Using multiple alignment methods to assess the quality of genomic data analysis

Cédric Notredame and Chantal Abergel

Information Génétique et Structurale

UMR 1889

31 Chemin Joseph Aiguier

13 006 Marseille

Email: cedric.notredame@igs.cnrs-mrs.fr, chantal.abergel@igs.cnrs-mrs.fr

ABSTRACT

The analysis of multiple sequence alignments can generate essential clues in genomic data analysis. Yet, to be informative such analyses require some mean of estimating the reliability of a multiple alignment. In this chapter we describe a novel method allowing the unambiguous identification of the residues correctly aligned within a multiple alignment. This method uses an index named CORE (Consistency of the Overall Residue Evaluation) based on the T-Coffee multiple sequence alignment algorithm. We provide two examples of applications: one where the CORE index is used to identify correct blocks within a difficult multiple alignment and another where the CORE index is used on genomic data to identify the proper start codon and a frame-shift within one of the sequence.

1 Introduction

Biological analysis largely relies on the assembly of elaborate models meant to summarize our knowledge of life complex mechanisms. For that purpose, vast amounts of data are collected, analyzed, validated and then integrated within a model. In an ideal world, an existing model would be available to explain every bit of experimental data. In the real world, this is rarely the case, and every day, existing models need to be modified to accommodate new findings. Sometimes, data that cannot be explained is kept at bay until the accumulation of new evidences prompts the design of an entirely new model. Unaccountable data can be viewed as the stuff inflating an inconsistency bubble. Eventually, the bubble bursts and a new model is designed.

A multiple alignment is nothing less than such a model. Given a series of sequences and an alignment criteria (structure similarity, common phylogenetic origin) the multiple alignment contains a series of hypothesis regarding the relationship between the sequences it is made of. This alignment can accommodate data generated experimentally (e.g. alignment of two homologous catalytic residues) or combine the results of various sequence analysis methods. The importance of the use of multiple sequence alignments in the context of sequence analysis has been recognized for a long time and it is so well established that most bioinformatics protocols make use of it. Multiple alignments have been turned into profiles (Gribskov et al., 1987) and hidden Markov models (Krogh et al., 1994) to enhance the sensitivity and the specificity of database searches (Altschul et al., 1997). State of the art methods for protein structure prediction depend on the proper assembly of a multiple

sequence alignment (Jones, 1999) as do phylogenetic analysis (Duret et al., 1994). Over the last years multiple sequence alignment techniques have been instrumental to improvements made in almost every key area of sequence analysis. Yet, despite its importance, the accurate assembly of a multiple sequence alignment is a complex process, the biological knowledge and the computational abilities it requires are far beyond our current capacities. As a consequence, biologists are left to use approximate programs that attempt to assemble proper alignments without providing any guaranty they may do so. The lack of a 'perfect' or at least reasonably robust method explains why so many multiple sequence alignment packages exist. The variations among these packages are not only cosmetic; they include the use of very different algorithms, different parameters and generally speaking different paradigms. For a recent review of state-of-the-art techniques, see (Duret and Abdeddaim, 2000).

Database searches, structure predictions, phylogenetic analysis are enough on their own to make multiple alignment compulsory in a genome analysis task. Yet, thanks to the sanity checks they provide, multiple alignments can also be instrumental at tackling the plague of genomic analysis: faulty data. When dealing with genomes, faulty data arises from two major sources: sequencing errors and wrong predictions. The consequence is that a predicted protein sequence may have accumulated errors both at the DNA level and when its frame was predicted (this will be especially true in eukaryotic genes where exons may be missed, added or improperly predicted). In the worst cases, the effect of such errors will be amplified in the high level analysis, leading to an improper analysis of the available data. On the other hand, once they have been identified, these errors are usually easily corrected either by extra sequencing or data extrapolation. Therefore, any method providing a reasonable sanity-check that earmarks areas of a genome likely to be problematic would be a major improvement. In this chapter we will show how multiple sequence alignments can be used to carry out part of

this task. For that purpose we will focus on the applications of T-Coffee, a recently described method (Notredame et al., 2000).

