Improving the accuracy and the efficiency of multiple sequence alignment methods

Carsten Kemena

TESI DOCTORAL UPF / ANY 2012

DIRECTOR DE LA TESI

Cédric Notredame
Bioinformatics and Genomics
Centre for Genomic Regulation

Department of Experimental and Health Sciences

Acknowledgement

I would like to thank my supervisor Cédric Notredame for all the help I received over the last four years, Jean-François Taly for reading the first draft and giving a lot of helpful suggestions for this thesis as well as Ionas Erb for the proof reading and Meritxell Oliva Pavia for helping with the Spanish translation. Further thanks go to Toni Gabaldon, Mar Alba and Matthieu Louis who accompanied me on my way as my Thesis Committee. I further thank all my co-authors, especially Fyodor Kondrashov. Moreover, I would like to thank Javier Herrero for allowing me to visit his group for three months. Of course I also thank my current and former colleagues with which I have been working, Romina Garrido Enamorado for always organizing everything so perfectly and our system administrators for keeping the system working. Finally I would like to thank my family and friends for supporting me the whole time.

Abstract

Sequence alignment is one of the basic methods to compare biological sequences and the cornerstone of a wide range of different analyses. Due to this privileged position at the beginning of many studies its accuracy is of great importance, in fact, each result based on an alignment is depending on the alignment quality. This has been confirmed in several recent papers investigating the effect of alignment methods on phylogenetic reconstruction and the estimation of positive selection. In this thesis, I present several projects dedicated to the problem of developing more accurate multiple sequence alignments and how to evaluate them. I addressed the problem of structural protein alignment evaluation, the accurate structural alignment of RNA sequences and the alignment of large sequence data sets.

Resumen

El alineamiento es uno de los métodos básicos en la comparación de secuencias biológicas, y a menudo el primer paso en análisis posteriores. Por su posición privilegiada al principio de muchos estudios, la calidad del alineamiento es de gran importancia, de hecho cada resultado basado en un alineamiento depende en gran medida de la calidad de éste. Este hecho se ha confirmado en diversos artículos recientes, en los cuales se ha investigado los efectos de la elección del método de alineamiento en la reconstrucción filogenética y la estimación de la selección positiva. En esta tesis, presento varios proyectos enfocados en la implementación de mejoras tanto en los métodos de alineamiento múltiple de secuencias como en la evaluación de estos. Concretamente, he tratado problemas como la evaluación de alineamientos estructurales de proteínas, la construcción de alineamientos estructurales y precisos de ARN y también el alineamiento de grandes conjuntos de secuencias.

Preface

Medical and biological research has changed a lot in the last 10-20 years. With the possibility of sequencing whole genomes new areas of research have been opened, allowing a more complete understanding of the molecular basis of life. Now it is possible to trace genetic diseases back to their location in the genome, it is feasible to perform large-scale analyses of gene expression, and the availability of different sequencing methods allows many other lines of investigation. These approaches produce huge amounts of data which need to be analyzed. Hence, computational approaches are needed an idea which is still relatively new to biology. When I started studying few people had an idea about bioinformatics and its utility. Today, bioinformatics is a well-established field of research. In this thesis I contribute to the improvement of the first method in sequence analysis, the multiple sequence alignment, which, due to its difficulty, still provides challenges.

Contents

1 Introduction

The following text in the extended version of a review (Kemena and Notredame, 2009), written at the beginning of my thesis and further adapted so as to include more recent developments in the field.

During the time of my thesis I have focused my interest on the development of methods for the comparison of biological sequences. In this introduction I will describe the biological and algorithmic basis of my work as well as the utility of the tools I developed. The work presented relies on the theory of evolution that was first published in the book "The origin of species" by Charles Darwin (1859). The core of this theory is that all living organisms have developed over millions of years starting from a single common ancestor and slowly evolved into different species. Darwin got this idea during a stay on the Galápagos Islands, where he noticed that finches living on this group of islands have different characteristics on different islands. From this observation he concluded that the different finches adapted themselves to the specific conditions of each island. He called this process of adaptation natural selection, a process in which only those organisms survive that are best adapted to their surrounding environment. Herbert Spencer phrased it in the famous words "Survival of the fittest". Of course at that time, the deoxyribonucleic acid (DNA) was still an unknown molecule, as was its involvement in the mechanisms of evolution.

Today it is known that the cells of each living organism contain large molecules of DNA, the genome. It contains the blueprint of the organism, coding for functional complexes like genes or regulatory motifs necessary for the production and the regulation of the cell machinery. When an organism reproduces, a copy of the DNA is passed on to its offspring. The copying mechanism is not perfect, mutations can occur, which together with other mechanisms like recombination and horizontal gene transfer can alter the DNA. These changes may be beneficial, neutral or detrimental. Under certain conditions, these

mutations can lead to the emergence of new species. According to the theory of speciation, a population can develop into two separate species when two groups of the same population stop to interbreed with each other, for example when a population is divided into two subpopulations in distinct areas due to some geographic event. From generation to generation these populations diverge more and more until they cannot produce offspring with each other anymore, a new species has emerged. All species and all organism are therefore related with each other, and the less time has passed since the last common ancestor, the closer two species or organisms are related. These relationships are reflected on the molecular level and thus comparisons between biological sequences, being it DNA, RNA or protein sequences allow to infer biological knowledge. A good way to perform this comparison and to detect specific areas of similarity and dissimilarity is the construction of a sequence alignment.

They are of importance for an ever increasing number of biological modeling methods. Chapter 1.2 describes the most important of these applications. While the vast majority of published applications are based on protein sequence alignments, recent biological discoveries coupled with the massive delivery of functional, structural and genomic data are rapidly expanding the potential scope of alignment methods. In order to make the best of the available data, sequence aligners will have to evolve and become able to deal with a very large number of sequences because the known quantity of sequences in public sequence databases like the ones from the European Molecular Biology Laboratory (EMBL) or its American counterpart, the National Center for Biotechnology Information (NCBI) (Karsch-Mizrachi et al., 2012) is growing exponentially (Figure 1.1). Merely aligning all the known orthologues of a given gene will soon require aligning several thousand sequences, and the massive re-sequencing effort currently underway (Siva, 2008) could even mean that within a few decades, multiple comparison methods may be required to align billions of closely related sequences. Besides this alignment programs will need to be able to integrate highly heterogeneous information types such as evolutionary, structural and functional data.

During my thesis I engaged in several projects concerned with some of these problems. The projects were not limited to the problem of computing highly accurate alignments but also included the estimation of their accuracy. The following sections contain a description of the alignment problem, as well as an introduction to the current state-of-the-art approaches to solve it under different contexts. The next chapters describe the results of

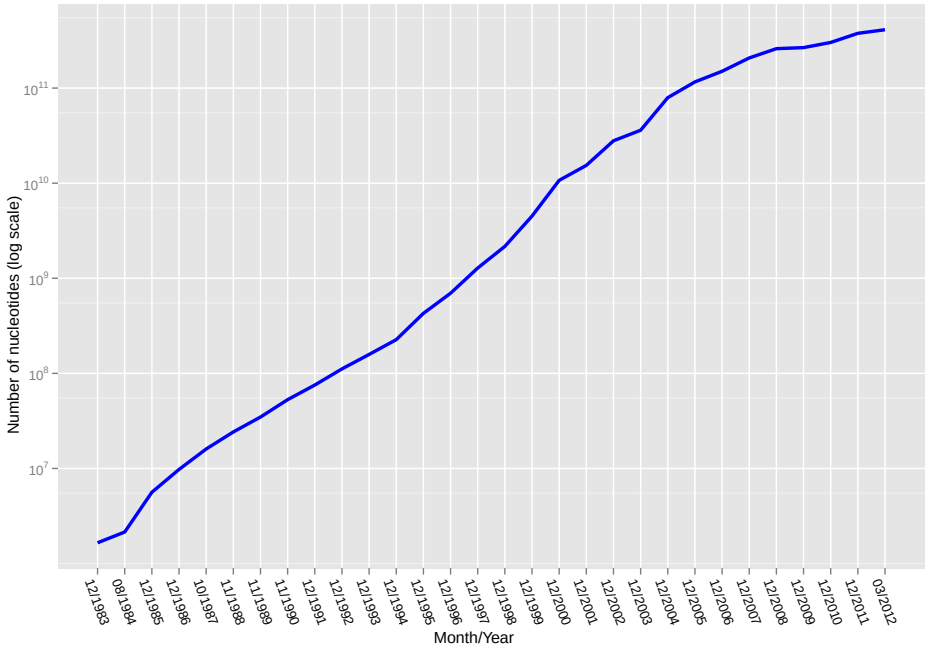


Figure 1.1: Number of nucleotides in the EBML-bank in the latest release of each year.

several of my projects. Chapter 2 deals with the problem of evaluating protein alignments using a single experimental structure. The subsequent chapters present new algorithms I developed to increase the accuracy of an alignment. I especially addressed the problem of multiple structural RNA alignments (Chapter 3) and the alignment of large protein datasets (Chapter 4). The last chapter contains a discussion of the results presented here.

1.1 What is a sequence alignment?

Starting from a set of sequences, an alignment of these can be constructed by inserting gap characters into the sequences without changing the order of characters in them. An alignment has two technical properties which need to be fulfilled: (i) all gap-extended sequences have the same length and (ii) no column consists of gap characters only. An example of an alignment can be seen in Figure 1.2. It has been produced using the default mode of T-Coffee (Notredame et al., 2000). The coloring schema represents the consistency of the multiple sequence alignment (MSA) with the pairwise alignments.

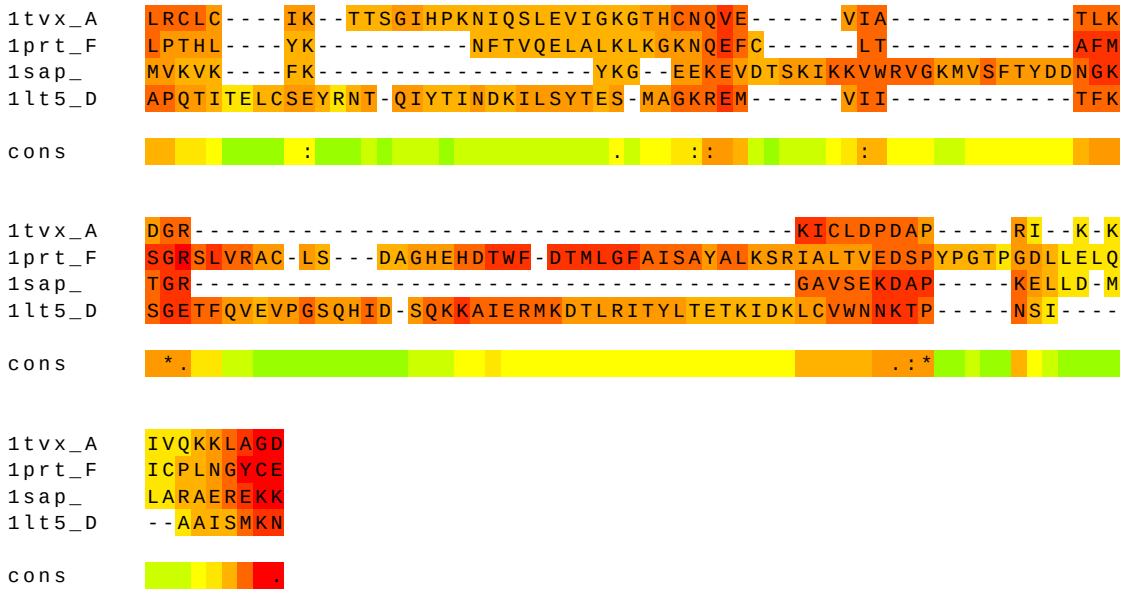


Figure 1.2: A multiple sequence alignment produced by T-Coffee. The coloring of the amino acids represents consistency with pairwise alignments. Green nucleotides correspond to low consistency, red nucleotides to high consistency.

Biologically, an alignment is a method to detect the evolutionary or functional relationship between two or more biological sequences. It organizes data so that similar sequence features are aligned together. A feature can be any relevant biological information: structure, function or homology to the common ancestor. The goal is either to reveal patterns that may be shared by many sequences, or identify modifications that may explain functional and phenotypic variability. The features one is interested in and the way in which these features are described ultimately define the correct alignment, and in theory, given a set of sequences, each feature type may define a distinct optimal alignment. For instance, a structurally correct alignment is an alignment where aligned residues play similar role in the 3D structure. Given a set of distantly related sequences, there may be more than one alignment equally optimal from a structural point of view. An alternative to structural conservation is homology (meant in a phylogenetic sense). In that case, the alignment of two residues is a statement that these two residues share a similar relation to their closest common ancestor. Aside from evolutionary inertia, there is no well defined reason why a structure and a homology based alignment of the same sequences should be identical. Likewise, in a functionally correct alignment, residues having the same function need to

be aligned, even if their similarity results from convergent evolution. Overall, a multiple sequence alignment is simply a way of confronting and organizing the specific data one is interested in.

These relationships can be either detected between two sequences in which case one speaks of pairwise alignments or in case of three or more sequences of multiple sequence alignments. In either case, the homology does not need to span the whole length of the sequences in which case the computation of a global alignments which spans the whole sequences, would result in inaccurate alignments. In this case preferably a local alignment method is used which is able to identify homologous segments inside the sequences and only align these. This is useful for example when aligning genes to a genome or when aligning proteins with each other. Proteins are often composed of several different domains, which are often associated with a specific function and have the capacity to fold independently from the rest of the protein. These domains are often reused in different proteins but the composition of each protein varies. Thus, when aligning proteins sharing only a subset of common domains, a local alignment might be the method of choice. A different classification of alignments can be achieved when classifying alignments according to the used information type. For example one can distinguish sequence alignments, profile alignments and structural alignments. The most basic alignment type is the simple sequence alignment. It can be used for all kinds of sequences (DNA, RNA and proteins) but as it uses only the information incorporated in the sequence, the accuracy of the resulting alignment greatly depends on the sequence identity. Sequence identity denotes the number of identical residue pairings compared to the overall number of pairs of two aligned sequences. Generally, the higher the identity the easier it is to derive an accurate alignment. Depending on the sequence type, the limit for achieving meaningful alignments varies. The minimum sequence identity for protein sequences is considered to be about 30% (Rost, 1999), while in contrast, for RNA alignments the limit is assumed to be around 60% identity (Abraham et al., 2008; Capriotti and Marti-Renom, 2010). The areas around these thresholds are called twilight zone because the signal derived from sequence identity gets disturbed. The identity threshold for DNA and RNA sequences is higher because the alphabet of nucleotide sequences is much smaller than for amino acid sequences (4 versus 20), hence it is less informative. The meta-alphabet of proteins helps to keep accurate track of sequence homology over time as often several codons can code for the same amino acid. For many amino acid the third base of the codon, the wobble base, can change without changing the amino acid. Furthermore, coding sequences are

often more conserved than non-coding sequences due to a higher evolutionary pressure to keep functionality. Moreover, in non-coding RNA the secondary structure plays an important role for its functionality and is often conserved while the nucleotides involved change more frequently. When reaching the twilight zone, additional information, which can be acquired from different sources, is needed to construct meaningful alignments, for example structural information for RNA or protein sequences.

The amount of data that could be integrated when building an MSA is rising by the day. It includes new sequences coming from large scale genome sequencing, with a density of information that will make it more and more possible to reconstruct evolutionary correct alignments (Frazer et al., 2007). Other high throughput based projects are delivering functional data in the form of transcript structure (The ENCODE Project Consortium, 2007) and structural data is following a similar trend thanks to coordinated efforts like targetDB (Chandonia et al., 2006). Another ongoing trend is the delivery of large-scale functional data, resulting from the use of robotic techniques. These make it possible to gather large amounts of functional information associated with homologous sequences (Fabian et al., 2005). This data is usually applied to Quantitative Structure and Activity Relationships (QSAR) analysis, but it could just as well be used when comparing protein sequences. Finally, the massive use of ChIp-Chip data makes it possible to reveal important protein/DNA interaction, thus allowing the enrichment of genomic data with functional data, an extra layer of information that could certainly be incorporated in sequence alignment strategies such as the ones described in the following chapters.

These trends have not gone unnoticed and over the last years, regular efforts have been made at developing and improving multiple sequence alignments methods so that they could take advantage of newly available data. Three areas have been actively explored: (i) accuracy improvement, achieved through the use of consistency based methods (Notredame et al., 2000; Do et al., 2005) (ii) an expansion of MSA methods scope, thanks to the development of template-based approaches (Armougom et al., 2006b; Pei and Grishin, 2007; Pei et al., 2008; Wallace et al., 2006; Wilm et al., 2008), a natural development of consistency based methods that makes it possible to efficiently integrate alternative methods and alternative types of data. (iii) large-scale alignments (Edgar, 2004a; Lassmann and Sonnhammer, 2005b; Katoh and Toh, 2008; Sievers et al., 2011). Most of the MSA methods currently available have been described and compared at length in several very complete reviews (Wallace et al., 2005a; Edgar and Batzoglou, 2006; Notredame, 2007;

Pei, 2008).

1.2 The purpose of calculating alignments

Alignments are calculated to discover similarities and dissimilarities between sequences to obtain information allowing to infer relationships between sequences, or, on a lower level, between characters of these sequences. This chapter will cover the most important applications of how this information can be used, giving examples from various areas, beginning with evolutionary biology.

One of the objectives of evolutionary biology is to determine the relationships between species or genes. A common way to infer these relationships is to identify mutations in a set of sequences and construct a phylogenetic tree using this information. An accurate method to detect these changes is via the construction of an MSA. Different methods exist to reconstruct a phylogentic tree from an alignment. The first methods developed turn an alignment into a matrix of pairwise distances and use a clustering algorithms to construct the tree (e.g. UPGMA (Sokal and Michener, 1958) or Neighbour Joining (Saitou and Nei, 1987)). These distance based methods regard the sequences as a whole and represent all differences between two sequences in a single value. The maximum parsimony method on the other hand belongs to the character-based methods, methods which treat residues of a sequence individually. For instance, the Fitch algorithm (Fitch, 1971) constructs a tree minimizing the overall number of changes necessary to explain the differences in the sequences as denoted by the alignment. More accurate character-based methods try to model the evolutionary process using probabilities. To this class of methods belong maximum likelihood estimation methods (e.g. RAxML (Stamatakis, 2006) or PhyML (Guindon et al., 2010)) and Bayesian approaches (e.g. MRBAYES (Huelsenbeck and Ronquist, 2001)). The increased accuracy of these methods is accompanied by longer running times when compared to distance or maximum parsimony methods. Some nodes in the reconstructed tree can have higher support from the alignment than others, information which can help to estimate the accuracy of a specific node. A widely used method for this purpose is the bootstrap value (Felsenstein, 1985), which reports the fraction of trees computed from sampled alignment columns supporting a given topology.

Another example from evolutionary studies is the inference of homology and orthology

between genes. One of the most commonly used methods to infer homology is BLAST (Basic local alignment search tool) (Altschul et al., 1990). A widely used ad-hoc method to determine orthology between sequences is the reciprocal best hit method in which two genes are considered to be ortholog when they find each other as the best hit when one is blasted against the genome of the other and vice versa. The establishment of these relationships allow for large-scale transfer of functional annotations from one genome to another.

However, annotating whole genomes still remains difficult and many functional elements remain unidentified. Using multiple genome alignments can help as they allow the identification of positions under purifying, positive and neutral selection. Widely used programs for the identification of elements under purifying selections are Gerp/Gerp++ (Cooper et al., 2005; Davydov et al., 2010) and PhastCons (Siepel et al., 2005). Gerp measures the difference between expected and observed rate of mutation to detect conserved sites in an alignment. PhastCons models conserved elements using a phylogenetic HMM. This HMM consists of two states, one for conserved regions and one for unconserved regions. Not only can regions under purifying selection be of interest but regions under positive selection as well as they may indicate newly functional elements. An alignment based program to discover these segments has been developed by Massingham and Goldman (2005). Whereas purifying and positive selection is detected by comparing sequences from different species with each other, alignments of sequences from the same species can help to detect single nucleotide polymorphisms, variations in a sequence occurring in a subgroup of a population. The identification of SNPs is important because certain SNPs have been associated with higher risks for certain diseases like diabetes (Qi et al., 2009) or Huntington's disease (Weydt et al., 2009). Thus identifying SNPs causing an illness may increase the probability of understanding it and may facilitate the search for a treatment. The SNPsFinder (Song et al., 2005) program, for example, is able to find SNPs inside a genome alignment.

