Aguileta, G., Marthey, S., Chiapello, H., Lebrun, M.-H., Rodolphe, F., Fournier, E., Gendrault-Jacquemard, a., and Giraud, T. (2008). Assessing the performance of single-copy genes for recovering robust phylogenies. *Systematic biology*, 57(4):613–27.

Ajawatanawong, P., Atkinson, G. C., Watson-Haigh, N. S., Mackenzie, B., and Baldauf, S. L. (2012). SeqFIRE: a web application for automated extraction of indel regions and conserved blocks from protein multiple sequence alignments. *Nucleic acids research*, 40(Web Server issue):W340–7.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Altenhoff, A. M. and Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS computational biology*, 5(1):e1000262.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10.

Baldauf, S. L. (2003). Phylogeny for the faint of heart: a tutorial. *Trends in genetics : TIG*, 19(6):345–51.

Blackburne, B. P. and Whelan, S. (2012). Measuring the distance between multiple sequence alignments. *Bioinformatics (Oxford, England)*, 28(4):495–502.

Brown, D. and Sjölander, K. (2006). Functional classification using phylogenomic inference. *PLoS computational biology*, 2(6):e77.

Bruno, W. J. and Halpern, A. L. (1999). Topological bias and inconsistency of maximum likelihood using wrong models. *Molecular biology and evolution*, 16(4):564–6.

Capella-Gutiérrez, S., Marcet-Houben, M., and Gabaldón, T. (2012). Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC biology*, 10(1):47.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, 25(15):1972–3.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17:540–552.

Chakrabarti, S., Lanczycki, C. J., Panchenko, A. R., Przytycka, T. M., Thiessen, P. A., and Bryant, S. H. (2006). Refining multiple sequence alignments with conserved core regions. *Nucleic acids research*, 34(9):2598–606.

Chaudhary, R., Bansal, M. S., Wehe, A., Fernández-Baca, D., and Eulenstein, O. (2010). iGTP: a software package for large-scale gene tree parsimony analysis. *BMC bioinformatics*, 11:574.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, 311(5765):1283–7.

Corradi, N. and Keeling, P. J. (2009). Microsporidia: a journey through radical taxonomical revisions. *Fungal Biology Reviews*, 23(1-2):1–8.

Creevey, C. J. and McInerney, J. O. (2005). Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics (Oxford, England)*, 21(3):390–2.

Criscuolo, A. and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, 10:210.

Darwin, C. R. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray, London, 1st edition.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). *A model of evolutionary change in proteins*. National Biomedical Research Foundation, Washington, DC.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature reviews. Genetics*, 6(5):361–75.

Dessimoz, C. and Gil, M. (2010). Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome biology*, 11(4):R37.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research*, 15(2):330–40.

Dress, A. W. M., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S. J., and Stadler, P. F. (2008). Noisy: identification of problematic columns in multiple sequence alignments. *Algorithms for molecular biology : AMB*, 3(1):7.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids*. Press, Cambridge University, paperback edition.

Eddy, S. R. (1995). Multiple alignment using hidden Markov models. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 3:114–20.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10):e1002195.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5:113.

Edgar, R. C. (2010). Quality measures for protein alignment benchmarks. *Nucleic acids research*, pages 1–9.

Eirín-López, J. M., Rebordinos, L., Rooney, A. P., and Rozas, J. (2010). The birth-and-death evolution of multigene families revisited. *Genome dynamics*, 7:170–96.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome research*, 8(3):163–7.

Eisen, J. A. and Fraser, C. M. (2003). Phylogenomics: intersection of evolution and genomics. *Science (New York, N.Y.)*, 300(5626):1706–7.

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Syst Biol*, 27(4):401–410.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2):99–113.

Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science (New York, N.Y.)*, 155(3760):279–84.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., and Searle, S. M. J. (2012). Ensembl 2012. *Nucleic acids research*, 40(Database issue):D84–90.

Frickey, T. and Lupas, A. N. (2004). PhyloGenie: automated phylome generation and analysis. *Nucleic acids research*, 32(17):5231–8.