In the examples above, the main source of information is the primary sequence but often sequences have a structural component which is of great importance for its correct functionality as, for example, in proteins or non-coding RNAs. Non-coding RNA is a class of RNA with a rapidly increasing number of known transcripts (Gardner et al., 2011; Djebali et al., 2012; Harrow et al., 2012). These RNAs have in common that inside an RNA family the secondary structure is often maintained while the primary sequence is

evolving rapidly. Knowing the structure of an RNA may help understanding the functional properties of it. Several approaches to predict the secondary structure from a single sequence have been proposed. Most of them are based on thermodynamics and predict the secondary structure by minimizing the free energy as implemented in Mfold (Zuker, 1989). Still, predicting the correct secondary structure from a single sequence is very difficult. As a consequence, different strategies to estimate the common RNA structure from a set of aligned homologus sequences have been proposed. The two most prevalent strategies are the prediction using stochastic context-free grammars (SCFGs) and the prediction using hybrid methods, methods which are a combination of thermodynamics and compensated mutation approaches. An often used representative of the first group is the PFOLD program (Knudsen and Hein, 2003) which besides SCFGs uses additionally an evolutionary model. RNAalifold (Hofacker et al., 2002) on the other hand calculates the linear combination of base pairing energies and covariation of base pairs found in the alignment. Compensated mutations cannot only be used to predict a secondary structure but can be used as well to identify so far unknown RNA genes. Rivas and Eddy (2001) for example applied three different hidden Markov models (HMMs) on pairwise alignments to distinguish which segment of the sequences can be attributed to RNA, protein coding and the NULL model. The protein coding model checks for mutations in the third base of the DNA, whereas the RNA model detects compensated mutations inside a secondary structure. All these methods depend on accurate RNA alignments, a difficult problem I addressed in Chapter 3 proposing a new algorithm to produce multiple RNA structure alignments.

Similar to RNA, the structure of proteins is important and usually conserved during evolution, to keep its function. Thus, knowing the structure of a protein can help in its functional analysis. Experimental determination of a protein structure is often possible but is still too expensive for a large-scale application. Hence, computational approaches on different levels have been developed. An important strategy is to predict contacts in protein sequences. A contact is the bond formed between two amino acids, due to their physicochemical properties. These contacts introduce evolutionary constraints onto the amino acids involved because mutations in only one of them need to be compensated by a mutation in the corresponding amino acid. This can be used to predict contacts by detecting compensated mutations using MSAs as, for example, described by Goebel et al. (1994). Similar approches have been developed to detect interactions between proteins (e.g. (Pazos and Valencia, 2002)) or using Bayesian networks (Burger and van Nimwe-

gen, 2008). On the level of secondary structure prediction, Cuff and Barton (2000) used alignments in combination with neural networks. Many algorithms were proposed to predict the tertiary structure of a sequence including 3D-Jigsaw (Bates et al., 2001), SWISS-MODEL (Schwede et al., 2003) and Modeller (Eswar et al., 2007). These commonly predict a three-dimensional structure of a given protein sequence using an alignment between the target sequence and one or more template sequences with a known structure as basis. While these methods need accurate alignments to predict a structure, one can also use a structure to evaluate the accuracy of an alignment. An example is the CAO score (Lin et al., 2003) or the STRIKE score proposed in Chapter 2.

1.3 Basic approaches of sequence alignment computation

Multiple sequence alignment computation stands at a cross-road between computation and biology. The computational issue is as complex to solve as it is straightforward to describe: given any sensible biological criterion, the computation of an exact MSA is NP-Complete and therefore impossible for all but unrealistically small datasets (Wang and Jiang, 1994). MSA computation therefore depends on approximate algorithms or heuristics and it is worth mentioning that almost every conceivable optimization technique has been adapted into a heuristic multiple sequence aligner. Over the last 30 years, more than a 100 multiple sequence alignment methods have been published, based on all kind of heuristics, including simulated annealing (Abhiman et al., 2006), genetic algorithms (Gondro and Kinghorn, 2007; Notredame and Higgins, 1996), Tabu search (Riaz et al., 2005), branch and bound algorithms (Reinert et al., 1997), Hidden Markov Modeling (Eddy, 1995) and countless agglomerative approaches including the progressive alignment algorithm (Hogeweg and Hesper, 1984), by far the most widely used nowadays. The biological issue surrounding MSAs is even more complex: given a set of sequences, we do not know how to estimate similarity in a way that will guaranty the biological correctness of an alignment, whether this correctness is defined in evolutionary, structural or functional terms. In fact, one could argue that being able to compare the biological features coded by a DNA sequence implies having solved most of the ab-initio problems associated with genetic information interpretation, including protein structure prediction. These problems are not solved and in practice multiple alignments are estimated by maximizing identity, in the hope that this simplistic criterion will be sufficiently informative

to yield models usable for most type of biological inference.

The objective function thus maximized is usually defined with a substitution matrix and a gap penalty scheme, the first one modeling mutations or the lack of them by assigning scores for aligning two residues with each other using a scoring scheme and a second one modeling insertions and deletions by assigning costs to a gap, where a gap is a consecutive row of gap characters. Most aligners do not distinguish between insertions and deletions as it is usually unknown which of the two is being handled and consequently assign the same costs to both. An exception from this general rule is PRANK (Löytynoja and Goldman, 2005). For nucleotide alignments the scoring schema is usually kept very simple by using an identity matrix assigning a single score for all kinds of matches and another score for all kinds of mismatches. Exceptions from this rule are for example BLASTZ (Schwartz et al., 2003), which uses two different values to score mismatches, and Pro-Coffee (Erb et al., 2012), which uses a scoring matrix to score matches of neighboring nucleotides. In the case of amino acids a more elaborate scoring schema is used to reflect the more complex relationships between amino acids. Each match/mismatch of two amino acids has its own score reflecting the similarity or dissimilarity between the physicochemical properties of the amino acids involved. The most common scoring matrices are the BLOcks SUbstitution Matrix (BLOSUM) (Henikoff and Henikoff, 1992) and the Point Accepted Mutations matrices (PAM) (Dayhoff et al., 1978). Different versions of these matrices are available to reflect the evolutionary distance between the sequences to align, as the probabilities to see mismatches increases with time of evolution. The BLOSUM matrix has been computed on blocks of local alignments and represents the log-odds score of observed frequencies of mutation versus the expected rate. The PAM matrices are computed very similarly on the mutation rate. These matrices are context independent, they assume that the pairing of a pair of residues is independent of any other residues in the sequence. This simplification often does not reflect biological reality for example in RNA and proteins the structure can introduce dependencies between different amino acids, but ignoring it allows for a faster and simpler computation.

Several gap cost functions have been developed to model insertions and deletions (gaps in general) in sequences. The simplest schema treats all gap characters the same way, each gap character is given a score which is then added to the objective function. This linear model is called homologous gap costs. Although the biological truth of how insertions and deletions appear is not known, this model is generally considered to be too simplistic,

as it assumes the same costs for each gap character. A better model would incorporate the assumption that the event of introducing a gap is more important than the length of it, as even short gaps can lead to frameshifts in proteins. This idea is modeled in the so called affine gap costs, which does not only consist of costs given to each character (gap extension costs), but gives an extra cost for each opening of a gap, the gap opening costs. As a consequence, when compared to homologous gaps, it is less likely to open a gap but more probable to produce longer gaps which should represent the biological assumptions better, as the probability of a frameshift drops. In endgap free alignments, gaps at the end and the beginning of an alignment are not penalized to avoid stretching of sequences of different lengths. In Probcons (Do et al., 2005), a biphasic gap cost scheme is proposed, scoring short and long gaps differently. Short gaps are more probable of being inserted but have higher extension costs compared to large gaps which have a low probability of being inserted.

Independent of scoring matrix and gap cost schema most aligners, use a dynamic programming approach to maximize the objective function. Needleman and Wunsch (1970) were the first to adopt this approach from computer science to the pairwise alignment problem and proposed an algorithm that is able to produce the alignment in $O(n^2)$ time and space using homologous gap costs. In dynamic programming, intermediate results are calculated once and then stored, so that they can be reused in later steps. In the case of alignment computation these intermediate results are alignments of prefixes which are reused to compute the next elongation. This kind of approach guarantees the discovery of the optimal solution given an objective function with context independent matching scores. Although methods became more sophisticated over time, the basic idea is still the same and is used in most alignment programs in one way or another. The importance of its development can also be seen in the many extensions and variations which have been developed for this algorithm. With slight changes in the algorithm it is possible to produce local alignments (Smith and Waterman, 1981) or alignments in linear space (Myers and Miller, 1988). A version which is able to incorporate affine gapcosts and keeping time and space requirements quadratic was developed by Gotoh (Gotoh, 1982). Even when including secondary structure predictions while computing alignments the basic algorithm is very similar to the original one (Sankoff, 1985). However, due to the incorporation of a non-independent scoring schema the running time strongly increases.

Recently probabilistic techniques have been implemented that rely on pair-Hidden Markov

Models (HMMs) and take advantage of a better defined statistical framework (Durbin et al., 1998). While DP and HMM based approaches are mostly interchangeable, the latter ones make it easier to explore the parameter space using off-the-shelves statistical tools such as Baum-Welch and Viterbi training. HMM modeling also offers easy access to a wider range of scoring possibilities, thanks to posterior decoding, thus making it possible to assess complex alignment scoring schemes. For instance, a significant share of the improvements measured in the ProbCons (Do et al., 2005) algorithm over other consistency based packages seems to result from the use of a bi-phasic penalty scheme (Table 1.1), predefined as a finite state automata (FSA) and parameterized by applying the Baum-Welsch algorithm on BALiBASE.

Sequence identity is only a crude substitute to biological homology, and in practice, it has often been argued that structurally correct alignments are those more likely to be useful for further biological modeling. Similarity based alignment methods have therefore been carefully tuned in order to produce structurally correct alignments. This tuning (or validation) has relied on the systematic usage of structure based reference multiple sequence alignments. This procedure has now been in use for more than a decade and has been a major shaping force on this entire field of research. We will now review the most common validation procedures with their associated databases.

1.4 Accuracy Estimation using reference alignments

The first systematic validation of a multiple sequence alignment using reference alignments was carried out by McClure (1994). McClure was evaluating her alignments by assessing the correct alignment of predefined functional motifs. Shortly after, Notredame and Higgins made the first attempt to systematically use structure based alignments while evaluating the biological accuracy of the SAGA package (Notredame and Higgins, 1996). The validation was carried out on a relatively small dataset named 3d-ali (Pascarella et al., 1996). A few years later Thompson developed a purpose built dataset named BALiBASE I (Thompson et al., 1999). The main specificity of BALiBASE was to address a wide range of different issues related to multiple sequence alignments. This included the alignment of distantly related homologues, the ability of alternative methods to deal with long insertions/deletions and their ability to properly integrate outliers. Its main weakness was the questionable accuracy of some alignments and the relatively small size (82) of the dataset.

Most of these issues have been addressed in the latest version of BALiBASE (BALiBASE 3) (Thompson et al., 2005) and this database is now one of the most widely used reference standard. Nonetheless, BALiBASE remains a handmade dataset, with potential arbitrary and uneven biases resulting from human intervention. The main alternative to BALiBASE is Prefab (Edgar, 2004b), a very extensive collection of over a 1000 pairs of homologous structures, each embedded in a collection of about 50 homologues (25 for each structure) gathered by PSI-BLAST. In Prefab, the reference alignment is defined as the portions of alignments consistently aligned by two structural aligners: CE (Shindyalov and Bourne, 1998) and DALI (Holm and Sander, 1995). Prefab, however, is not a multiple sequence alignment collection since each dataset only contains a pair of structures thus making it a less stringent than BALiBASE where accuracy can be tested on entire multiple alignment columns rather than pairs of residues. Other commonly used databases for protein multiple sequence alignments include HOMSTRAD (Stebbings and Mizuguchi, 2004) and SABmark (Van Walle et al., 2005). One may ask why so many resources for addressing an apparently simple question. The answer probably lies in the complexity of structural alignments. While reasonably accurate structure based alignments are easy enough to generate, owing to the strength of the structural signal, it is nonetheless very hard to objectively assess the relative merits of alternative structure based alignments (Kolodny et al., 2005). Several alternative references are therefore available and no simple way exists to objectively evaluate their relative merits. In practice, the authors have taken the habit of running their methods on two or three datasets, verifying trend agreement. Recently Blackshield and Higgins(2006) produced an extensive benchmarking, comparing the 10 main MSA methods using 6 available datasets. The main trend uncovered by this analysis is that all the empirical reference datasets tend to yield similar results, quite significantly distinct from those measured on artificial datasets such as IRMbase (Subramanian et al., 2005), a collection of artificially generated alignments with local similarity. We checked by re-analyzing some of the Blackshield and Higgins benchmark data (Table 1.2). The methodology is very straightforward: each reference dataset is divided in subcategories, and altogether the 6 datasets make a total of 77 subcategories (68 for the empirical datasets, 9 for the artificial). Given two MSA methods A and B, we counted how many times the ranking suggested by one subcategory is in agreement with the ranking suggested by another subcategories (Agreement in Table 1.2). We then compared all the subcategories of a dataset against all the subcategories of the other datasets and reported the average figure in Table 1.2. We also computed the average agreement within

Table 1.1: Benchmarking of a selection of methods on the RV11 Balibase dataset. BaliBase RV11 is made of 38 datasests consisting of 7 or more highly divergent protein sequences (less than 20% pairwise identity on the reference alignment). All packages were ran using the default parameters. Servers were ran in August 2008.

Method	Version	Score	Mode	Templates	RV11	Server
3DPSI-Coffee	7.05	Consistency	Accurate	Profile + Structure	61.00	www.tcoffee.org
PROMAL-3D	server	Consistency	Default	Profile + Structure	58.66	prodata.swemd.edu/promals3d
PROMALS	server	Consistency	Default	Profile	55.80	prodata.swemd.edu/promals3d
PSI-Coffee	7.05	Consistency	Psicoffee	Profile	53.71	www.tcoffee.org
M-Coffee	7.05	Consistency	Muscl+Kal.+ ProbC +TC	—	41.63	www.tcoffee.org
T-Coffee	7.05	Consistency	Default	—	42.30	www.tcoffee.org
ProbCons	1.1	Consistency	Default	—	40.80	probcons.stanford.edu
ProbCons	1.1	Consistency	Monophsic Penalty	—	37.53	probcons.stanford.edu
Kalign	2.03	It + Matrix	Default	—	33.82	msa.cgb.ki.se
MUSCLE	3.7	It + Matrix	Default	—	31.37	www.drive5.com/muscle
Mafft	6.603b	It + Matrix	Default	—	26.21	align.genome.jp/mafft
Prank	0.080715	Matrix	Default	—	26.18	www.ebi.ac.uk
Prank	0.080715	Matrix	+F	—	24.82	www.ebi.ac.uk
ClustalW	2.0.9	Matrix	Default	—	22.74	www.ebi.ac.uk/clustalw

every dataset by measuring the agreement across different categories within a dataset. The results on Table 1.2 suggest that the 5 main empirical datasets are on average 72.4% consistent with one another. It means that any prediction of accuracy made on the basis of a single reference dataset is likely to be supported by 72.4% of similar measurements made on the 5 other empirical reference datasets. A striking observation is the lower agreement between the artificial dataset (IRMdb) and the empirical ones. Observations made on IRMdb are on average only supported by 58.1% of the observations made on the empirical datasets. Two factors could explain this discrepancy: the local nature of IRMdb, mostly designed for assessing local alignment capacities, or its artificial nature. The fact that empirical datasets biased toward local similarity (BALiBASE RV50, long indels, 76.8% agreement) do not show a similar trend suggest that the discrepancy between IRMdb and the empirical datasets owes much to its simulated component. Furthermore, at least three other studies reported similar findings, with results established on artificial datasets conflicting with empirical ones (Lassmann and Sonnhammer, 2002, 2005b; Loytynoja and Goldman, 2008).

While there is no clear consensus on this matter, we would argue here that the discrepancy between artificial and empirical datasets pleads in favor on not using the artificial ones. The use of artificial dataset should probably be restricted to situations where the process responsible for the sequence generation is well known and properly modeled, as happens in sequence assembly for instance. It is interesting to note that some sub-categories of BALiBASE are extremely informative albeit relatively small. RV11 for instance is 77.4% consistent with the entire collection of empirical dataset which makes it one of the most compact and informative dataset. This is not so surprising if one considers the nature of RV11, made of 38 highly divergent sequences (less than 25% id in the reference alignment). So far, this dataset has proven fairly resistant to heavy tuning and over-fitting and it is a striking observation that ProbCons, the only package explicitly trained on BALiBASE is not the most accurate (as shown on Table 1.1). Table 1.1 shows a systematic benchmarking of most methods discussed here on the RV11 dataset. Results are in broad agreement with those reported in most benchmarking studies published over these last 10 years, but the challenging nature of the dataset makes it easier to reveal significant difference in accuracy that are otherwise blurred by other less challenging datasets.

BALiBASE has had a strong influence on the field, prompting the design of novel reference datasets for sequences other than proteins. Similar to BALiBASE a reference

Table 1.2: Comparison of alternative reference datasets (adapted from Blackshield and Higgins). Blackshield and Higgins published the average accuracy of 10 MSA packages (Mafft, Muscle, POA, Dialign-T, Dialign2, PCMA,align_m, T-Coffee, Clustalw, ProbCons) on 6 reference databases. This table shows a new analysis of the original data. *Dataset* indicates the considered dataset. In this column, RV11 and RV50 are two BALiBASE categories, *Empirical Dataset* refers to the 5 empirical datasets (BALiBASE3, SabMark, Oxbench and Prefab). All datasets includes IRMdb as well. *#Categories* indicates the number of sub-categories contained in the considered datasets. *Agreement:* average agreement between all the considered categories of a given dataset and all the categories of the other databases. The agreement is defined as the number of times two given databases subcategories agree on the relative accuracy of two methods. The *Empirical dataset* average is obtained by considering all possible pairs of methods across all possible pairs of categories within the empirical datasets (i.e. all datasets except IRMdb). *Self-agreement:* same measure but restricted to a single database (i.e. each category in turn against all the other categories of the considered database).

Dataset	#Categories	Agreement (%)	Self-agreement
BaliBase	11	71.4	82.9
RV11	1	77.4	83.3
RV50	1	76.8	80.6
SabMark	4	69.8	81.3
Oxbench	10	65.0	70.8
Prefab	5	64.6	72.3
Homstrad	4	66.8	76.9
IRMdb	9	58.1	88.1
Empirical datasets	68	72.4	—
All datasets	77	66.1	—

dataset exists to validate ncRNA alignment methods, called BRAliBase (Wilm et al., 2006). BRAliBase works along the same lines as BALiBASE and relies on a comparison between an RNA alignment and its structure based counterpart. There is, nonetheless, a clear difference between these two reference datasets: in BRAliBase the reference structures are only predicted, and the final evaluation combines a comparison with the reference and an estimation of the predictive capacity of the new alignment. As such, BRAliBase is at the same time more sophisticated than BALiBASE (because it evaluates the prediction capacity of the alignment) and less powerful because it is not based on a sequence-independent method (unlike BALiBASE that uses structural comparison). This limitation results from the relative lack of RNA 3D structures in databases. Current benchmarking strategies have some short comings and cannot address all the situations relevant to MSA evaluation. These methods have nonetheless been used to validate all the currently available multiple sequence alignment packages and can certainly be cred-

ited (or blamed...) for having refocused the entire methodological development toward the production of structurally correct alignments. Well standardized reference datasets have also gradually pushed the MSA field toward becoming a fairly codified discipline, where all contenders try to improve over each other's methods by developing increasingly sophisticated algorithms, all tested in the same arena. Given the increased accuracies reported these last years, one may either consider the case closed, or suspect that time has come to change arena.

1.5 Common frameworks for MSA computation

An interesting consequence of the systematic use of benchmarking methods has been the gradual phase-off of most packages not based on the progressive algorithm (Hogeweg and Hesper, 1984). With the exception of POA (Lee et al., 2002), most of the methods commonly used nowadays are built around the progressive alignment. This popular MSA assembly algorithm is a straightforward agglomerative procedure. Sequences are first compared two by two in order to fill up a distance matrix, containing the percent identity. A clustering algorithm (UPGMA or NJ) is then applied onto this distance matrix to generate a rooted binary tree (guide tree). The agglomerative algorithm follows the tree topology thus defined and works its way from the leaf to the root, aligning two by two each sequence pair (or profile) associated with each encountered node. The procedure can be applied using any algorithm able to align two sequences or two alignments. In most packages, this algorithm is the Needleman and Wunsch (1970) or more recently the Viterbi algorithm (Durbin et al., 1998). As simple as it may seem, the progressive alignment strategy affords many possible adjustments, the most notable ones being the tree computing algorithm, the sequence weighting method and the gap weighting scheme. In recent work (Wheeler and Kececioglu, 2007), the authors have shown that that a proper tuning of these various components can take a standard method up to the level of the most accurate ones. ClustalW (Thompson et al., 1994) is often considered to be the archetype of progressive alignments. It is a bit paradoxical since its implementation of the progressive alignment significantly differs from the canonical one, in that it delays the incorporation of the most distantly related sequences until the second and unique iteration. This delaying procedure was incorporated in ClustalW in order to address the main drawback of the progressive alignment strategy: the greediness. When progressing from the leaves toward the root, a

progressive aligner ignores most of the information contained in the dataset, especially at the early stage. Whenever mistakes are made on these initial alignments, they cannot be corrected and tend to propagate in the entire alignment, thus affecting the entire process. With a large number of sequences, the propagation and the resulting degradation can have extreme effects. This is a well known problem, usually addressed via an iterative strategy. In an iterative scheme, groups of sequences are realigned a certain number of time, using either random splits or splits suggested by the guide tree. The most sophisticated iterative strategies (incorporated in Muscle and PRRP (Gotoh, 1996)), involve two nested iterative loops, an inner one in which the alignment is optimized with the respect to a guide tree, and an outer one in which the current MSA is used to re-estimate the guide tree. The procedure keeps going until both the alignment and the guide tree converge. It was recently shown that these iterations almost always improve the MSA accuracy (Wallace et al., 2005b), especially when they are deeply embedded within the assembly algorithm.

1.5.1 Consistency based MSA Methods

The greediness of progressive aligners limits their accuracy, and even when using sophisticated iteration schemes, it can be very hard to correct mistakes committed early in the alignment process. In theory, these mistakes could easily be avoided if all the information contained in the sequences was simultaneously used. Unfortunately this goal is computationally unrealistic, a limitation that has prompted the development of consistency based methods. In their vast majority, algorithms based on consistency are also greedy heuristics (with the exception of the Maximum Weight Trace problem (MWT) formulation of Kececioglu (Kececioglu, 1993), but even so, they have been designed to incorporate a larger fraction of the available information at a reasonable computational cost. The use of consistency for improved alignment accuracy was originally described Gotoh (1990) and later refined by Vingron and Argos (1991). Kececioglu provided an exact solution to this problem, reformulated as a maximum weigh trace problem. This exact approach is limited to small datasets but was further expanded by Morgenstern who proposed the first heuristic to solve this problem for large instances, thanks to the concept of overlapping weights (Morgenstern et al., 1996). While the notions developed in these four approaches are not totally identical, they have in common the idea of evaluating pairwise alignments through the comparison of a third sequences (i.e. considering an intermediate sequence). In practice, Gotoh did not use consistency to construct alignments, but rather to evaluate

them, and only considering three sequences. The consistency described by Vingron is very strict because it results from dot-matrices multiplications, therefore requiring strict triplet consistency in order to deliver an alignment. The overlapping weights described by Morgenstern also involve considering the support given by an intermediate sequence to a pairwise alignment, but in this context, the goal is to help guiding the incorporation of pairwise segments into the final MSA. While the overlapping weights bear a strong resemblance to the most commonly used definition of consistency, it is important to point out that Morgenstern also uses the term consistency but gives it a different meaning to describe the compatibility of a pair of matched segments within the rest of a partly defined multiple sequence alignments. The first combination of a consistency based scoring scheme with the progressive alignment algorithm was later developed in the T-Coffee package (Notredame et al., 2000). The main feature of a consistency based algorithm is its scoring scheme, largely inspired by the Dialign overlapping weights. Regular scoring schemes are based on a substitution matrix, used to reward identities and penalize mismatches. In a consistency based algorithm, the reward for aligning two residues is estimated from a collection of pairwise residue alignments named the library. Given the library, any pair of residues receives an alignment score equal to the number of time these two residues have been found aligned, either directly or indirectly through a third residue (Figure 1.3). The indirect alignments are estimated by combining every possible pair of pairwise alignments (i.e. XY + YZ = X-Y-Z). Each observation can be weighted with a score reflecting the expected accuracy of the alignment on which the observation was made. In the original T-Coffee, the residue pairs contained in the library were generated using a global (ClustalW) and a local (Lalign) method applied on each pair of sequences. At the time, the T-Coffee protocol resulted in a significant improvement over all alternative methods. This protocol was later brought into a probabilistic framework with the package ProbCons. In ProbCons, the sequences are compared using a pair HMM with a bi-phasic gap penalty (i.e. a gap extension penalty higher for short gaps than long gaps). A posterior HMM decoding of this HMM is then used to identify the high scoring pairs that are incorporated in the library, using their posterior probability as a weight. The library is then used to score the alignment with the T-Coffee triplet extension. Because it uses a library generated with a probabilistic method, this protocols is often referred to as ”probabilistic consistency” and has been incorporated in several packages, including SPEM (Zhou and Zhou, 2005), MUMMALS and PROMMAL (Pei and Grishin, 2006, 2007) as well as the latest versions of T-Coffee (version 6.00 and higher). Interestingly,

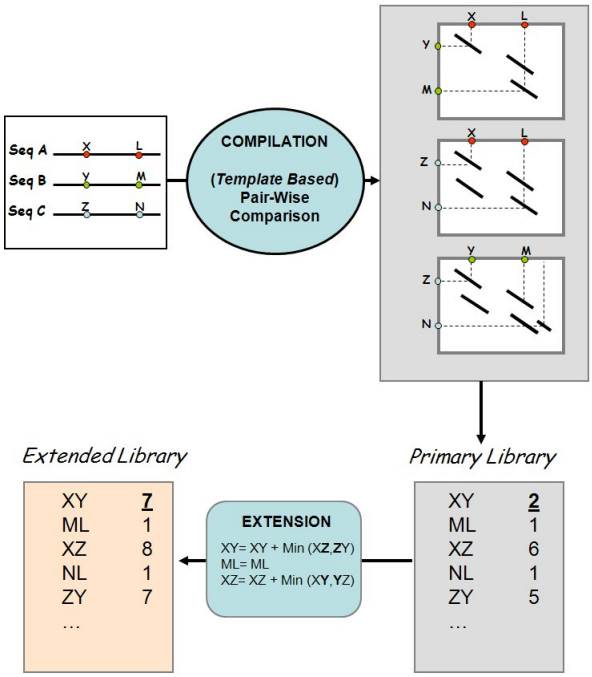


Figure 1.3: Generic overview for the derivation of a consistency based scoring scheme. The sequences are originally compared two by two using any suitable methods. The second box shows the projection of pairwise comparisons. These projections may equally come from multiple sequence alignments, pairwise comparison, or any method able to generate such projections, including posterior decoding of an HMM. They may also come from a template based comparison such as the one described in Figure 1.4. Pairs thus identified are incorporated in the primary library. These pairs are then associated with weights used during the extension. The figure shows the T-Coffee extension protocol. When using probabilistic consistency, the probabilities are treated as weights and triplet extension is made by multiplying the weights rather than taking the minimum.

the improvement is usually considered to be a consequence of the probabilistic framework when in fact it seems to result mostly from the use of a more appropriate gap penalty scheme at the pairwise level. For instance, Table 1.1 shows the effect of applying a regular gap penalty scheme (monophasic) when compared to the bi-phasic gap penalty scheme that ProbCons uses by default. This improvement has also been observed when incorporating the bi-phasic scheme in T-Coffee. Consistency based methods are typically 40% accurate when considering the column score measured on the RV11 dataset. This makes consistency based aligners about 10 points more accurate than regular iterative progressive aligners like ClustalW, Kalign, Muscle or Mafft. This increased accuracy comes at a cost and consistency based methods require on average N times more CPU time (N being the number of sequences) than a regular progressive aligner.

Aside from improved accuracy, an important aspect of consistency based scheme is the conceptual separation it defines between the computation of the original alignments, merged into a library, and the final transformation of this library into a multiple sequence alignment. This procedure made it straightforward to combine seemingly heterogeneous algorithms, such as ClustalW and Lalign in the original T-Coffee package, but it also opened the way towards a more generic combination of aligners. For instance, the latest versions of T-Coffee (Version 6.00 and newer) is able to combine up to 15 different alignment methods, including pairwise structural aligners, regular multiple sequence alignment methods, and even RNA alignment methods such as Consan (Dowell and Eddy, 2006). From the start, the T-Coffee framework made it possible to turn any pairwise method into a multiple alignment method, thus opening the way to two major developments undergone by multiple aligners these last years: meta alignment methods and template based alignments.

1.5.2 Meta-methods as an alternative to regular MSA methods

The wealth of available methods and the lack of a globally accepted solution make it harder than ever for biologists to choose a specific method. This dilemma is real and has recently received some renewed attention with a high impact report establishing the tight dependency of phylogenetic modeling on the chosen aligner. According to Wong and collaborators, phylogenetic trees may significantly vary depending on the methods used to compute the underlying alignment (Wong et al., 2008). In a similar way, several editions of the CASP (Battey et al., 2007) contest have revealed that a proper multiple alignment is an essential component of any successful structural modeling approach. A commonly advocated strategy is to use the method performing best on average, as estimated by benchmarking against structure based reference datasets. It is a reasonable martingale, like betting on the horse with the best odds. One wins on average, but not always... Unsurprisingly, benchmarks also make it clear that no method outperforms all the others, and that it is almost impossible to predict with enough certainty which method will outperform all the others on a specific dataset. It is quite clear that the chosen method is irrelevant on datasets made of sufficiently similar sequences (more than 50% pairwise identity). Yet, whenever remote homologues need to be considered, the accuracy drops and one would like to run all the available methods before selecting the best resulting alignment. This can be achieved when enough structural data is available (by selecting

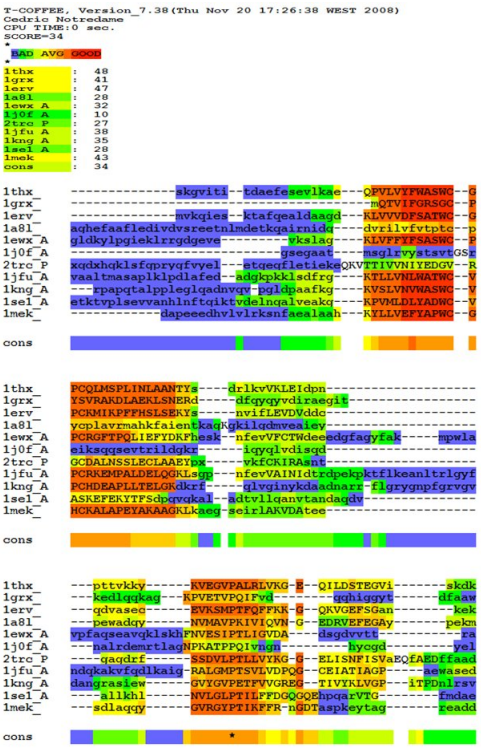


Figure 1.4: Typical colored output of M-Coffee. This output was obtained on the RV11033 BaliBase dataset, made of 11 distantly related bacterial NADH dehydrogenases. The alignment was obtained by combining Muscle, T-Coffee, Kalign and Mafft with M-Coffee. Correctly aligned residues (correctly aligned with 50% of their column, as judged from the reference) are in upper case, non-correct ones are in lower case. In this colored output, each residue has a color that indicates the agreement of the four initial MSAs with respect to the alignment of that specific residue. Dark red indicates residues aligned in a similar fashion among all the individual MSAs, blue indicates a very low agreement. Dark yellow, orange and red residues can be considered to be reliably aligned.

the alignment supporting the best structural superposition), or when functional information is at hand (by evaluating the alignment of similar features, such as catalytic residues). Unfortunately, experimental data is rarely available in sufficient amount, and when using several packages, one is usually left with a collection of alignments whose respective value is hard to assess in absolute terms. Chapter 2 presents a new method able to do this. Another method to address this issue are meta-methods. So far, M-Coffee (Wallace et al., 2006) has been the only package explicitly engineered to be used as a meta-method, although in theory all consistency based packages could follow suit. Given a multiple sequence dataset, M-Coffee computes alternative MSAs using any selected method. Each of the alignments thus produced is then turned into a primary library and merged to the main T-Coffee library. The resulting library is used to compute an MSA consistent with the

original alignments. This final MSA may be considered as some sort of average of all the considered alignments. When combining 8 of the most accurate and distinct MSA packages, M-Coffee produces alignments that are on average better than any of the individual methods. The improvement is not very high (1-2 point percent) but relatively consistent since the meta-method outperforms the best individual method (ProbCons) on about 2/3 of the 2000 considered datasets (HOMSTRAD, Prefab and BALiBASE)(Wallace et al., 2006). On a dataset like RV11, the improvement is much less marked (M-Coffee delivered alignments having an average accuracy of 37.5%) and one needs to restrict the combination to the 4 best non template based methods in order to obtain alignments with accuracy comparable to the best methods (Table 1.1). Yet, as desirable as it may be, the improved accuracy is not the main goal of M-Coffee and one may argue that rather than its accuracy, M-Coffee's main advantage is its ability to provide an estimate of local consistency between the final alignment and the combined MSAs. This measure (the CORE index (Notredame and Abergel, 2003)) not only estimates the agreement among the various methods (Figure 1.3) in a graphical way but it also gives precious indication on the local structural correctness (Notredame and Abergel, 2003; Lassmann and Sonnhammer, 2005a) and can therefore be considered as a good predictor of alignment accuracy. Previous benchmarking made on the original CORE measure suggest that a position with a consistency score of 50% or higher (i.e. 50% of the methods agreeing on a position) is 90% likely to be correct from a structural point of view. These results are consistent with those reported by Lassman and Sonnhammer (2005a) who recently re-implemented this measure while basing it on libraries made of alternative multiple sequence alignments. Even though these predictions are only restricted to a subset of the alignment, they can be an invaluable asset whenever a modeling process is very sensitive to alignment accuracy. For instance, the CORE index is used by the CASPER server to guide molecular replacement (Claude et al., 2004). From a computational point of view, meta-methods are relatively efficient. Provided fast methods are used to generate the original alignment, the meta-alignment procedure of M-Coffee can use a sparse dynamic programming procedure that takes advantage of the strong agreement between the considered alignments. A recent re-implementation of M-Coffee in the SeqAn (Döring et al., 2008) alignment library shows that the multiple alignment step of M-Coffee is about twice faster than standard consistency based aligners based on pairwise alignments like ProbCons or Promals (Rausch et al., 2008). Yet, all things considered, meta-methods only offer a marginal improvement over single methods, and they even suggest that the current state of the art

aligners are reaching a limit that may hard to break without some novel development in the field of sequence alignment. While waiting for a method able to accurately align two remote homologues in an ab-initio fashion (i.e. without using any other information than the sequences themselves), the best alternative is to use extra information, evolutionary, structural or functional. Template based MSA methods have been design to precisely address this aspect of data integration.

1.5.3 Template based MSA methods

The word template based alignment was originally coined by Taylor (1986) with reference to sequence/structure alignments. The notion was later extended within the T-Coffee package in a series of publications dedicated to protein and RNA alignments (Armougom et al., 2006b; Notredame and Higgins, 1996; O'Sullivan et al., 2004; Wilm et al., 2008). Template base alignment refers to the notion of enriching a sequence with the information contained in a template (Figure 1.5). The template can either be a 3D-structure, a profile, or a prediction of any kind. Once the template is precisely mapped onto the sequence, its information content can be used to guide the sequence alignment in a sequence independent fashion. Depending on the nature of the template one refers to its usage as structural extension or homology extension (sequence profile). Structural extension is the most straightforward protocol. It takes advantage of the increasing number of sequences with an experimentally characterized homologue in the PDB database. Given two sequences with a homologue in PDB, one can accurately superpose the PDB structures (Templates) and map the resulting alignment onto the original sequences. Provided the sequence/template alignment is unambiguous, this protocol yields an alignment of the original sequences having all the properties of a structure based sequence alignment. This approach only defines pair-wise alignments, but the alignment thus compiled can be integrated into a T-Coffee library and turned into a consistency based multiple sequence alignment (Figure 1.3 and 1.5). Structural extension was initially implemented in 3D-Coffee (O'Sullivan et al., 2004). EXPRESSO, a special mode of 3D-Coffee was then designed so that templates could be automatically selected via a BLAST against the PDB database. This protocol has recently been reimplemented in the PROMAL-3D (Pei et al., 2008) package. Structural extension is not limited to proteins, and recently several approaches have been described using RNA secondary structures as templates, these include T-Lara (Bauer et al., 2005), MARNA (Siebert and Backofen, 2005) and R-Coffee (Wilm

et al., 2008). In all these packages, sequences are associated with a predicted structural template (RNA secondary structure). The templates are then used by add-hoc algorithms to accurately align the sequences while taking into account the predicted structures (templates). The resulting pairwise alignments are combined into a regular T-Coffee library and fed to T-Coffee. Chapter 3 describes a new method SARA-Coffee which is able to produce structural informed RNA alignments. Homology extension works along the same principle as structural extension but uses profiles rather than structures. In practice, each sequence is replaced with a profile containing homologues. The profiles could be built using any available techniques although fast methods like PSI-BLAST have been favored. The first homology extension protocol was described by Heringa and implemented in the PRALINE package (Simossis and Heringa, 2005). PROMALS was described shortly afterwards (Pei and Grishin, 2007). PROMALS is a consistency based aligner, using libraries generated with the ProbCons pair-HMM posterior decoding strategy. PROMALS also uses secondary structure predictions in order to increase the alignment accuracy, although this extra information seems to only have a limited effect on the alignment accuracy. In Praline and PROMALS sequences are associated with a PSI-BLAST profile. A similar mode is also available in T-Coffee (Version 6.00+, mode=psicoffee) based on BLAST profiles (Table 1.1). The use of structural and homology extended templates results in increased accuracy in all cases. For instance, the combination of RNAplfold (Bernhart, et al., 2006) predicted secondary structures made R-Coffee more accurate at aligning RNA sequences than any of the alternative regular aligners, with a 4 point net improvement as estimated on BRAliBase (Wilm et al., 2008). The improvements resulting from homology extension on proteins are even more significant. On Prefab, the authors of PROMALS reported 9 points of improvement over the next best method (ProbCons). A similar usage of PROMALS or PSI-Coffee on category RV11 (distant homologues) of BALiBASE resulted in more than 10 points of improvement over the next best regular non template based aligner (Table 1.1). Of course, the most accurate alignments are obtained when using structural extension. In a recent work, Grishin and collaborators reported an extensive validation using a combination of structure and homology extension (Pei, 2008). Their results suggest that template based alignments achieve the best results when using structural extension. They also indicate that the choice of the structural aligner can make a difference, with DALI-Lite possibly more accurate than SAP. Given the same structural extension protocol, the authors report similar results between 3D-Coffee and PROMALS-3D, suggesting that the structural aligner is the most important component of

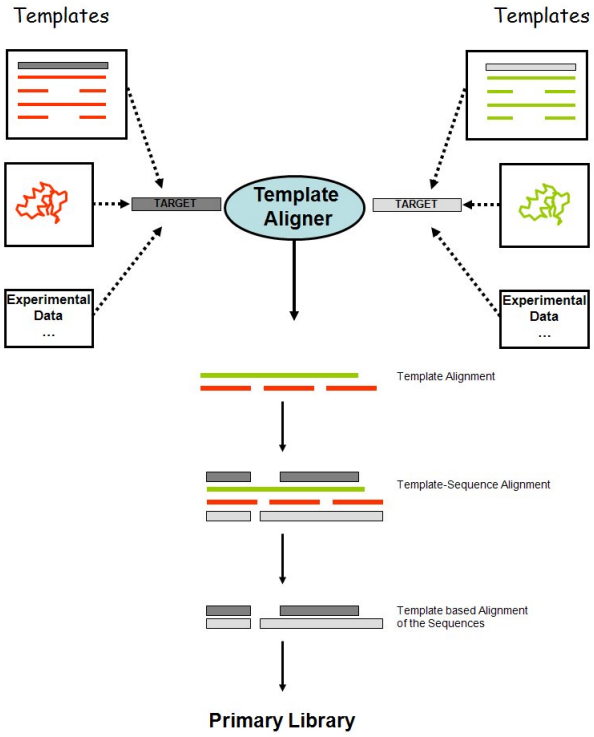


Figure 1.5: Overview of template-based protocols. Templates are identified and mapped onto the target sequences. The figure shows three possible types of templates: homology extension, structure and functional annotation. The templates are then compared with a suitable method (profile aligner, structural aligner, etc.) and the resulting alignment (or comparison) is mapped onto the final alignment of the original target sequences. The residue pairs thus identified are then incorporated in the primary library.

the protocol. The improvement is very significant, and on Prefab for instance, the combined use of DaliLite with homology extension resulted in nearly 30 points improvement over alternative non template based protocols. Results in Table 1.1 confirm these claims and suggest that the use of structural extension is the best way to obtain highly accurate alignments. This very high accuracy, obtained when using structural information is, however, to be interpreted with some caution. On the one hand, these high figures suggest a broad agreement between PROMALS-3D or 3D-Coffee alignments with the references. On the other hand, one should not forget that these methods use 3D information. As such, they are not any different from the methods used to derive the reference benchmarks themselves. It therefore means that PROAMLS-3D or 3D-Coffee/Expresso alignments may be seen as new reference datasets, generated with a different structural alignment protocol. Whether these are more or less accurate than the benchmarks themselves is open to interpretations, as it amounts to comparing alternative multiple structure based sequence

alignments.