Gabaldón, T. (2008a). Comparative genomics-based prediction of protein function. *Methods in molecular biology (Clifton, N.J.)*, 439:387–401.

Gabaldón, T. (2008b). Large-scale assignment of orthology: back to phylogenetics? *Genome biology*, 9(10):235.

Gabaldón, T., Dessimoz, C., Huxley-Jones, J., Vilella, A. J., Sonnhammer, E. L., and Lewis, S. (2009). Joining forces in the quest for orthologs. *Genome biology*, 10(9):403.

Gabaldón, T., Marcet-Houben, M., and Huerta-Cepas, J. (2008). Reconstruction and analysis of large-scale phylogenetic data, challenges and opportunities. In Russe, A., editor, *Computational Biology: New Research*, pages 129–146. Nova Sciences Publishers, New York.

Galperin, M. Y. and Koonin, E. V. (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In silico biology*, 1(1):55–67.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution*, 14(7):685–95.

Giovannoni, S. J., Turner, S., Olsen, G. J., Barns, S., Lane, D. J., and Pace, N. R. (1988). Evolutionary relationships among cyanobacteria and green chloroplasts. *Journal of bacteriology*, 170(8):3584–92.

Gogarten, J. P. and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature reviews. Microbiology*, 3(9):679–87.

Golubchik, T., Wise, M. J., Easteal, S., and Jermiin, L. S. (2007). Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Molecular biology and evolution*, 24(11):2433–42.

Gonnet, G. H., Hallett, M. T., Korostensky, C., and Bernardin, L. (2000). Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics (Oxford, England)*, 16(2):101–3.

Goodstadt, L. and Ponting, C. P. (2006). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS computational biology*, 2(9):e133.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3):307–21.

Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174.

Hawksworth, D. L. (2001). The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycological Research*, 105(12):1422–1432.

Henson, B. J., Watson, L. E., and Barnum, S. R. (2004). The evolutionary history of nitrogen fixation, as assessed by NifD. *Journal of molecular evolution*, 58(4):390–9.

Hogeweg, P. and Hesper, B. (1984). The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of molecular evolution*, 20(2):175–86.

Holton, T. A. and Pisani, D. (2010). Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. *Genome biology and evolution*, 2:310–24.

Huerta-Cepas, J., Bueno, A., Dopazo, J., and Gabaldón, T. (2008). PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic acids research*, 36(Database issue):D491–6.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L. P., Denisov, I., Kormes, D., Marcet-Houben, M., and Gabaldón, T. (2011). PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic acids research*, 39(Database issue):D556–60.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldón, T. (2007). The human phylome. *Genome biology*, 8(6):R109.

Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python Environment for Tree Exploration. *BMC bioinformatics*, 11(1):24.

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J., Lumbsch, H. T., Rauhut, A., Reeb, V., Arnold, a. E., Amtoft, A., Stajich, J. E., Hosaka, K., Sung, G.-H., Johnson, D., O'Rourke, B., Crockett, M., Binder, M., Curtis, J. M., Slot, J. C., Wang, Z., Wilson, A. W., Schüssler, A., Longcore, J. E., O'Donnell, K., Mozley-Standridge, S., Porter, D., Letcher, P. M., Powell, M. J., Taylor, J. W., White, M. M., Griffith, G. W., Davies, D. R., Humber, R. a., Morton, J. B., Sugiyama, J., Rossman, A. Y., Rogers, J. D., Pfister, D. H., Hewitt, D., Hansen, K., Hambleton, S., Shoemaker, R. a., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Spotts, R. a., Serdani, M., Crous, P. W., Hughes, K. W., Matsuura, K., Langer, E., Langer, G., Untereiner, W. a., Lücking, R., Büdel, B., Geiser, D. M., Aptroot, A., Diederich, P., Schmitt, I., Schultz, M., Yahr, R., Hibbett, D. S., Lutzoni, F., McLaughlin, D. J., Spatafora, J. W., and Vilgalys, R. (2006). Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, 443(7113):818–22.