1.5.4 New issues with the validation of template based methods

As reported by Kolodny et al. (2005), the task of comparing alternative structure based alignments is complex. In order to address it, authors have recently started using alignment free evaluation methods. These methods consider the target alignment as a list of structurally equivalent residues and estimate how good would be the resulting structural superposition. These measures are either based on the RMSD (Root Mean Squared Deviation: average squared distance between homologous alpha carbons) or the dRMSD (distance RMSD: average square difference of distances between equivalent pairs of amino acids) like the DALI score (Holm and Sander, 1995), APDB (O'Sullivan et al., 2003) or the iRMSD (Armougom et al., 2006a). So far, three extensive studies (Armougom et al., 2006a; O'Sullivan et al., 2003; Pei et al., 2008) have suggested that the results obtained with these alignment-free benchmarking methods are in broad agreement with those reported when using regular benchmarks. The main drawback of these alignment free evaluation methods is their reliance on distance measures strongly correlated with the methodology used by some structural aligners (Dali in particular) thus raising the question whether they might be biased toward this particular structural aligner. A simpler and not yet widely used alternative would be to evaluate the modeling potential of the alignments, by measuring the accuracy of structural predictions based upon it. This could probably achieved by recycling some components of the CASP evaluation pipelines.

1.5.5 Alignment of very large datasets

Accuracy has been a traditional limitation of multiple sequence alignments for the last 20 years, and it is no surprise that this issue has been the most actively addressed, if only because inaccurate alignments are simply useless. The other interesting development has been the increase of the number of sequences. Traditionally, the length of the sequences (L) was greater than the number of sequences (N), and most methods were tuned so that they could deal with any value of L, assuming N would not be a problem. This is especially true of consistency based methods that are cubic in complexity with N, but only quadratic with L. With N<<L, the extra-cost incurred by consistency remains

manageable, but things degrade rapidly when N becomes big. Yet, it is now clear that L is bounded, at most by the average length of a genome. N, on the other hand, has no foreseeable limit and could reflect the total number of species or the total number of individuals (past and present) in a population or even the total number of haplotypes in a system. Dealing with large values of N should therefore be considered a prime goal. In the context of a progressive algorithm, the first easy step is to speed up the guide tree estimation, for instance using a ktup based method, as most packages currently do (-quicktree option in ClustalW). The second step is to use an efficient tree reconstruction algorithm. The default UPGMA and NJ algorithms are cubic with the number of sequences, but these algorithms can be adapted in order to become quadratic, as is the case with the current ClustalW implementation. Even so, quadratic algorithms will not be efficient enough when dealing with very large datasets and more efficient data compression methods (such as those used to decrease redundancy in databases) will probably need to be used in the close future (Blackshields et al., 2008). The next step for decreasing CPU requirements is to use an efficient dynamic programming strategy. This is the strategy used by MAFFT that relies on a very efficient dynamic programming. Consistency based methods have a disadvantage because of the N-cubic requirement of consistency. Yet, the protocol is relatively flexible and heuristics can probably be designed to estimate the original library more efficiently. For instance, PCMA (Pei et al., 2003) starts by identifying subgroup of sequences closely related enough to be prealigned. SeqAn (Rausch et al., 2008) takes advantage of the sparse matrix defined by the extended library and only does the minimum required computation to guarantee optimality. SeqAn also makes an attempt to treat the sequences as a chain of segments rather than a chain of residues thus considerably reducing the CPU requirements for closely related sequences. The SeqAn library has been designed to be linked with any of the consistency based aligners. Even so, the complexity of most consistency based aligners remains too high to deal with the very large datasets that are expected to come. Chapter 4 presents a solution to reduce the complexity of the consistency approach on large datasets. Phylogeny being one of the main application of large-scale alignments, it will also be worth evaluating the phylogenetic potential of these large-scale methods. Doing so is far from trivial as it connects with the delicate issue of establishing reference tree collections. More generally it addresses the problem of predicting accurate trees from multiple sequence alignments.

1.5.6 Phylogenetic relevance of multiple sequence alignments

The pace of accumulation of new entire bacterial genomes (and to a lesser extend eukaryotic genomes) can only be compared with the discovery of new species along the XIXth century. Never have we had so much molecular data at hand to reconstruct the natural history of life, a real challenge for intelligent design supporters. Multiple sequence alignments constitute the ideal compost on which to grow these trees, and although there have been a few reports of alignment free tree reconstruction methods (Ferragina et al., 2007), the difficulty of aligning distantly related sequences probably means that unless a breakthrough happens in the field of sequence alignments and guarantees error free pairwise alignments, MSAs will remain the starting point for most phylogenetic analysis. An interesting paradox of the whole MSA field is that although most methods are defined within some sort of phylogenetic framework (progressive alignment), they are only evaluated for their capacity of producing structurally correct MSAs. As a consequence, we do not really know how MSA methods fare with respect to phylogenetic reconstruction and, assuming the current structural benchmarks reflect well enough the evolutionary relation among proteins, we do not really know if this analysis can be safely extrapolated to ncRNAs. Recent work suggest (Katoh and Toh, 2008) that the accuracy ranking of the best packages is roughly the same when benchmarking on RNA sequences (BRAliBase) or protein sequences, but little is known about the accurate reconstruction of RNA based phylogenetic tree. This is a paradoxical situation when considering that most trees of life are derived from a multiple sequence alignment of ribosomal RNA sequences. Two high impact publications have made an attempt to raise the attention of the community on the issue of phylogenetic reconstruction (Loytynoja and Goldman, 2008; Wong et al., 2008). The work by Wong shows that phylogenetic reconstruction can be very sensitive to the MSA method used to deliver the alignment. The authors stopped short of proposing a way for selecting the best phylogenetic trees, but they make it clear that various methods can lead to different models, a new concept in a field where MSAs had always been considered to be data rather than models. It is a context where meta-methods could certainly provide an element of answer, mostly by helping selecting the sites on the basis of their expected accuracy, using the CORE index or any related method. In this context the main advantage of the CORE index is to provide a filter independent from sequence conservation, as opposed to other accuracy predictors. An MSA region can have a low level of conservation but a high CORE index, provided all the pairwise alignments are in agreement with respect to the considered position. Regions where conservation is low

and consistency high may be considered prime targets for phylogenetic reconstruction. The PRANK+F (Loytynoja and Goldman, 2008) algorithm was described shortly afterward and also addresses the issue of accurate phylogenetic reconstruction seen from an MSA perspective. PRANK+F is a novel attempt to model the gap insertion required by the alignment process in a phylogenetically meaningful way. This new approach opens up the possibility of incorporating the indel signal in the reconstruction of evolutionary scenarios, but it also raises an equally important question: given that alternative aligners lead to different trees, and given that the signal contained in the alignments can be used in many different ways, how are we going to evaluate the phylogenetic potential of multiple sequence alignment methods? Building reference datasets is a very difficult task in phylogeny where an objective, independent source of information for establishing the correct history of a set of sequences is usually lacking. Fossil records provide little help when it comes to selecting true orthologous sequences. Given a sequence dataset, it is therefore very hard, and maybe impossible to establish a correct reference tree. So far, the validation of tree reconstruction methods has therefore focused on the method's ability to optimize a mathematical model (Guindon and Gascuel, 2003). Even when this optimization is highly successful, the only guarantee is the mathematical correctness of the final tree with no clear guarantee on its biological relevance, except that provided by expert diagnostic of the tree (i.e. the observation that related species are grouped by the tree in a biologically meaningful way). This situation is very similar to that encountered with MSA computation where one has on the one hand the mathematical correctness of a method, estimated by its capacity to optimize a given objective function (sums-of-pairs, viterbi, etc...) and on the other hand, the biological accuracy, estimated by comparison with a reference alignment. In the context of MSA analysis, the use of structure made it clear that there could be a significant discrepancy between mathematical correctness and biological accuracy. Unfortunately, the equivalent of structural information is not available in phylogeny, and most current strategies, including Prank+F are validated on simulated data. The simplest approaches simulate both the data and the trees using generators like ROSE (Stoye et al., 1997). As pointed out earlier, the results obtained on simulated data differ significantly from those measured on empirical data, and for instance, PRANK outperforms all alternative packages on phylogenetic simulated data, but performs poorly when it comes to reconstructing structural alignments. Assuming the relevance of results established on the simulated data, this suggests there could major differences between phylogenetically accurate alignments and structurally accurate ones, an hypothesis that

remains to be further tested and confirmed.

1.6 Genome alignments

Alignments of genomes allow the discovery of conserved regions like enhancers and similar motifs which are often located outside of genetic regions, and can also detect large genome rearrangements. Now that the number of newly sequenced genomes increases rapidly, there is a high demand of the community for accurate genome alignments. For several reasons, computing genome alignments is much more challenging than the computation of short sequence alignments of proteins, RNA, promoter or genes. Obviously, the first difficulty encountered is the size of genomic sequences which can reach several hundreds of millions base pairs (Mbp). For example the human genome consists of around 3,137 Mbp in its version GRCh37.p7. Currently, bacteria, with several hundred new sequences each year, are the domain for which the number of new genomes is growing the fastest (Karsch-Mizrachi et al., 2012). Of course compared to eukaryotic genomes their genome size is much smaller but can as well reach the size of more than 10Mbp (eg. Sorangium cellulosum (Schneiker et al., 2007)). Thus applying algorithms developed for much shorter sequences (not more than a few kb) is generally unfeasible due to time and memory requirements. As a consequence heuristics have been proposed to reduce the search space and with it the time and memory needed. After these more technical considerations, the second and arguably more difficult problem is biological. Beside point mutations, insertions and deletions, which can all be handled by MSAs, genomes gather during the course of evolution a much larger variety of changes and rearrangements on a scale which can involve whole chromosomes. These rearrangements range from the shifting of sequence parts to a new position up to the fusion and fission of chromosomes. Figure 1.6 shows several of these rearrangements which are not possible to be solved with MSA programs because they generally assume co-linearity of the sequences. Although the figure shows the rearrangement of genes they can affect any kind of sequence segment of any length. Even otherwise very similar genomes can have large-scale rearrangements. For example the human and chimpanzee genomes are very similar (The Chimpanzee Sequencing Analysis Consortium, 2005), but the human genome has one chromosome less because two ancestor chromosomes fused in the human lineage into a single one and several inversions have been found (Yunis et al., 1980; The Chimpanzee Sequencing Analy-

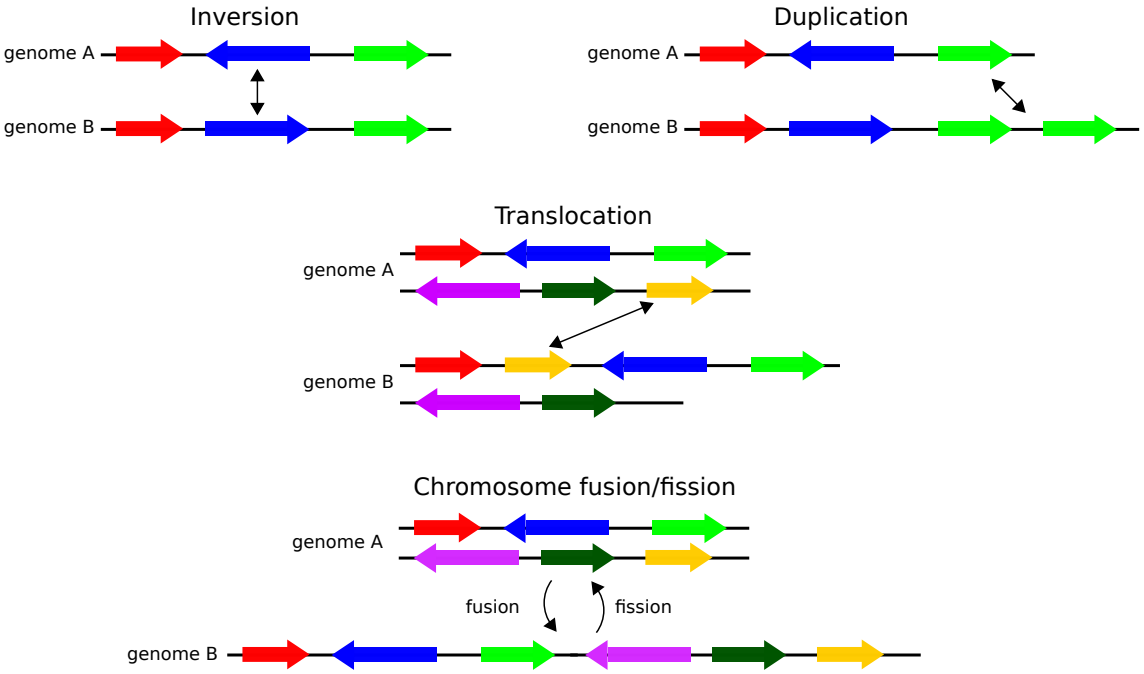


Figure 1.6: Different genome rearrangements. Black lines denote chromosomes and arrows genes.

sis Consortium, 2005). Normal linear aligners would completely misalign these segments, resulting in a much higher difference between the genomes than supported by reality. The number of rearrangements increases with the evolutionary distance between the genomes; hence, it is a critical point that a genome aligner has to solve. Further problems include the low information content of the DNA alphabet, and the larger divergence of genomes in non-coding regions, which makes it harder to identify corresponding nucleotides. Furthermore, approaches associated to MSAs like the usage of structural information or sequence profiles cannot be applied to genome alignments as DNA has no fine grained three dimensional structures as have proteins. Profiles might be of help but would heavily increase the computational time and at the current state not enough organisms have been fully sequenced.

A common method to address the problem of aligning large sequences is the anchor strategy. The first step is to find sequence segments, the anchor points, which are considered to be homologous and are aligned with each other. Then the sequences between the anchor points are aligned using normal dynamic programming. AVID (Bray et al., 2003), LAGAN (Brudno et al., 2003a), M-GCAT (Treangen and Messeguer, 2006) and

Pecan (Paten et al., 2009) are examples for alignment programs using this approach. Special care has to be taken when choosing the anchor points as they dramatically reduce the search space by determining which sequence parts can be aligned with each other. Thus the accuracy of an alignment can be strongly reduced if the anchor points are badly chosen and do not reflect segments which would be aligned in full dynamic programming. Hence, maximal extended matches (MEMs) and maximal unique matches (MUMs) are widely used approaches. Maximal unique matches appear only once in the sequences and thus can be regarded as secure anchor points if they are of sufficient length. MUMs and MEMs are found efficiently with suffix arrays or suffix trees as done for example in MUMmer (Kurtz et al., 2004). In order to find good anchor points Cgaln (Nakato and Gotoh, 2010) splits the genomes into equally sized blocks and uses k-mer frequencies between blocks to find similar segments. Often the first round of anchor detection is not sufficient to decrease the space for standard dynamic programming. To solve this, the anchor detection step is recursively repeated with less restrictive parameters in large inter-anchor region. GLASS (Batzoglou et al., 2000) and many other aligners use this recursive step to increase anchor point detection. This very general anchor strategy has been refined in different ways. The GS-Aligner (Shih and Li, 2003) encodes the DNA sequence into a row of two bits and uses lookup tables to efficiently calculate anchor points. Pecan defines a small frame around the anchor points in which standard dynamic programming is applied. Furthermore it uses a consistency approach to improve alignment accuracy. BlastZ (Schwartz et al., 2003) uses a slightly different approach. Instead of performing dynamic programming between anchor points it uses an approach similar to Gapped BLAST (Altschul et al., 1997) end extends the found anchor points until a score drop of a certain value is reached.

The first approaches to solve rearrangements were based on the usage of a reference sequence. When aligning two genomes, one of them is used as a reference (or target) sequence and the other is split into pieces which can be aligned linearly to the reference. Shuffle-LAGAN (Brudno et al., 2003b) for example uses local alignments to find homologous segments. These local alignments are chained together according to their occurrence in the reference and merged into larger blocks when having the same direction. These blocks are then aligned with the reference sequence using the LAGAN algorithm (Brudno et al., 2003a). The usage of a reference alignment is especially disadvantageous when aligning more than two genomes, because sequence segments are only aligned when appearing in the reference, thus common insertions are missed. Furthermore, duplications

are not treated or are only treated in one of the sequences. The Shuffle-Lagan approach was later extended by the VISTA pipeline (Dubchak et al., 2009) which is able to align multiple genomes without a reference alignment. The chains of local blocks were not only computed according to one sequence but to both, and the segments were chosen according to an out-group. TBA (Blanchette et al., 2004) was one of the first methods to think of a genome alignment as a collection of blocks each representing homologous segments rather than a single linear alignment. An advantage of such a description is the possibility to sort the blocks according to any of the included sequences depending on the need. However, TBA was only able to contain a limited amount of blocks, no inversions or duplications were treated. Mercator (Dewey, 2007) and Enredo (Paten et al., 2008) were the first programs to consequently think of a genome alignment to be a collection of blocks without any use of a reference sequence. Both programs use a similar strategy to split the sequences into homologous blocks. Mercator uses gene annotations and the BLAST program to find corresponding proteins between the genomes. The principle is to merge several proteins into a single block when they appear consecutively in the same direction. As only orthologous protein exons are considered, duplications and rearrangements within non-protein regions are not treated in the Mercator algorithm. Another disadvantage is that due to the usage of BLAST it does not take non-coding RNA regions into account. These problems have been addressed in the Enredo program, which uses the more general concept of genome point anchor (GPA) instead of restricting itself to exons. These GPAs can be any kind of similar segments between two or more sequences and can be produced, for example, by local alignment programs. Similar to the Mercator approach a graph is build using a set of non-overlapping GPAs whose nodes are merged as long as colinearity is given. The resulting blocks of Mercator or Enredo can then be aligned using any kind of normal MSA program. A schematic overview of this process can be seen in Figure 1.7. Other algorithms use the A-Bruijn graph (Zhang and Waterman, 2003; Raphael et al., 2004) to model the different rearrangements. More recently the cactus graph (Paten et al., 2011) has been proposed to take the place of the A-Bruijn graphs. These graph methods have in common that nodes represent conserved regions between two or more genomes whereas the edges represent sequences from one to the next block. These graphs can contain loops to allow the representation of duplications inside the sequences.