Jänicke, R. U., Sohn, D., and Schulze-Osthoff, K. (2008). The dark side of a tumor suppressor: anti-apoptotic p53. *Cell death and differentiation*, 15(6):959–76.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, III:21 – 132.

Katoh, K. and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics*, 9(4):286–98.

Kauff, F. and Büdel, B. (2011). Phylogeny of Cyanobacteria: An Overview. In Lüttge, U. E., Beyschlag, W., Büdel, B., Francis, D., and Lüttge, U., editors, *Progress in Botany*, volume 72 of *Progress in Botany*, chapter 4, pages 209–224. Springer Berlin Heidelberg, 72 edition.

Keeling, P. J. and Fast, N. M. (2002). Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annual review of microbiology*, 56:93–116.

Kemena, C. and Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics (Oxford, England)*, 25(19):2455–65.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome research*, 12(4):656–64.

Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B. M., Wägele, J. W., and Misof, B. (2010). Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in zoology*, 7:10.

Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L., and Tamura, K. (2011). Statistics and Truth in Phylogenomics. *Molecular Biology and Evolution*, 29(2):457–472.

Lake, J. A. (1991). The order of sequence alignment can bias the selection of tree topology. *Molecular biology and evolution*, 8(3):378–85.

Landan, G. and Graur, D. (2007). Heads or tails: a simple reliability check for multiple sequence alignments. *Molecular biology and evolution*, 24(6):1380–3.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21):2947–8.

Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics (Oxford, England)*, 25(17):2286–8.

Lassmann, T., Frings, O., and Sonnhammer, E. L. L. (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic acids research*, 37(3):858–65.

Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–20.

Lemaitre, C., Barré, A., Citti, C., Tardy, F., Thiaucourt, F., Sirand-Pugnet, P., and Thébault, P. (2011). A novel substitution matrix fitted to the compositional bias in Mollicutes improves the prediction of homologous relationships. *BMC bioinformatics*, 12:457.

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., and Linder, C. R. (2012). SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic biology*, 61(1):90–106.

Löytynoja, A. and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science (New York, N.Y.)*, 320(5883):1632–5.

Löytynoja, A. and Milinkovitch, M. C. (2001). SOAP, cleaning multiple alignments from unstable blocks. *Bioinformatics (Oxford, England)*, 17(6):573–4.

Marcet-Houben, M. and Gabaldón, T. (2009). The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *PloS one*, 4(2):e4357.

Murtagh, F. (1984). Complexities of hierarchic clustering algorithms: state of the art. *Computational Statistics Quarterly*, 1(2):101–113.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS computational biology*, 3(8):e123.

Notredame, C. and Higgins, D. G. (1996). SAGA: sequence alignment by genetic algorithm. *Nucleic acids research*, 24(8):1515–24.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17.

Notredame, C., Holm, L., and Higgins, D. G. (1998). COFFEE: an objective function for multiple sequence alignments. *Bioinformatics (Oxford, England)*, 14(5):407–22.

Page, R. D. (1998). GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics (Oxford, England)*, 14(9):819–20.

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., and Pupko, T. (2010). GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic acids research*, 38(Web Server issue):W23–8.

Philippe, H. and Roure, B. (2011). Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC biology*, 9:91.

Posada, D. (2003). Selecting models of evolution. *The phylogenetic handbook*, pages 256–282.

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

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)*, 19(12):1572–4.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–25.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7.