Even with the many available genome aligners it is still a difficult process to calculate a genome alignment due to necessary preparations of the sequences (e.g. repeat filter-

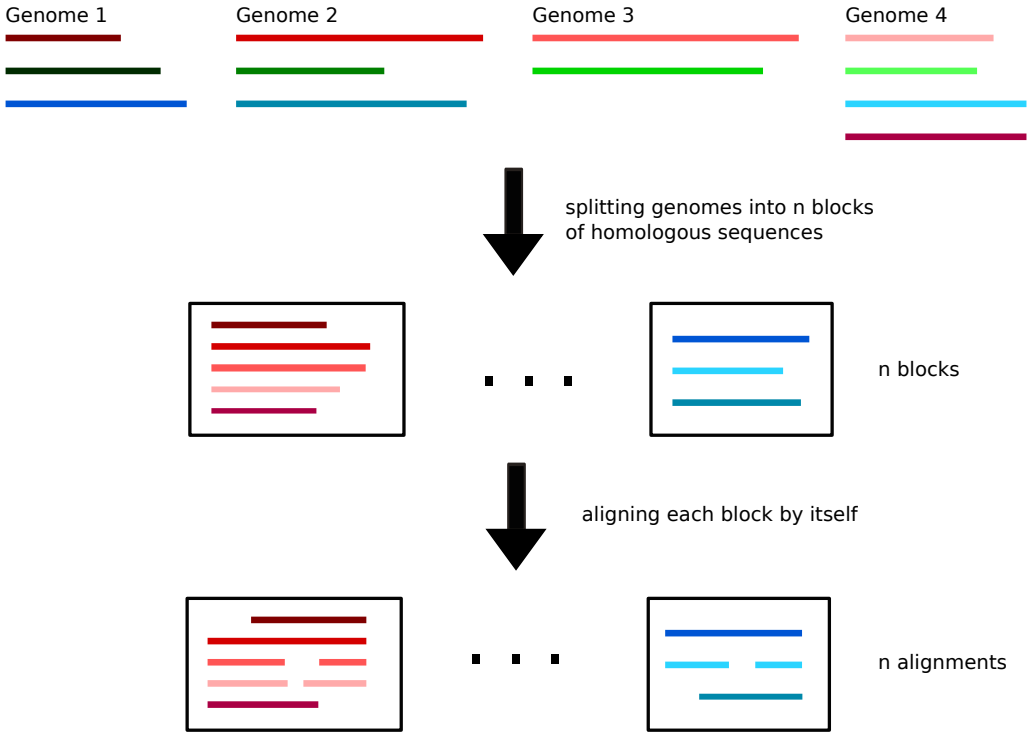


Figure 1.7: Shematic overview of the two steps usually executed to produce multiple genome alignment.

ing) and the often high computational resource demands. Because of this several genome alignments of the most commonly used species are made available by projects like Ensembl (Flicek et al., 2012) or the UCSC Genome Browser (Kuhn et al., 2012).

1.6.1 Genome alignment evaluation

For the asessment of protein alignments several gold standard benchmarks exist and are regularly used for the evaluation of new programs, but not a single benchmark exists for genome aligners turning the evaluation of an aligner into a complicated undertaking. Due to this situation many different metrics have been developed for the task of evaluating genome alignments each with different advantages and disadvantages. A common method is to evaluate the overall alignment based on the quality of the alignment of exons as done in LAGAN and the VISTA pipeline (Brudno et al., 2003a; Dubchak et al., 2009). In this approach one determines orthologous exons and measures how accurately they are aligned. This method only examines coding regions which is a big drawback because

they are known to be easier to align than non-coding regions due to a higher conservation rate. A similar approach using ancestral repeats instead of exons has been used to validate the Enredo-Pecan pipeline (Paten et al., 2008). The percentage of complete and partial alignments of ancestral repeats is calculated to estimate the accuracy of the alignment. Finally, Bray et al. measured the coverage of coding regions and UTRs in their alignments (2003).

The measurements described so far are confined to very specific segments of sequences but other metrics exists taking the whole alignment into account. One of these methods is the accuracy estimation using simulated data sets. Different programs (e.g. Rose (Stoye et al., 1998), evolver (Edgar et al.)) have been developed to model the evolutionary process. Using a simulateddatasethas the advantage that all the information needed to measure the accuracy of an alignment is known. As the programs evolve the sequences, they keep track of the changes introduced and thus are able to output not only the generated sequences but additionally the correct alignment as well. After aligning the sequences with an alignment program the resulting alignments can then be simply compared to the reference alignment produced by the simulator software. This approach has been used to validate several algorithms (Blanchette et al., 2004; Darling et al., 2004, 2010; Paten et al., 2011). A problem of this approach is that the real process of evolution is unknown, thus the model that is used might not reflect the reality. An additional problem arises if the same or a similar model is used in the alignment program which used to align the sequences as this introduces circularity to the benchmark and the result might only reflect the similarity of the models used and not the accuracy of the alignment. This effect has been already shown on proteins (Blackshields et al., 2006). Several programs have been developed to measure the accuracy of an alignment without knowing the actual true alignment. Examples for these programs are PSAR (Kim and Ma, 2011) and StatSigma-W (Prakash and Tompa, 2007; Chen and Tompa, 2010). For a given alignment PSAR removes iteratively each sequence from it, realigns it to the remaining aligned genomes using an HMM and then measures the agreement between the two alignments. A score is thus computed for each column representing the probability that they are correctly aligned. This approach is related to the alignment uncertainty issue explained in the next section. StatSigma-w uses a different approach. It tries to identify a branch of a given phylogenetic tree which separates the alignment into two subsets which might be misaligned. For this purpose it uses a so-called discordance score, which has high similarity to a p-value, a measure of the significance of a statistical observation.

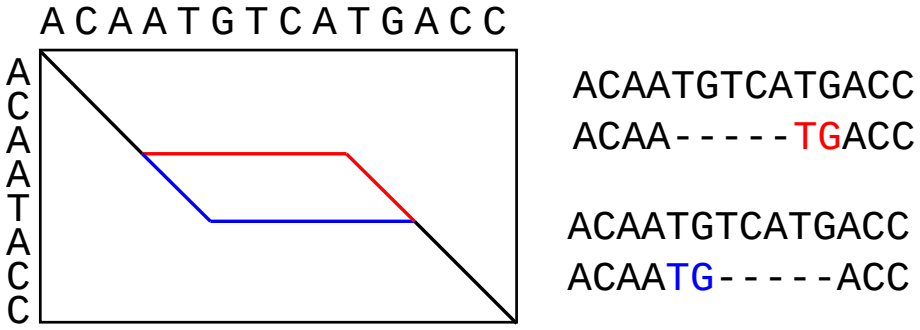


Figure 1.8: The left side shows the dynamic programming path of an alignment of two sequences. Most aligners choose arbitrarily which of the two paths to follow. The two possible alignments are shown on the right side. The substring TG switches position depending on which path is followed.

1.7 Alignment uncertainty

As Wong et al. (2008) stated, a multiple alignment is not a true observation (i.e. data), but the result of applying a specific model to a givendatasetresulting in an alignment in dependence of the used model. Several times it has been shown that this can have a large influence on downstream analysis. In the above-mentioned paper the authors study the influence of the aligner on the reconstructed phylogeny. Another study (Markova-Raina and Petrov, 2011) shows the influence of alignments on the detection of positive selection. While it is obvious that different aligners produce different alignments, it is less recognized, but not less important, that the same aligner may calculate a different alignment on the same set of sequences. The reason is that most alignment programs are sensitive to the direction of the sequence, so that reversing the order of characters in all sequences would give a different alignment (Landan and Graur, 2007). A similar effect can also be observed when changing the order of sequences in the input file. Figure 1.8 shows the explanation for this behavior. In the alignment process points exist in which one can align characters to one position or another with equal score. Usually the choice is taken arbitrarily, always in the same direction. Thus when changing from aligning sequence A with sequence B to aligning B with A the tie is broken differently. Most of the alignment programs suffer from this kind of problems. An exception is the FSA (Bradley et al., 2009) program which bypasses the issue by aligning nucleotides in the order of their probability to be aligned. As pairs of nucleotides are added, their consistency with the set of already chosen pairs needs to be checked, resulting in high running times.

Several approaches have been suggested to measure the stability of an alignment. The T-Coffee package (Notredame et al., 2000) uses the core index, a score reflecting the agreement of an alignment position with the different pairwise alignments. The heads and tails method (Landan and Graur, 2007) compares alignments constructed with the original set (heads) and constructed with the sequences in the reverse order (tails) using the column and sum of pairs score as described by Thompson et al. (1999). The GUIDANCE score (Penn et al., 2010) calculates the robustness of an alignment by comparing alignments computed with different guide trees which are used in progressive multiple sequence alignment methods. It is generally possible to apply the methods to multiple genome alignments as well but problems arise when e.g. a different guide tree leads to a different splitting of the genomes making a comparison of the resulting alignments difficult. Often post-processing steps are undertaken to eliminate unreliable columns from the alignment as they might cause artifacts in downstream analyses. TrimAl (Capella-Gutiérrez et al., 2009) for example has been developed to identify and remove columns from an alignment according to different gap criteria allowing automated post-processing in large-scale approaches. A similar trimming approach is proposed in Chapter 4 for large data sets, which allows to identify sequences responsible for introducing uncertainty around gaps.

2 MSA evaluation using structural information: STRIKE

Kemena C, Taly JF, Kleinjung J, NotredameC, "STRIKE: evaluation of protein MSAs using a single 3D structure.", Bioinformatics, vol. 27, no. 24, pp. 3385–3391, 2011.

3 RNA structural alignment: Sara-Coffee

Title: Using tertiary structure for the computation of highly accurate multiple RNA alignments with the SARA-Coffee package

Authors: Carsten Kemena+, Giovanni Bussotti+, Emidio Capriotti, Marc A. Marti-Renom, Cedric Notredame (+ These two authors contributed equally to this work)

Status: (submitted)

Contribution: CK has been doing the NiRMSD evaluation as well as the implementation in the framework of T-Coffee. GB has been doing the benchmark preparation and the 3SP evaluation.

3.1 Abstract

Motivation: Aligning RNAs is useful to search for homologous genes, study evolutionary relationships, detect conserved regions and identify any patterns that may be of biological relevance. Comparing RNA sequences is, however, difficult owing to the poor level of conservation among homologues, even when considering evolutionary related sequences.

Results: We describe SARA-Coffee a tertiary structure based multiple RNA aligner and validate it using BRAliDARTS, a new benchmark framework designed for evaluating tertiary structure based multiple RNA aligners. We provide two methods to measure the capacity of alignments to match corresponding secondary and tertiary structure features. On this benchmark, SARA-Coffee outperforms both regular aligners and those using secondary structure information. Furthermore we show that on sequences in which less than 60% of the nucleotides form base pairs, primary sequence methods usually perform better than secondary structure aware aligners.

Availability and implementation: The package and the datasets are available from: http://www.tcoffee.org and http://structure.biofold.org/sara/.

3.2 Introduction

Recent reports of a large number of previously unknown RNA coding genes (Guttman et al., 2009) have prompted a renewed interest in the field of non-coding RNA (ncRNAs) analysis. This shows well in the growing number of scientific reports uncovering a rapidly expanding range of new functions and it now appears that non-coding RNAs are involved in most essential parts of the cell machinery, including X inactivation (Xists (Brown et al., 1992)), genome integrity maintenance (piRNA (Farazi et al., 2008)), transcript knock-down and cell differentiation (miRNA (Lee et al., 1993)) as well as nuclear trafficking (NRON (Willingham et al., 2005)), among others. From a functional standpoint, the main consequence of high throughput sequencing has certainly been the discovery of long ncRNA (lncRNA), simultaneously identified as un-reported non-coding ENCODE transcripts (Orom et al., 2010) and as conserved genomic regions with active promoter chromatin signatures (Guttman et al., 2009). The exact function of this new class remains a matter of debate though mounting bodies of evidence suggest their involvement in gene regulation, either through trans (Rinn et al., 2007) or cis-acting (Orom et al., 2010) mechanisms. Other reports are also suggesting the potential usage of lncRNAs as biomarkers (Romanuik et al., 2009). In human only, the latest ENCODE catalogue lists more than 10,000 lncRNA genes, and probably more have to come as a wider range of tissues get deep sequenced. Such a pace of discovery makes the elucidation of lncRNA a promising future milestone in biological research.

Making sense of so much information will depend on our ability to build homology-based models (Capriotti and Marti-Renom, 2008a). Alignment methods rely on the notion that key features are usually preserved by evolution through purifying selection. Multiple comparison models can therefore reveal functional elements that would otherwise be difficult to identify on a single sequence. This is especially true for structured RNA molecules where compensated mutations are frequent signatures for evolutionarily maintained stem loops. This strategy has been extensively used for the successful elucidation of ribosomal RNA secondary structures (Gutell and Fox, 1988). Unfortunately, producing alignments accurate enough to be used for secondary structure prediction can be a challenging

task, especially when dealing with distantly related sequences. Two main obstacles exist that prevent the computation of informative homology based models. First of all, RNA sequences use a four-letter alphabet, with no higher order meta-alphabet (like protein's amino acid code) that would help powering statistical analysis. As a consequence, structure similarity becomes hard to infer when sequences have less than 60% identity (Abraham et al., 2008; Capriotti and Marti-Renom, 2010). Secondly, RNA sequence evolution is mostly constrained by the maintenance of secondary structure elements stabilized through a combination of canonical and non-canonical base-pairings. Under such constraints, it has been shown that sequences can evolve rapidly while exploring so-called neutral networks (Huynen et al., 1996). The combination of a small alphabet with rapid evolution makes it difficult to use standard alignment tools like BLAST-based approaches (Altschul et al., 1990). To address these limitations, one can tap into the evolutionary signal contained in di-nucleotides that results from the co-evolution of adjacent bases. This approach has been recently shown to be effective enough for the improvement of database search accuracy (Bussotti et al., 2011). Unfortunately, the signal thus uncovered is very modest and unlikely to result in significantly improved alignments. A more convincing solution involves the simultaneous estimation of sequence and structural conservation using Sankoff's algorithm (Sankoff, 1985). As effective as it may be in theory, this approach is hampered by prohibitive memory and CPU requirements, a limitation that has prompted the development of a large number of faster approximate heuristics for the inclusion of secondary structure information when aligning RNA. Some of the most popular tools include R-Coffee (Wilm et al., 2008), LocARNA (Will et al., 2007) and Consan (Dowell and Eddy, 2006). Consan combines expectation maximization with a sophisticated banded dynamic programming strategy, which results in a heuristic approximation of Sankoff's algorithm. The Consan algorithm that only aligns two sequences at a time, can easily be combined with a consistency based multiple sequence aligner like T-Coffee (Notredame et al., 2000) or R-Coffee (Wilm et al., 2008) in order to assemble MSAs.

Consistency based aligners (Do et al., 2005; Notredame et al., 2000; Roshan and Livesay, 2006; Wilm et al., 2008) rely on the compilation of an exhaustive library of all-against-all pairwise alignments. This library is extended in order to derive a position specific scoring scheme, used to compute a standard progressive alignment. The main strength of multiple aligners like T-Coffee is to allow any third party pairwise aligner to be used for the library generation. This property was previously used to generate structure based protein alignments (O'Sullivan et al., 2004) by combining structural pairwise aligners like SAP

(Taylor and Orengo, 1989). We show here how this approach, originally developed for proteins, can easily be extended to RNA sequence alignments provided suitable pairwise tools are used to build the pairwise library. Structure-based RNA alignment algorithms include SARA (Capriotti and Marti-Renom, 2008b, 2009), DIAL (Ferre et al., 2007), ARTS (Dror et al., 2005), LaJolla (Bauer et al., 2009), R3D Align (Rahrig et al., 2010) and SARSA (Chang et al., 2008). These tools all belongs to a recently described class of aligners that make use of experimentally derived three-dimensional structures. In this study, we chose SARA but in practice, any of the above mentioned tools could be used either as a replacement or in combination with SARA.

To derive a structure based alignment, SARA calculates a series of unit vectors between consecutive C3' atoms and aligns them using dynamic programming, to maximize the set of superimposed atoms within a root mean square deviation. As a stand-alone pairwise structural aligner, SARA is directly usable within the T-Coffee framework. In this work, we describe and benchmark a combination of these two packages named SARA-Coffee and able to generate multiple structure based RNA alignments.

The main motivation of this work is not only to describe a procedure for structure based RNA multiple structural alignments but also to assess the relative benefits of using experimental RNA tertiary and secondary structure, and to determine to which extend these expensive approaches can benefit modeling projects. Bypassing experimental 3D structures is very important, since determining RNA structures are much harder to determine than proteins (hence the much lower number of RNA structures in the PDB). We therefore took the opportunity of a pure structure based validation in order to estimate whether experimental structure is effectively improving modeling accuracy, either at the 2D or the 3D level. This question is especially relevant in a context where it will soon be relatively easy to use next-generation sequencing in order to do massive secondary structure estimation at minimal cost (Kertesz et al., 2010; Wan et al., 2012).

For this purpose we built a new benchmarking framework containing enough tertiary structure data. This framework is named BRAliDARTS. Its main characteristic is to be independent from any reference multiple sequence alignment and therefore totally unbiased towards one method or another. BRAliDARTS only contains sequences with a known 3D-structure and is used to evaluate MSAs for their capacity to match homologous structural features. MSAs are evaluated using the NiRMSD (Armougom et al., 2006a), a method designed to estimate MSAs structural accuracy (see Methods). BRAliDARTS is not the

only benchmark framework for RNA MSA analysis and other approaches have been described like BraliBase (Gardner et al., 2005) or Rfam (Griffiths-Jones et al., 2005). Both benchmarks explicitly rely on reference alignments and are therefore potentially biased towards specific aligning strategies. It is the lack of such potential bias, together with full reliance on 3D information that sets BRAliDARTS aside from other benchmarking frameworks.

3.3 Methods

3.3.1 Benchmarking Dataset

Our benchmark is a collection of 41 dataset each made of several unaligned homologous structures. These datasets were compiled from the DARTS database (Abraham et al., 2008). DARTS stores 1,333 RNA structures that can be clustered in 94 structurally homogenous subgroups using ARTS (Dror et al., 2005). Not all DARTS sequences are suitable for the approach described here and some filtering was needed in order to define a usable subset. The initial dataset was filtered by: (i) removing all sequences tagged as fragments by DARTS, (ii) converting all non-canonical residue symbols into an N; (iii) updating outdated PDB structures with their newer versions; (iv) removing RNA-DNA hybrids and structures including heteroatoms; (v) removing structures containing less than 9 nucleotides; (vi) removing clusters in which X3DNA (Lu and Olson, 2003) failed to extracting at least one secondary structure; (v) removing structures with discrepancies between the ATOM and the SEQRES PDB fields; and (vi) removing clusters with less than 3 sequences. The final dataset resulted in a total of 41 distinct sequence sets containing a total of 486 structures (see Suplementary Materials). We named this dataset collection BRAliDARTS, by reference to BRAliBase (Gardner et al., 2005), a popular reference dataset used for RNA aligners benchmarks. BRAliDARTS can be downloaded from http://www.tcoffee.org/Projects/saracoffee. Note that all the results presented here are based on readouts obtained on 31 datasets, since 10 BRAliDARTS datasets had to be discarded for being either invariant across all considered methods or for inducing out of range readouts when using the iRMSD (cf. Result section). We furthermore extracted a high quality subset (BRAliDARTS-HQ) from the initial dataset which contains only X-ray structures with resolution lower than 2.85 Å. Besides that we discarded all the RNAs

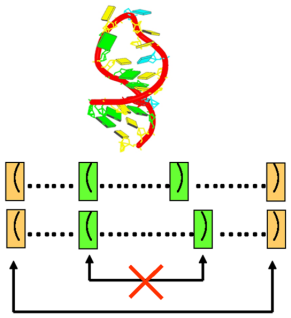


Figure 3.1: Schema of the 3SP score computation. Base pairs are colored in orange when they match on both side, green indicates partial matches.

with low secondary structure density (BP-index < 0.48) representing inter-molecular interactions. This resulted in a set of 10 clusters with a total of 79 sequences.

3.3.2 Benchmark

In BRAliDARTS, reference datasets do not come along with reference alignments, and are merely defined as sets of homologous sequences each with an associated 3D structure. It is important to stress that the benchmark strategy described here relies on all the considered sequences having an experimentally known 3D structure. We used two MSA method independent metrics. The first one is adapted from the NiRMSD citepArmougom2006, a measure originally defined to evaluate protein MSAs by comparing the variation in intra-molecular distances (as inferred from the evaluated MSA itself and the 3D structure of the considered sequences). The NiRMSD can be described as a normalized form of the distance RMSD. In this work, the original package was adapted in order to evaluate intra-molecular distances using the RNA ribose C3' instead of the peptidic alpha carbons. The principle of a distance RMSD is to compare variations of distances between pairs of aligned residues. Its main advantage over a standard RMSD is its non-reliance on a structural superposition. Equivalent residues are declared by the alignments and intra-molecular distances are directly estimated within the non superposed 3D structures. The second metric is named Secondary Structure Sum of Pairs (3SP). 3SP is a simple measure estimating the number base pairs where each side of the pair is aligned with equally contacting residues. We used the m3 implementation of this measure originally described

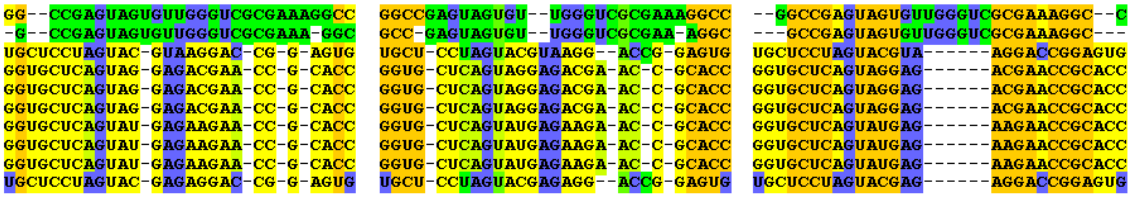


Figure 3.2: Example of 3 alignments showing the agreement across known secondary structure. From left to right: T-Coffee, R-Coffee, SARA-Coffee. Maximal agreement across all pairs is indicated in orange, minimal agreement is in dark green. Blue nucleotides are not involved in the secondary structure.

in (Notredame et al., 1997) and formalized as follows:

$$3SP = \frac{\sum_{i,j} P_{i,j}}{\sum_{i,j} min(p_i, p_j)} \tag{3.1}$$

where $P_{i,j}$ is the number of residue pairs found to be in agreement when considering the pairwise alignment of sequences i and j. Figure 3.1 shows a schematic overview of this metrics. When using a newick-like representation of RNA secondary structures, the 3SP metrics amounts to estimating the number of matching parenthesis aligned with equally matching parenthesis and normalize this value by its theoretical maximum. Figure 3.2 shows an example of three colored alignments using this metric.

3.3.3 Sara-Coffee

Our new method is called SARA-Coffee and is based on R-Coffee. R-Coffee is a consistency aligner for RNA. It can be described as a modified version of T-Coffee able to incorporate predicted (RNAPlfold (Bernhart et al., 2006)) secondary structure . One advantage of consistency aligners is their ability to re-construct a multiple sequence alignment out of any collection of pairwise sequence alignments. SARA-Coffee has two important improvements over R-Coffee. First of all, SARA-Coffee uses SARA, a pairwise RNA 3D structure alignment method to assemble its library. Secondly, it does the sequence alignment using true (rather than predicted) secondary structures, as estimated by applying 3DNA onto the PDB files. To determine the influence of 2D/3D structure information

we also designed two additional T-Coffee: R-CoffeeReal and BestPairs. R-CoffeeReal is the default R-Coffee ran using experimental secondary structures. BestPair is a mixture of SARA-Coffee and R-Coffee that runs SARA on a single pair of sequences (the most closely related) and ignores true structural information for all the other pairs of sequences.

3.3.4 Alignment Comparison

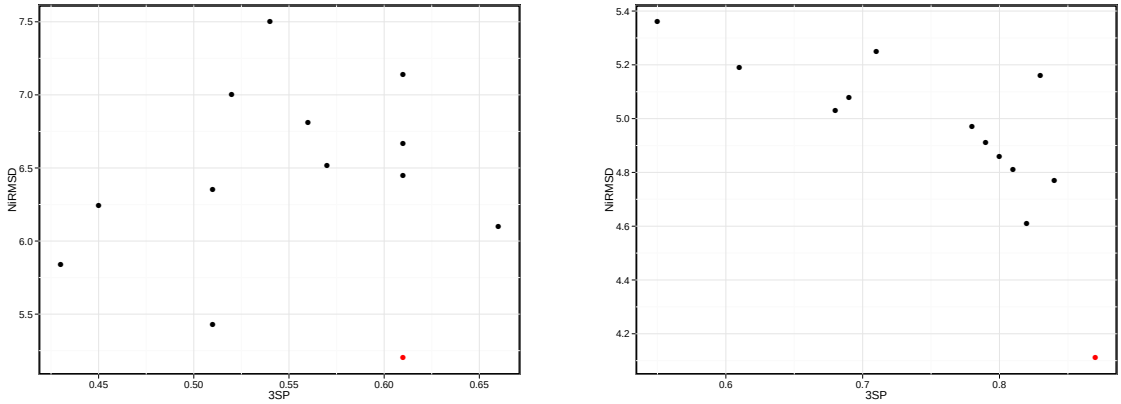
We compared SARA-Coffee to both generic and structure aware aligners. Generic aligners include: ClustalW 1.82 (Larkin et al., 2007), MAFFT (default) 6.624b (Katoh et al., 2005) Probalign 1.4 (Roshan and Livesay, 2006), ProbconsRNA 1.1 (Do et al., 2005) and T-Coffee 8.28 (Notredame et al., 2000). Structure aware aligners use structural information while assembling an MSA. Structural information can either be predicted from single sequences LocARNA 1.6.2 (Will et al., 2007), MAFFT-qinisi 6.864b, MXSCARNA 2.1 (Tabei et al., 2008) and R-Coffee 8.28 (Wilm et al., 2008), or using compensated mutations as in Consan-Coffee 8.28 (Dowell and Eddy, 2006).

3.4 Implementation/Distribution

SARA-Coffee is part of the standard T-Coffee distribution, an open-source freeware available from http://www.tcoffee.org. It requires the SARA program as a plugin, which is available from http://structure.biofold.org/sara/. The benchmark dataset including the evaluation procedure is available from http://www.tcoffee.org/Projects/saracoffee.

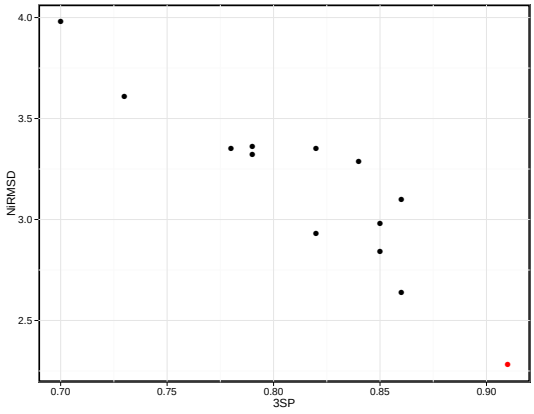
3.5 Results

Our main goal is to estimate the effectiveness of structural information incorporation when assembling RNA multiple sequence alignments. We were especially interested in quantifying the usefulness of three-dimensional information and its relative merits in comparison with inferred secondary structures. To address this problem we focused a large part of this work on the design of BRAliDARTS, a structure based benchmark system. Aside from its full reliance on 3D information, BRAliDARTS' main strength is its total independence from any reference MSA. This independence makes it possible to avoid any



(a) Low Density Structures

(b) High Density Structures



(c) BRAliDARTS-HQ

Figure 3.3: Correlation between the NiRMSD and the 3SP measure. Each point represents one of the 13 MSA methods tested here. SARA-Coffee is shown in red.

bias towards specific alignment methods. Having assembled BRAliDARTS, we tested five generic aligners, five secondary structure aware aligners and the three new methods described here on the 41 datasets of BRAliDARTS. We then calculated 3SP and NiRMSD, the two metrics developed for BRAliDARTS. The 3SP estimates the fraction of base pairs aligned with a potentially homologous pair. This metrics merely requires knowing the secondary structure of the considered sequences. The NiRMSD is used to estimate the variation of intra-molecular distances across homologous pairs of residue pairs (as defined by the MSA one evaluates). It relies on the notion that in a correct alignment, the distance between two residues in a structure should be as similar as possible to the distance between homologous residues in another structure. The NiRMSD is independent from any structural superposition, a key property to avoid any circularity when bench-

marking methods like SARA that depends on a 3D superposition. After evaluating the 3SP and the NiRMSD on the full BRAliDARTS benchmark (Supplementary Table 3.1), we decided to remove several datasets on the basis of their readouts. We first excluded 6 datasets for which the NiRMSD values were on average higher than 10 Å across the 13 aligners. Such high values suggest either a combination of non-homologous structures, or some lower level problem in the PDB files. We also excluded 4 more datasets for which all 3SP or NiRMSD readouts were completely identical across the 13 aligners. This left us with a total of 31 datasets on which we based all subsequent analysis. Since our aim was to quantify the importance of structural information when assembling an RNA MSA, we first estimated the fraction of residues involved in a base pair in each dataset. This measure, referred to as base pair index (BP-index) in the rest of the text, varies significantly across datasets and ranges from 6% to 90% with a median close to 73%. We split the BRAliDARTS accordingly in two subsets, one containing datasets made of low-density secondary structures (14 datasets, BP-index $<$73%) and a second one containing high-density structures (17 datasets, BP-index $\geq$73%) (Supplementary Table 3.1). We then averaged readouts for each alignment method in both the high and the low-density bin (Table 3.1). On the low-density dataset, we found the 3SP readouts to behave roughly according to expectations, with primary methods delivering results about 10 percent point lower than their secondary counterparts (0.48 vs 0.59). Tertiary methods like SARA-Coffee were among the best. These observations are in stark disagreements with similar readouts measured using the NiRMSD where we found the primary methods to outperform the secondary (6.16 Å vs 6.80 Å, with the lowest values being the best ones). By contrast, SARA-Coffee delivers the best performance, with the lowest NiRMSD measured across all methods. This result suggests that secondary structure information does not help building an MSA as they are not a dominant feature in the considered sequences. In this context, the lack of any strong correlation between the 3SP and the NiRMSD (Figure 3.3(a)) suggests that secondary structure accuracy (3SP) is a poor proxy for the overall MSA accuracy when dealing with low-density structures.

Our measures on the other subset of BRAliDARTS, the one with highly connected structures gave very different results. On this dataset, the 3SP and the NiRMSD measures are strongly correlated (-0.74, Figure 3.3(b)). The differences between primary and secondary or tertiary methods are also much more pronounced. We found more than 20 point percent improvement on 3SP between the primary methods and SARA-Coffee and more than 1 Å on the NiRMSD. On that same metrics, 5 out of 6 secondary methods outper-

Table 3.1: The table lists for each method the results of the benchmark as well as the CPU time needed to align all datasets of BRAliDARTS. The best readout in each column is indicated with a bold case.

Structural Information	Method	low structured		high structured		Time (s)
		3SP	NiRMSD	3SP	NiRMSD	
Primary	ClustalW	0.43	5.84	0.55	5.36	**2**
	T-Coffee	0.51	6.35	0.69	5.08	37
	Mafft	0.51	5.43	0.68	5.03	4
	Probalign	0.45	6.24	0.61	5.19	12
	ProbconsRNA	0.52	7.00	0.71	5.25	14
	Average	*0.48*	*6.16*	*0.64*	*5.18*	-
Secondary	Mafft-qinsi	0.54	7.50	0.78	4.97	20
	LocARNA	**0.66**	6.10	0.81	4.81	601
	MXSCARNA	0.61	7.14	0.83	5.16	15
	R-Coffee	0.56	6.81	0.80	4.86	229
	R-CoffeeReal	0.61	6.45	0.84	4.77	511
	Consan-Coffee	0.57	6.52	0.79	4.91	1168135
	Average	*0.59*	*6.80*	*0.81*	*4.91*	-
Tert./Sec.	BestPair	0.61	6.67	0.82	4.61	551
Tertiary	SARA-Coffee	0.61	**5.20**	**0.87**	**4.11**	19324

form all the primary methods. This result suggests that on densely structured sequences, one can improve MSA accuracy by using secondary or tertiary structural information with the best results being achieved with 3D information. A possible confounding factor when observing this correlation might be the effect of low-resolution structures, in which the BP-Index could have been underestimated. In that case, the correlation might have to do more with structural data quality than with base-pairing density. To rule out this possibility we used BRALiDARTS-HQ a third reference dataset made of a small number of carefully selected high quality structures and found the correlation to be even stronger (-0.91, Figure 3.3(c)).

Of course, one may argue that the superiority of SARA-Coffee is merely the result from using experimental secondary structures instead of predicted ones and not the result of using tertiary structure information. We addressed this question with R-CoffeeReal, an adaptation of R-Coffee, using experimental rather than predicted secondary structures. Results (Table 3.1) show the effectiveness of this approach. On the high structural density dataset, R-CoffeeReal manages to outperform all the other non-tertiary based methods. It remains nonetheless significantly less accurate than SARA-Coffee. The burden of requir-

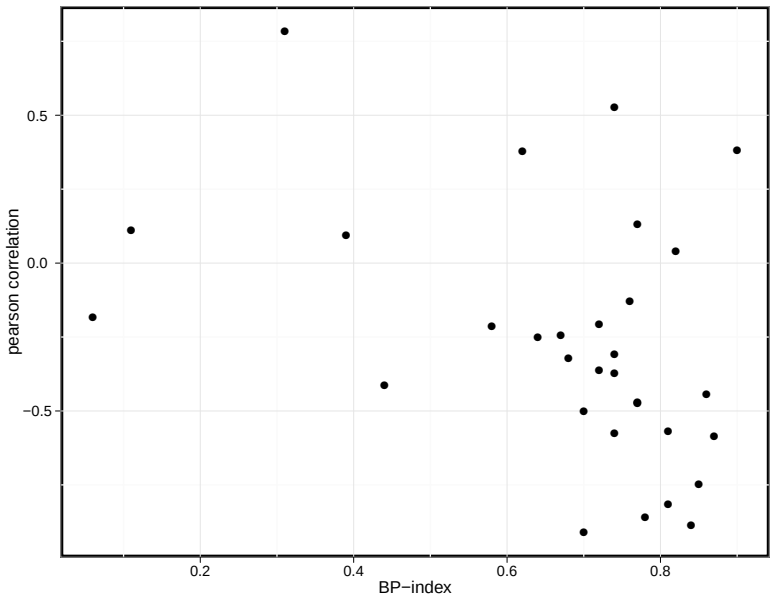
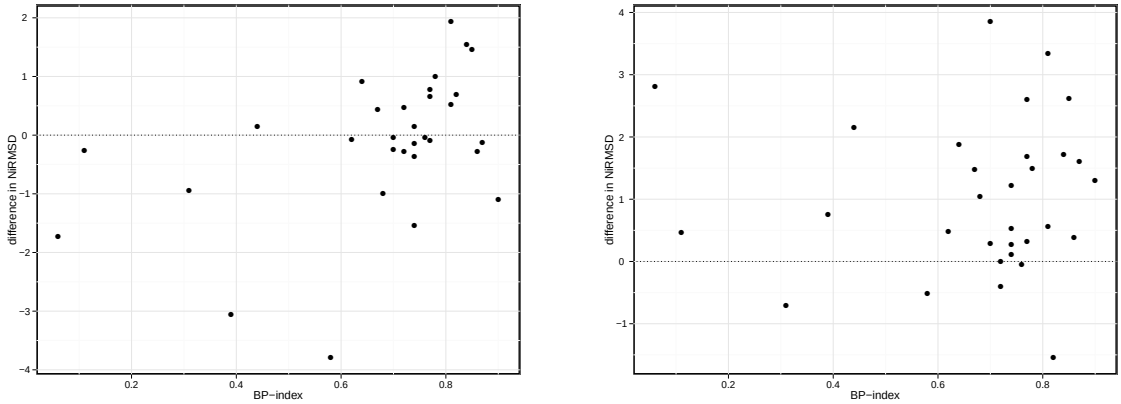


Figure 3.4: Dependency of the NiRMSD/3SP correlation on structural density Each point corresponds to one of the 31 datasets used for benchmark. The horizontal axis indicates the fraction of nucleotides involved in a base pair in the considered dataset. The vertical axis corresponds to the Pearson correlation coefficient between the NiRMSD and the 3SP readouts measured on the MSAs produced using the 13 alternative methods displayed on Table 3.1

ing an experimental for each RNA sequence one wants to align dramatically limits the scope of SARA-Coffee. We therefore asked whether using only a handful of structures might be enough to significantly improve MSA modeling (BestPair method). BestPair results are in par with those achieved with other methods (Table 3.1). On the high structural density subset, its 3SP score is the second best one (ignoring R-CoffeeReal) and its NiRMSD is better than that of any secondary structure albeit significantly less good than that measured on SARA-Coffee. All combined together, the results measured on the low and the high-density BRAliDARTS subsets suggest some heterogeneity and a strong sensitivity to datasets structural composition. When dealing with highly structured sequences, secondary and tertiary methods perform better, and result in highly correlated secondary (3SP) and tertiary readouts (NiRMSD). This correlation seems to disappear when analyzing low-density structures. We tested this hypothesis a bit further by taking advantage of the availability of 13 alternative MSAs for each single dataset. This variety allowed us to estimate a Pearson correlation coefficient between the 3SP and the NiRMSD of each single dataset and plot the resulting values against the BP-index (Figure 3.4). Despite a rather weak correlation, the trend shows an increasing correlation above a BP-index



(a) Secondary methods vs primary methods

(b) SARA-Coffee vs primary methods

Figure 3.5: Relative usefulness of structural information. Each point corresponds to one the 31 datasets used for benchmarking.

of 60%, with most datasets above 80% BP-index having very strong correlations. This result suggests that secondary structure-based methods to be best suited for datasets having a BP-index higher than 80%. We tested this hypothesis by measuring on each dataset, the difference in NiRMSD readouts between primary and secondary methods (Figure 3.5(a)). As expected, we found that below a BP-index of 60%, primary methods tend to give better results, while above this value, secondary structure aware methods often result in an improvement. A similar analysis carried out by comparing primary and the tertiary methods (Figure 3.5(b)) shows that SARA-Coffee yields its most significant improvements on datasets with a BP-index above 60% but rarely degrades the MSAs below this value.

We finished our analysis by doing a pairwise comparison of all the methods considered here and by counting, for each metrics, the number of time any method outperform any other method (Supplementary Figure 3.1 and 3.2). Such a comparison is important as it makes it possible to estimate the statistical support for the observed differences. We found most differences to be statistically significant on the 3SP method, while on the NiRMSD, SARA-Coffee is the only aligner whose behavior appears to be statistically different from most alternatives on most datasets. These comparisons, that reflect individual dataset readouts also support the notion of secondary structure information being more useful when dealing with highly structured sequences.