Schoch, C. L., Sung, G.-H., López-Giráldez, F., Townsend, J. P., Miadlikowska, J., Hofstetter, V., Robbertse, B., Matheny, P. B., Kauff, F., Wang, Z., Gueidan, C., Andrie, R. M., Trippe, K., Ciufetti, L. M., Wynns, A., Fraker, E., Hodkinson, B. P., Bonito, G., Groenewald, J. Z., Arzanlou, M., de Hoog, G. S., Crous, P. W., Hewitt, D., Pfister, D. H., Peterson, K., Gryzenhout, M., Wingfield, M. J., Aptroot, A., Suh, S.-O., Blackwell, M., Hillis, D. M., Griffith, G. W., Castlebury, L. A., Rossman, A. Y., Lumbsch, H. T., Lücking, R., Büdel, B., Rauhut, A., Diederich, P., Ertz, D., Geiser, D. M., Hosaka, K., Inderbitzin, P., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Mostert, L., O'Donnell, K., Sipman, H., Rogers, J. D., Shoemaker, R. A., Sugiyama, J., Summerbell, R. C., Untereiner, W., Johnston, P. R., Stenroos, S., Zuccaro, A., Dyer, P. S., Crittenden, P. D., Cole, M. S., Hansen, K., Trappe, J. M., Yahr, R., Lutzoni, F., and Spatafora, J. W. (2009). The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systematic biology*, 58(2):224–39.

Schopf, J. (2002). The Fossil Record: Tracing the Roots of the Cyanobacterial Lineage. In Whitton, B. A. and Potts, M., editors, *The Ecology of Cyanobacteria*, chapter 2, pages 13–35. Kluwer Academic Publishers, Dordrecht.

Seo, P.-S. and Yokota, A. (2003). The phylogenetic relationships of cyanobacteria inferred from 16S rRNA, gyrB, rpoC1 and rpoD1 gene sequences. *The Journal of general and applied microbiology*, 49(3):191–203.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–45.

Shimodaira, H. and Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17(12):1246.

Sicheritz-Pontén, T. and Andersson, S. G. (2001). A phylogenomic approach to microbial evolution. *Nucleic acids research*, 29(2):545–52.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7:539.

Snel, B., Huynen, M. a., and Dutilh, B. E. (2005). Genome trees and the nature of genome evolution. *Annual review of microbiology*, 59:191–209.

Sokal, R. and Michener, C. (1958). A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.*, 38:1409 – 1438.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, 22(21):2688–90.

Stoye, J., Evers, D., and Meyer, F. (1998). Rose: generating sequence families. *Bioinformatics (Oxford, England)*, 14(2):157–63.

Subramanian, A. R., Kaufmann, M., and Morgenstern, B. (2008). DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for molecular biology : AMB*, 3(6):6.

Swingley, W., Blankenship, R., and Raymond, J. (2008). Insights into cyanobacterial evolution from comparative genomics. *The cyanobacteria: molecular biology, genomics, and evolution*, page 2.

Talavera, G. and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, 56(4):564–77.

Thompson, J. D. (2003). RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, 19(9):1155–1161.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1):127–36.

Thompson, J. D., Plewniak, F., and Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic acids research*, 27(13):2682–90.

Vilella, A. J., Severin, J., Ureta-vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, 19(2):327–35.

Walker, D. M., Castlebury, L. A., Rossman, A. Y., and White, J. F. (2012). New molecular markers for fungal phylogenetics: two genes for species-level systematics in the Sordariomycetes (Ascomycota). *Molecular phylogenetics and evolution*, 64(3):500–12.

Wallace, I. M., O'Sullivan, O., and Higgins, D. G. (2005). Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics (Oxford, England)*, 21(8):1408–14.

Wallace, I. M., O'Sullivan, O., Higgins, D. G., and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic acids research*, 34(6):1692–9.

Wang, H., Xu, Z., Gao, L., and Hao, B. (2009). A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC evolutionary biology*, 9:195.

Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of computational biology : a journal of computational molecular cell biology*, 1(4):337–48.

Whelan, N. V. (2011). Species tree inference in the age of genomics. *Trends in Evolutionary Biology*, 3(1).

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5088–90.

Wolf, Y., Rogozin, I., and Grishin, N. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary ...*, 1:8.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science (New York, N.Y.)*, 319(5862):473–6.

Wu, M., Chatterji, S., and Eisen, J. A. (2012). Accounting for alignment uncertainty in phylogenomics. *PloS one*, 7(1):e30288.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–91.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of theoretical biology*, 8(2):357–66.