3.6 Conclusion-Discussion

In this work we introduce SARA-Coffee, a new tool for generating multiple RNA structure alignments and we show how the usage of tertiary information can result in significantly improved RNA alignments. We quantified these improvements using a purpose built benchmark framework named BRAliDARTS. BRAliDARTS is made of 41 collections of homologous RNA sequences with known 3D structures and two evaluation metrics independent from any reference alignment. An important focus of our work has been the precise quantification of structural information usefulness when assembling an RNA MSA. By doing in parallel similar analysis on 3 categories of methods that use primary, secondary and tertiary structure information, we have shown that aligners using secondary structure information are rarely suitable when dealing with sequences in which less than 80% of the nucleotides are involved in a base pair. Below this figure, methods that rely on predictions appear to induce a degradation of MSA accuracy. By contrast, tertiary methods like SARA-Coffee, almost always manage to improve MSA models accuracy, regardless of the fraction of structured nucleotides.

In the real world, RNA tertiary information is rather scarce and we therefore had to ask whether alternative sources of information could be reasonable substitutes for tertiary data. For instance, it is now possible to do large scale secondary structure prediction using high throughput sequencing and the two leading technologies for single molecule sequencing techniques, PacBio and Nanopore, have been announcing kits dedicated to large scale secondary structure determination. It is therefore realistic to consider that a wide amount of secondary structure information will soon be available. We tested the effect of using this information of a variation of the R-Coffee method named R-CoffeeReal. Our results are encouraging. They show that when dealing with highly structured RNAs ($\geq$73%) the use of experimental secondary structure results in MSAs significantly better than those obtained with alternative secondary methods, even though accuracy does not reach the level of pure tertiary structure based alignments. This result was also supported by the high correlation (0.91) observed when measured on BRAliDARTS-HQ.

The main limitation of our work is probably its reliance on a rather small collection of dataset. Indeed, starting from 41 BRAliDARTS datasets, we ended up doing the validation on 31 collections only. Furthermore, sequences making up this dataset are also rather short (44 nucleotides on average, 213 at most). In that context, it is not entirely clear how

the behavior of methods relying on secondary structure prediction can be extrapolated to longer sequences, since it is well known that length tends to impact structure prediction accuracy (Ding et al., 2008; Doshi et al., 2004). By contrast, it is quite likely that the good performances measured on R-CoffeeReal and all methods using experimental data will hold reasonably well.

Overall, we conclude that no ideal substitute exists for experimental data when modeling RNA homology. Experimentally proven secondary structures are the next best thing after tertiary information, but they appear to merely provide models accurate enough for secondary structure analysis. This however, should already be enough for a variety of modeling applications, and especially when building stochastic context free grammars in order to look for remote homologues.

Acknowledgment

The authors thank Eric Westhof for the helpful discussions.

Funding: CN, and CK are funded by the Centro de Regulacio Genomica (CRG), the Plan Nacional (BFU2011-28575) and the EU (Quantomics, KBBE-2A-222664). GB is founded by the la Caixa PhD Fellowship Program. EC acknowledges support from the Marie Curie International Outgoing Fellowship program (PIOF-GA-2009-237225). MAM-R acknowledges support from the Spanish MINECO (BFU2010-19310).

3.7 Supplementary Material

Supplementary Table 3.1: DARTS-Sub clusters. This table lists for each cluster the average percent pairwise identity as estimated on the Sara-Coffee alignments, the number of sequences and the average sequence length.

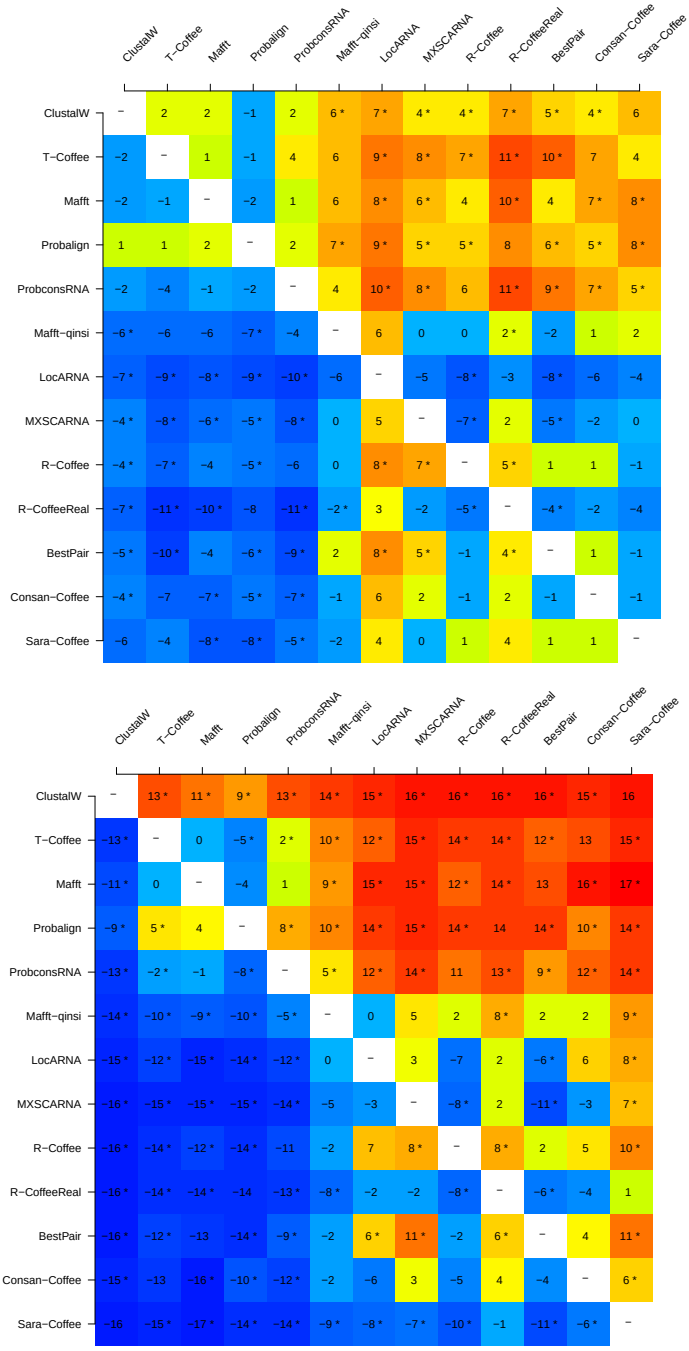
Cluster	% identity Sara-Coffee	# sequences	avg sequence length
1	43.43	8	27
2	78.87	11	36
3	67.73	8	47
4	66.14	4	31
5	40.56	16	21
6	42.25	6	28
7	61.52	10	31
8	57.66	71	76
9	63.23	11	29
10	30.76	16	16
11	51.56	22	22
12	39.94	4	24
13	39.27	13	23
14	61.34	12	39
15	80.81	11	213
16	61.06	6	31
17	86.36	4	23
18	44.62	10	49
19	75.85	8	26
20	83.92	10	29
21	36.52	24	16
22	31.71	3	25
23	55.80	7	47
24	50.72	5	62
25	44.40	24	15

Continued on next page

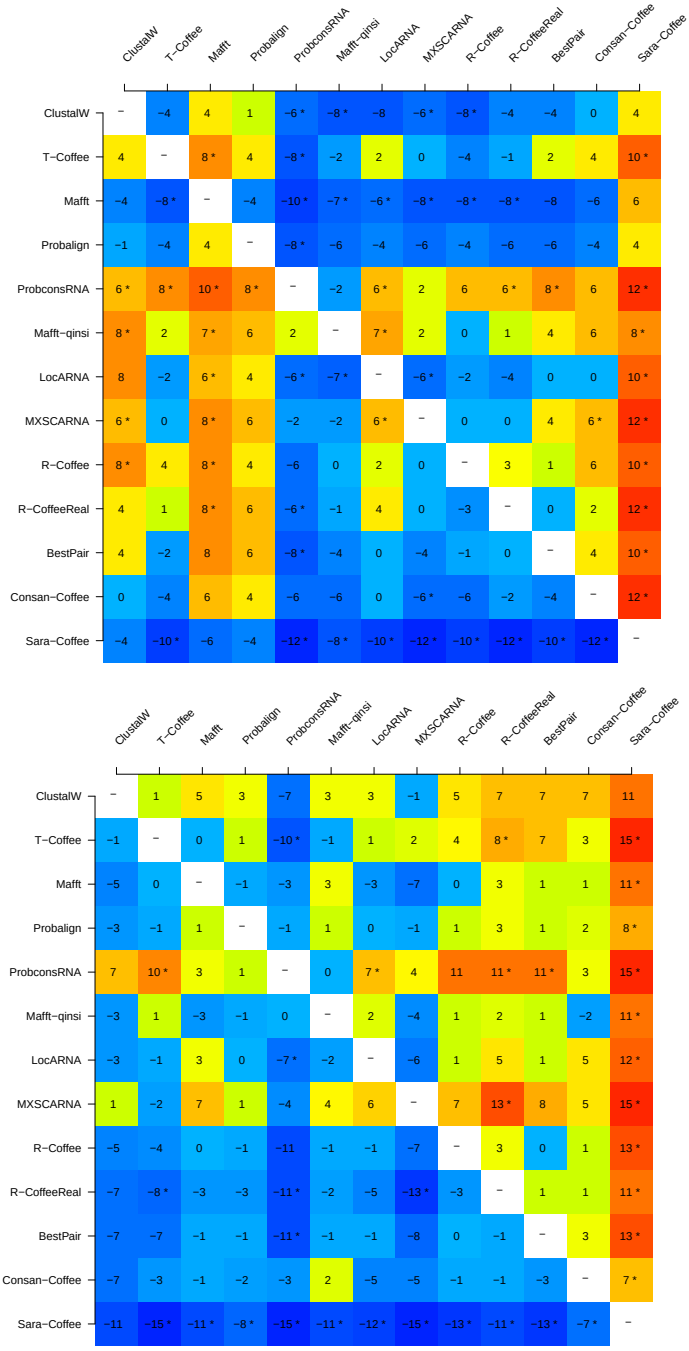
continued from previous page			
Cluster	% identity Sara-Coffee	# sequences	avg sequence length
26	35.86	5	18
27	49.33	3	31
28	72.12	43	24
29	51.05	3	157
30	55.56	4	20
31	32.81	3	27
32	62.16	5	30
33	91.39	6	23
34	65.23	5	36
35	85.91	3	59
36	99.21	12	75
37	72.07	48	122
38	85.17	7	81
39	33.04	3	36
40	71.93	5	61
41	54.76	7	30
AVG	58.87	11.9	44.2

Supplementary Table 3.2: The 31 clusters divided into two sets depending on the amount of structure included.

BP-index	Cluster Ids
<73	10, 28, 11, 21, 25, 32, 2, 26, 1, 27, 38, 16, 20, 40
≥ 73	17, 8, 37, 6, 15, 34, 7, 13, 41, 3, 19, 22, 5, 35, 9, 14, 4



Supplementary Figure 3.1: Number of times the upper method is winning against the horizontal method according to 3SP score. Stars denote a significant difference in the two distributions according to the Wilcoxon-Test with a p-value < 0.05. top: low-density secondary structures, bottom: high-density secondary structures.



Supplementary Figure 3.2: Number of times the upper method is winning against the horizontal method according to NiRMSD score. Stars denote a significant difference in the two distributions according to the Wilcoxon-Test with a p-value < 0.05. top: low-density secondary structures, bottom: high-density secondary structures.

4 Large scale alignment: KM-Coffee

Breen M S, Kemena C, Vlasov P K, Notredame C, Kondrashov F A, "Epistasis as the primary factor in molecular evolution", Nature, 2012 ; 490(7421): 535-8.

Breen M S, Kemena C, Vlasov P K, Notredame C, Kondrashov F A. Epistasis as the primary factor in molecular evolution. Supplementary information. Nature. 2012; 490(7421): 535-8.

5 Discussion

The development of new algorithms comes along with the need establish their accuracy as compared with existing algorithms. In the area of multiple sequence alignments, the method of choice is the usage of community accepted reference alignments like BALiBASE (Thompson et al., 2005) used as benchmarks. They provide an objective criterion to decide which alignment program performs best and allows the developers to evaluate the new program against a known standard. Some important problems exist when using such a procedure, the main one being the risk to over-fit an algorithm to a specific benchmark or a certain kind of data (Blackshields et al., 2006; Boulesteix, 2010). Hence, some benchmark sets like BALiBASE contain subsets addressing different problems to prevent over-fitting to specific data properties. Another problem is the experimental design of the benchmark. Some level of automation is usually needed, that often results in reference alignments which are not completely trustworthy. To address this issue, the correctness of a method is only evaluated on a certain subset or subregion of the models. For example in BALiBASE only certain columns are evaluated when using the associated scoring program. It is also sometimes unclear to which extend the reference dataset reflects accuracy. For instance, in the Homfam benchmark (Sievers et al., 2011), each dataset contains on average 10 reference sequences coming along with more than 10,000 homologues. Such a setup will be highly sensitive to the order in which the sequences are being aligned and this order may be a strong confounding factor, making it hard to determine if the variations in accuracy result from true sequence matching improvements, or from subtle variation in the handling of the reference sequences. Moreover, reference alignment benchmarks only determine the program which is best on average but they do not predict which alignment program will perform best on a specific data set. Thus although very useful, reference benchmarks have limits and these limits need to be kept in mind when developing alignment tools or deciding which alignment program to use for a specific data set. Aniba et al. (2010) proposed with AlexSys a method to address this last problem. They try to predict, in dependence of certain sequence features, which

alignment methods will perform best on a given dataset. They show that using AlexSys decreases the running time with only a small decrease in the score compared to the best performing program (ProbCons) on the used data set.

With STRIKE we tried to address the same problem specifically for protein alignments using structural information. The goal of STRIKE is to detect the most accurate alignment among a set of alternative alignments of the same set of sequences. Although some methods exist, e.g. the sum-of-pairs score or the NiRMSD, they are not often used in practice as they have several drawbacks. The sum-of-pairs score performs a simple sequence similarity evaluation, therefore the results are only accurate for highly similar sequences. By contrast, the NiRMSD measures structural accuracy and allows distinguishing between alignments of sequences with very low similarity. This metrics, however, requires at least two structures and ideally one structure for each sequence, an unrealistic requirement in most cases. The STRIKE score tries to get the best of both metrics. It uses a single structure only, therefore it can be used on a much larger set of sequences while at the same time still being structural informative. Simultaneously, the projection of the structure over the whole alignment permits the evaluation of a much larger fraction of the alignment compared to the NiRMSD when given less than the full set of structures. STRIKE enables a user to evaluate his specificdataseton different alignment methods and chooses the one with the highest accuracy to be used in further steps. This is, with the growing number of protein structures in the PDB database, a realistic scenario. Our benchmarking shows that STRIKE is an accurate method and we believe that it will be useful for the community.

Another project where the evaluation of alignments played an important role was the SARA-Coffee project during which a new multiple RNA structural aligner was developed. The difficulty for the evaluation was that no benchmarking set for RNA tertiary structure alignment exists. Instead of providing reference alignments, which would introduce circularity, we decided to use objective criteria to avoid the problem of constructing our own reference alignments. We applied two different, albeit related, metrics to measure the accuracy of an alignment. The first one measures the agreement in the secondary structure alignment, a property widely used for accuracy estimation in RNA alignments. The second one compares the superposition of the three dimensional structures. Although these measurements are correlated because the secondary structure strongly influences the tertiary structure, we showed that differences can arise especially if no dense secondary structure exists. Additionally we demonstrated that our method performs very well on

both metrics, unfortunately the limited number of available RNA structures (currently less than 1000 structures are published in the PDB) limits is applicability. But as this number is steadily growing and new methods are developed to deduce the structural conformation of RNAs (e.g. experiment informed prediction (Ding et al., 2012)) the usability of Sara-Coffee will increase as well.

While for RNA relatively few sequences and even less structures are known making structural alignments a difficult undertaking, in proteins this problem is not as pronounced and the method of structural alignment is common and several programs exist. However, the amount of sequences is still much higher than the number of known structures rendering this method useless for large datasets. The largest family in the Pfam database (COX1) for example has almost 290,000 sequences but only 34 of them have known structures. This large datasets are in general very difficult to align and currently only very few programs exist being able to do it at all. With growing number of sequences it gets more and more difficult to align the sequences accurately (Sievers et al., 2011) as mistakes at the beginning propagate through the whole alignment with each following alignment step. We tried to address this problems in two ways, the first one is to adapt the consistency approach to be able to align large number of sequences and on the other to propose a general method to trim the alignment according to gap occurrence. Usually columns which contain a high number of gaps are deleted because it is known that columns with a lot of gaps are less trustworthy. For large datasets we propose a different procedure, instead of deleting columns with a high gap number we delete the sequences from the set which introduce these gaps and realign the reduced set. This approach is not useful for small datasets because the sequences are often carefully selected but for large datasets the loss of information is minimal.

Of course not all current problems in the alignment field can be addressed by a single thesis. Among these are the genome alignment problem and the alignment uncertainty, which were already mentioned in the introduction. Open problems in genome alignments include for example the splitting of the genomes, which even with recent improvements is still a challenge especially with only partly assembled genomes consisting of a high number of short scaffolds. Current genome alignment methods cannot incorporate these scaffolds into blocks because they are not able to find a sufficient number of anchors in them. Thus a large part of the sequence information is not included in the homologous blocks.

The problem of alignment uncertainty is a problem which cannot be solved conceptually. As we do not know how evolution really proceeded, for alignments of divergent enough sequences we will never know how the true alignment would look like. Thus different models will always produce different alignments. As this cannot be really solved, other methods need to be used. Approaches identifying those alignment parts which are more and which are less trustworthy are a good starting point. However, most of the time this information is not really used in the following steps. First approaches have been undertaken to keep track of the uncertainty in the alignment process but research in this area is still at the beginning.

All in all, computing accurate alignments is a challenge and will stay one as long as alignments are being used.

6 Conclusion

The following points give a summary of the presented projects:

1. The STRIKE score developed during this thesis allows to estimate the structural accuracy of an alignment using a single experimentally derived three dimensional structure. This method permits to use different aligners to align a set of sequences and choose the best alignment according to the STRIKE score.

2. The structural RNA alignment project did not only include a new algorithm to calculate RNA structural alignments but as well a new benchmarking system for RNA structure alignments. The benchmark comes with two reference-independent accuracy measurements which will be useful for future alignment method testing.

3. KM-Coffee addressed the problem of aligning large data sets. Beside the algorithm that produces large-scale alignments, a method is presented to clean up the initial data set, which results in improved alignments.

Bibliography

Abhiman S, Daub CO, and Sonnhammer EL. Prediction of function divergence in protein families using the substitution rate variation parameter alpha. *Mol Biol Evol*, 23(7):1406–13, 2006.

Abraham M, Dror O, Nussinov R, and Wolfson HJ. Analysis and classification of RNA tertiary structures. *RNA*, 14(11):2274–2289, 2008.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

Aniba MR, Poch O, Marchler-Bauer A, and Thompson JD. AlexSys: a knowledge-based expert system for multiple sequence alignment construction and analysis. *Nucleic Acids Res*, 38(19):6338–6349, 2010.

Armougom F, Moretti S, Keduas V, and Notredame C. The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, 22(14):e35–9, 2006a.

Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, and Notredame C. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res*, 34(Web Server issue):W604–W608, 2006b.

Bates PA, Kelley LA, MacCallum RM, and Sternberg MJ. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, Suppl 5:39–46, 2001.

Battey JN, Kopp J, Bordo-li L, Read RJ, Clarke ND, and Schwede T. Automated server predictions in CASP7. *Proteins*, 69 Suppl 8:68–82, 2007.

Batzoglou S, Pachter L, Mesirov JP, Berger B, and Lander ES. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res*, 10(7):950–958, 2000.

Bauer M, Klau G, and Reinert K. Multiple structural RNA alignment with lagrangian relaxation. *Lecture Notes in Computer Science*, 3692:303–314, 2005.

Bauer R, Rother K, Moor P, Reinert K, Steinke T, Bujnicki J, and Preissner R. Fast Structural Alignment of Biomolecules Using a Hash Table, N-Grams and String Descriptors. *Algorithms*, 2(2):692–709, 2009.

Bernhart SH, Hofacker IL, and Stadler PF. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–5, 2006.

Blackshields G, Larkin M, Wallace IM, Wilm A, and Higgins DG. Fast embedding methods for clustering tens of thousands of sequences. *Comput Biol Chem*, 32(4):282–6, 2008.

Blackshields G, Wallace IM, Larkin M, and Higgins DG. Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biol*, 6(4):321–339, 2006.

Blanchette M, Ken JW, Riemer C, Elnitski L, Smit AF, M RK, Baertsch R, Rosenbloom K, et al. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research*, 14:708–715, 2004.

Boulesteix AL. Over-optimism in bioinformatics research. *Bioinformatics*, 26(3):437–439, 2010.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, and Pachter L. Fast statistical alignment. *PLoS Comput Biol*, 5(5):e1000392, 2009.

Bray N, Dubchak I, and Pachter L. AVID: A global alignment program. *Genome Res*, 13(1):97–102, 2003.

Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, and Willard HF. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71(3):527–42, 1992.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, and Batzoglou S. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*, 13(4):721–731, 2003a.

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, and Batzoglou S. Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, 19 Suppl 1:i54–i62, 2003b.

Burger L and van Nimwegen E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol*, 4:165, 2008.

Bussotti G, Raineri E, Erb I, Zytnicki M, Wilm A, Beaudoing E, Bucher P, and Notredame C. BlastR–fast and accurate database searches for non-coding RNAs. *Nucleic Acids Res.*, 39(16):6886–95, 2011.

Capella-Gutiérrez S, Silla-Martínez JM, and Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, 2009.

Capriotti E and Marti-Renom MA. Computational RNA structure prediction. *Current Bioinformatics*, 3:32–45, 2008a.

Capriotti E and Marti-Renom MA. RNA structure alignment by a unit-vector approach. *Bioinformatics*, 24(16):i112–8, 2008b.

Capriotti E and Marti-Renom MA. SARA: a server for function annotation of RNA structures. *Nucleic Acids Res*, 37(Web Server issue):W260–5, 2009.

Capriotti E and Marti-Renom MA. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics*, 11:322, 2010.

Chandonia JM, Kim SH, and Brenner SE. Target selection and deselection at the Berkeley Structural Genomics Center. *Proteins*, 62(2):356–70, 2006.

Chang YF, Huang YL, and Lu CL. SARSA: a web tool for structural alignment of RNA using a structural alphabet. *Nucleic Acids Res*, 36(Web Server issue):W19–24, 2008.

Chen X and Tompa M. Comparative assessment of methods for aligning multiple genome sequences. *Nature Biotechnology*, 28(6):567–572, 2010.

Claude JB, Suhre K, No-tredame C, Claverie JM, and Abergel C. CaspR: a web server for automated mo-lecular replacement using homology modelling. *Nucleic Acids Res*, 32(Web Server issue):W606–9., 2004.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, and Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*, 15(7):901–913, 2005.

Cuff JA and Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, 40(3):502–511, 2000.

Darling A, Mau B, Blattner F, and Perna N. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7):1394–1403, 2004.

Darling AE, Mau B, and Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, 5(6):e11147, 2010.

Darwin C. *On the Origin of Species by Means of Natural Selection*. Murray, London, 1859.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, and Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, 6(12):e1001025, 2010.

Dayhoff MO, Schwartz RM, and Orcutt BC. A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5(suppl 3):345–351, 1978.

Dewey CN. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol Biol*, 395:221–236, 2007.

Ding F, Lavender CA, Weeks KM, and Dokholyan NV. Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat Methods*, 9(6):603–608, 2012.

Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, and Dokholyan NV. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *Rna*, 14(6):1164–73, 2008.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

Do CB, Mahabhashyam MSP, Brudno M, and Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, 15(2):330–340, 2005.

Döring A, Weese D, Rausch T, and Reinert K. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9:11, 2008.

Doshi KJ, Cannone JJ, Cobaugh CW, and Gutell RR. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105, 2004.

Dowell RD and Eddy SR. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, 7:400, 2006.

Dror O, Nussinov R, and Wolfson H. ARTS: alignment of RNA tertiary structures. *Bioinformatics*, 21 Suppl 2:ii47–53, 2005.

Dubchak I, Poliakov A, Kislyuk A, and Brudno M. Multiple whole-genome alignments without a reference organism. *Genome Res*, 19(4):682–689, 2009.

Durbin R, Eddy SR, Krogh A, and Mitchison G. *Biological Sequence Analysis*. Cambridge University Press, 1998.

Eddy SR. Multiple alignment using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol*, 3:114–120, 1995.

Edgar RC. MUSCLE: a multiple sequence align-ment method with reduced time and space com-plexity. *BMC Bioinformatics*, 5:113, 2004a.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7. Print 2004., 2004b.

Edgar RC, Asimenos G, Batzoglou S, and Sidow A. Evolver. URL `http://www.drive5.com/evolver`.

Edgar RC and Batzoglou S. Multiple sequence alignment. *Curr Opin Struct Biol*, 16(3):368–73, 2006.

Erb I, González-Vallinas JR, Bussotti G, Blanco E, Eyras E, and Notredame C. Use of ChIP-Seq data for the design of a multiple promoter-alignment method. *Nucleic Acids Res*, 40(7):e52, 2012.

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, and Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci*, Chapter 2:Unit 2.9, 2007.

Fabian MA, Biggs r W H, Treiber DK, Atteridge CE, Azi-mioara MD, Benedetti MG, Carter TA, Ciceri P, et al. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotechnol*, 23(3):329–36, 2005.

Farazi TA, Juranek SA, and Tuschl T. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development*, 135(7):1201–14, 2008.

Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4):783–791, 1985.

Ferragina P, Giancarlo R, Greco V, Manzini G, and Valiente G. Compression-based classification of bio-logical sequences and structures via the Universal Similarity Metric: experimental assessment. *BMC Bioinformatics*, 8:252, 2007.

Ferre F, Ponty Y, Lorenz WA, and Clote P. DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Res*, 35(Web Server issue):W659–68, 2007.

Fitch W. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Zool*, 20:406–416, 1971.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, et al. Ensembl 2012. *Nucleic Acids Res*, 40(Database issue):D84–D90, 2012.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61, 2007.

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, et al. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*, 39(Database issue):D141–D145, 2011.

Gardner PP, Wilm A, and Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, 33(8):2433–9, 2005.

Göbel U, Sander C, Schneider R, and Valencia A. Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–317, 1994.

Gondro C and Kinghorn BP. A simple genetic algorithm for multiple sequence alignment. *Genet Mol Res*, 6(4):964–982, 2007.

Gotoh O. An improved algorithm for matching biological sequences. *J Mol Biol*, 162(3):705–708, 1982.

Gotoh O. Consistency of optimal sequence align-ments. *Bull. Math. Biol.*, 52:509–525, 1990.

Gotoh O. Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Itera-tive Refinements as Assessed by Reference to Structural Alignments. *J. Mol. Biol.*, 264(4):823–838, 1996.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, and Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33(Database issue):D121–4, 2005.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*, 59(3):307–321, 2010.

Guindon S and Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelih-ood. *Syst Biol*, 52(5):696–704, 2003.

Gutell RR and Fox GE. A compilation of large subunit RNA sequences presented in a structural format. *Nucleic Acids Res*, 16 Suppl:r175–269, 1988.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235):223–7, 2009.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res*, 22(9):1760–1774, 2012.

Henikoff S and Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919, 1992.

Hofacker IL, Fekete M, and Stadler PF. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–1066, 2002.

Hogeweg P and Hesper B. The alignment of sets of sequences and the construction of phylogenetic trees. An inte-grated method. *J. Mol. Evol.*, 20:175–186, 1984.

Holm L and Sander C. Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, 20:478–480, 1995.

Huelsenbeck JP and Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

Huynen MA, Stadler PF, and Fontana W. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc Natl Acad Sci U S A*, 93(1):397–401, 1996.

Karsch-Mizrachi I, Nakamura Y, and Cochrane G. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res*, 40(Database issue):D33–D37, 2012.

Katoh K, Kuma K, Toh H, and Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–8, 2005.

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

Kececioglu J. The maximum weight trace problem in multiple sequence alignment. In A Apostolico, M Crochemore, Z Galil, and U Manber, editors, *Combinatorial Pattern Matching*, volume 684 of *Lecture Notes in Computer Science*, chapter 9, pages 106–119. Springer-Verlag, Berlin/Heidelberg, 1993.

Kemena C and Notredame C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19):2455–2465, 2009.

Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, and Segal E. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–7, 2010.

Kim J and Ma J. PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. *Nucleic Acids Res.*, 39(15):6359–68, 2011.

Knudsen B and Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, 2003.

Kolodny R, Koehl P, and Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol*, 346(4):1173–88, 2005.

Kuhn RM, Haussler D, and Kent WJ. The UCSC genome browser and associated tools. *Brief Bioinform*, 2012.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, and Salzberg SL. Versatile and open software for comparing large genomes. *Genome Biol*, 5(2):R12, 2004.

Landan G and Graur D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*, 24(6):1380–1383, 2007.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–8, 2007.

Lassmann T and Sonnhammer EL. Automatic assessment of alignment quality. *Nucleic Acids Res*, 33(22):7120–8, 2005a.

Lassmann T and Sonnhammer EL. Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6:298, 2005b.

Lassmann T and Sonnhammer ELL. Quality assessment of multiple alignment programs. *FEBS Lett*, 529(1):126–130, 2002.

Lee C, Grasso C, and Sharlow MF. Multiple sequence alignment using par-tial order graphs. *Bioinformatics*, 18(3):452–64., 2002.

Lee RC, Feinbaum RL, and Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–54, 1993.

Lin K, Kleinjung J, Taylor WR, and Heringa J. Testing homology with Contact Accepted mutatiOn (CAO): a contact-based Markov model of protein evolution. *Comput Biol Chem*, 27(2):93–102, 2003.

Löytynoja A and Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A*, 102(30):10557–10562, 2005.

Loytynoja A and Goldman N. Phylogeny-aware gap placement pre-vents errors in sequence alignment and evolutio-nary analysis. *Science*, 320(5883):1632–5, 2008.

Lu XJ and Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*, 31(17):5108–21, 2003.

Markova-Raina P and Petrov D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res*, 21(6):863–874, 2011.

Massingham T and Goldman N. Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169(3):1753–1762, 2005.

McClure M, Vasi T, and Fitch W. Comparative analysis of multiple pro-tein-sequence alignmnent methods. *Molecular Biology Evolution*, 11(4):571–592, 1994.

Morgenstern B, Dress A, and Wener T. Multiple DNA and Protein sequence based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA*, 93:12098–12103, 1996.

Myers EW and Miller W. Optimal alignments in linear space. *Comput Appl Biosci*, 4(1):11–17, 1988.

Nakato R and Gotoh O. Cgaln: fast and space-efficient whole-genome alignment. *BMC Bioinformatics*, 11:224, 2010.

Needleman SB and Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, 1970.

Notredame C. Recent Evolutions of Multiple Sequence Alignment. *PLoS Comput Biol*, 3(8):e123, 2007.

Notredame C and Abergel C. *Bioinformatics and genomes: current perspectives*, chapter Using multiple alignment methods to assess the quality of genomic data analysis. Horizon Scientific Press, 2003.

Notredame C and Higgins DG. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res*, 24(8):1515–1524, 1996.

Notredame C, Higgins DG, and Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217, 2000.

Notredame C, O'Brien EA, and Higgins DG. RAGA: RNA sequence alignment by genetic algorithm. *Nucleic Acids Res*, 25(22):4570–80, 1997.

Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143(1):46–58, 2010.

O'Sullivan O, Suhre K, Abergel C, Higgins DG, and Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol*, 340(2):385–95, 2004.

O'Sullivan O, Zehnder M, Higgins D, Bucher P, Grosdidier A, and Notredame C. APDB: a novel measure for benchmark-ing sequence alignment methods without refer-ence alignments. *Bioinformatics*, 19 Suppl 1:i215–21, 2003.

Pascarella S, Milpetz F, and Argos P. A databank (3D-ali) collecting related protein sequences and structures. *Protein Eng*, 9(3):249–251, 1996.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, and Haussler D. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.*, 21(9):1512–28, 2011.

Paten B, Herrero J, Beal K, and Birney E. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, 25(3):295–301, 2009.

Paten B, Herrero J, Beal K, Fitzgerald S, and Birney E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome research*, 18(11):1814–1828, 2008.

Pazos F and Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 47(2):219–227, 2002.

Pei J. Multiple protein sequence alignment. *Curr Opin Struct Biol*, 18(3):382–6, 2008.

Pei J and Grishin NV. MUMMALS: multiple sequence align-ment improved by using hidden Markov models with local structural information. *Nucleic Acids Res*, 34(16):4364–74, 2006.

Pei J and Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, 2007.

Pei J, Kim BH, and Grishin NV. PROMALS3D: a tool for multiple pro-tein sequence and structure alignments. *Nucleic Acids Res*, 36(7):2295–300, 2008.

Pei J, Sadreyev R, and Grishin NV. PCMA: fast and accurate multiple se-quence alignment based on profile consistency. *Bioinformatics*, 19(3):427–8., 2003.

Penn O, Privman E, Landan G, Graur D, and Pupko T. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol*, 27(8):1759–1767, 2010.

Prakash A and Tompa M. Measuring the accuracy of genome-size multiple alignments. *Genome Biology*, 8(6):R124+, 2007.

Qi L, Rifai N, and Hu FB. Interleukin-6 receptor gene, plasma C-reactive protein, and diabetes risk in women. *Diabetes*, 58(1):275–278, 2009.

Rahrig RR, Leontis NB, and Zirbel CL. R3D Align: global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics*, 26(21):2689–97, 2010.

Raphael B, Zhi D, Tang H, and Pevzner P. A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Res*, 14(11):2336–2346, 2004.

Rausch T, Emde AK, Weese D, Doring A, Notredame C, and Reinert K. Segment-based multiple sequence alignment. *Bioinformatics*, 24(16):i187–92, 2008.

Reinert K, Lenhof H, Mutzel P, Melhorn K, and Kececioglu J. A branch-and-cut Algorithm for multiple sequence alignment. *Recomb97*, pages 241–249, 1997.

Riaz T, Yi W, and Li KB. A tabu search algorithm for post-processing multiple sequence alignment. *J Bioinform Comput Biol*, 3(1):145–56, 2005.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–23, 2007.

Rivas E and Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8, 2001.

Romanuik TL, Ueda T, Le N, Haile S, Yong TM, Thomson T, Vessella RL, and Sadar MD. Novel biomarkers for prostate cancer including noncoding transcripts. *Am J Pathol*, 175(6):2264–76, 2009.

Roshan U and Livesay DR. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, 22(22):2715–21, 2006.

Rost B. Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94, 1999.

Saitou N and Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, 1987.

Sankoff D. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.

Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO, Bartels D, Bekel T, et al. Complete genome sequence of the myxobacterium Sorangium cellulosum. *Nat Biotechnol*, 25(11):1281–1289, 2007.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, and Miller W. Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107, 2003.

Schwede T, Kopp J, Guex N, and Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res*, 31(13):3381–3385, 2003.

Shih ACC and Li WH. GS-Aligner: a novel tool for aligning genomic sequences using bit-level operations. *Mol Biol Evol*, 20(8):1299–1309, 2003.

Shindyalov IN and Bourne PE. Protein structure alignment by incre-mental combinatorial extension (CE) of the op-timal path. *Protein Eng*, 11(9):739–47, 1998.

Siebert S and Backofen R. MARNA: multiple alignment and con-sensus structure prediction of RNAs based on se-quence structure comparisons. *Bioinformatics*, 21(16):3352–9, 2005.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, 2005.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*, 7:539, 2011.

Simossis VA and Heringa J. PRALINE: a multiple sequence align-ment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res*, 33(Web Server issue):W289–94, 2005.

Siva N. 1000 Genomes project. *Nat Biotechnol*, 26(3):256, 2008.

Smith TF and Waterman MS. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, 1981.

Sokal RR and Michener CD. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.

Song J, Xu Y, White S, Miller KWP, and Wolinsky M. SNPsFinder–a web-based application for genome-wide discovery of single nucleotide polymorphisms in microbial genomes. *Bioinformatics*, 21(9):2083–2084, 2005.

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

Stebbings LA and Mizuguchi K. HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Da-tabase. *Nucleic Acids Res*, 32(Database issue):D203–7, 2004.

Stoye J, Evers D, and Meyer F. Generating benchmarks for multiple se-quence alignments and phylogenetic reconstruc-tions. *Ismb*, 5:303–6, 1997.

Stoye J, Evers D, and Meyer F. Rose: generating sequence families. *Bioinformatics*, 14(2):157–163, 1998.

Subramanian AR, Weyer-Menkhoff J, Kaufmann M, and Morgenstern B. DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, 6(1):66, 2005.

Tabei Y, Kiryu H, Kin T, and Asai K. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, 9:33, 2008.

Taylor WR. Identification of protein sequence ho-mology by consensus template alignment. *J. Mol. Biol.*, 188:233–258, 1986.

Taylor WR and Orengo CA. Protein structure alignment. *J Mol Biol*, 208(1):1–22, 1989.

The Chimpanzee Sequencing Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

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

Thompson JD, Koehl P, Ripp R, and Poch O. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1):127–136, 2005.

Thompson JD, Plewniak F, and Poch O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1):87–88, 1999.

Treangen TJ and Messeguer X. M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. *BMC Bioinformatics*, 7:433, 2006.

Van Walle I, Lasters I, and Wyns L. SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, 21(7):1267–8, 2005.

Vingron M and Argos P. Motif recognition and alignment for many sequences by comparison of dot-matrices. *J. Mol. Biol.*, 218:33–43, 1991.

Wallace IM, Blackshields G, and Higgins DG. Multiple sequence alignments. *Curr Opin Struct Biol*, 15(3):261–6, 2005a.

Wallace IM, O'Sullivan O, and Higgins DG. Evaluation of iterative alignment algo-rithms for multiple alignment. *Bioinformatics*, 21(8):1408–14, 2005b.

Wallace IM, O'Sullivan O, Higgins DG, and Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*, 34(6):1692–9, 2006.

Wan Y, Qu K, Ouyang Z, Kertesz M, Li J, Tibshirani R, Makino DL, Nutter RC, et al. Genome-wide Measurement of RNA Folding Energies. *Mol Cell*, 2012.

Wang L and Jiang T. On the complexity of multiple sequence alignment. *Journal of computational biology*, 1(4):337–348, 1994.

Weydt P, Soyal SM, Gellera C, Didonato S, Weidinger C, Oberkofler H, Landwehrmeyer GB, and Patsch W. The gene coding for PGC-1alpha modifies age at onset in Huntington's Disease. *Mol Neurodegener*, 4:3, 2009.

Wheeler TJ and Kececioglu JD. Multiple alignment by aligning align-ments. *Bioinformatics*, 23(13):i559–68, 2007.

Will S, Reiche K, Hofacker IL, Stadler PF, and Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4):e65, 2007.

Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, and Schultz PG. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science*, 309(5740):1570–3, 2005.

Wilm A, Higgins DG, and Notredame C. R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res*, 36(9):e52, 2008.

Wilm A, Mainz I, and Steger G. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol*, 1:19, 2006.

Wong KM, Suchard MA, and Huelsenbeck JP. Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–476, 2008.

Yunis JJ, Sawyer JR, and Dunham K. The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. *Science*, 208(4448):1145–1148, 1980.

Zhang Y and Waterman MS. An Eulerian path approach to global multiple alignment for DNA sequences. *J Comput Biol*, 10(6):803–819, 2003.

Zhou H and Zhou Y. SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, 21(18):3615–21, 2005.

Zuker M. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989.