2 Generating Multiple Alignments With T-Coffee

Despite the large variety of multiple sequence alignment methods publicly available, the number of packages effectively used for data analysis is surprisingly small and a vast majority of the alignments found in the literature are produced using only two programs: ClustalW (Thompson et al., 1994) and its X-Window implementation ClustalX. ClustalW uses the progressive alignment strategy described by Taylor (Taylor, 1988) and Doolitle (Feng and Doolittle, 1987), refined in order to incorporate sequence weights and a local gap penalty scheme. Recently, the ClustalW algorithm was further modified in order to improve the accuracy of the produced alignments by making the evaluation of the substitution costs position dependant. This improved algorithm is implemented in the T-Coffee package (Notredame et al., 2000).

The aim of T-Coffee is to build a multiple alignment that has a high level of consistency with a library of pre-computed pair-wise alignments. This library may contain as many alignments as one wishes and it may also be redundant and inconsistent with itself. For instance it may contain several alternative alignments of the same sequences aligned using various gap penalties. It may also contain alternative alignments obtained by applying different methods onto the sequences. Overall, the library is a collection of alignments believed to be correct. Within this library, each alignment receives a weight that is an estimation of its biological likeliness (i.e. how much trust does one have in this alignment to be correct). For that

purpose, one may use any suitable criteria such as percent identity, P-Value estimation or any other appropriate method. The T-Coffee algorithm uses this library in order to compute the score for aligning two residues with one another in the multiple alignment. This score is named the extended weight because it requires an extension of the library. The extended weight takes into account the compatibility of the alignment of two residues with the rest of the alignments observed within the library, its derivation is extensively described in (Notredame et al., 2000). The principle is straightforward: in order to compute the extended weight associated with two residues R and S of two different sequences, one will consider whether when R is found aligned in the library with some residue X of a third sequence, S is also found aligned with that same residue X in another entry of the library. If that is the case, then the weight associated with R and S will be increased by the minimum of the two weights RX and SX. The final extended weight will be obtained when every possible X has been considered and the resulting contributions summed up. Although this operation seems to be very expensive from a computational point of view, its effective computational cost is kept low thanks to the scarceness of the primary library (i.e. for most pairs of residues RS, very few Xs need to be considered). In the end, a pair of residues is highly consistent (and has a high extended weight) if most of the other sequences contain at least one residue that is found aligned both to R and to S in two different pair-wise alignments. A key property of this weight extension procedure is to concentrate information: the extended score of RS incorporates some information coming from all the sequences in the set and not only from the two sequences contributing R and S.

The main advantage of the extended weights is that they can be used in place of a substitution matrix. While standard substitution matrices do not discriminate between two identical residues (e.g. all the cysteins are the same for a Pam (Dayhoff et al., 1979) or a Blosum

(Henikoff and Henikoff, 1992)), the extended weights are truly position specific and make it possible to discriminate between two identical residues that only differ by their positions. Once the library has been assembled (potential ways of assembling that library are described later) and the extended weights computed, T-Coffee closely follows the ClustalW procedure using the extended weights instead of a substitution matrix. The overall T-Coffee strategy is outlined in Figure 1. All the sequences are first aligned two by two, using dynamic programming (Needleman and Wunsch, 1970) and the extended library in place of a substitution matrix. The distance matrix thus obtained is then used to compute a neighbor-joining tree (Saitou and Nei, 1987). This tree guides the progressive assembly of a multiple sequence alignment: the two closest sequences are first aligned by normal dynamic programming using the extended weights to align the residues in the two sequences, no gap penalty is applied (because it has already been applied to generate the alignments contained in the library). This pair of sequences is then fixed and any gaps that have been introduced cannot be shifted later. Then the program aligns the next closest two sequences or adds a sequence to the existing alignment of the first two sequences, depending which is suggested by the guide tree. The procedure always joins the next two closest sequences or pre-aligned group of sequences. This continues until all the sequences have been aligned. To align two groups of pre-aligned sequences one uses the extended weights, as before, but taking the average library scores in each column of the existing alignments.

The key feature of T-Coffee is the freedom given to the user to build his own library following whatever protocol may seem appropriate. For this purpose, one may mix structural information with database results, knowledge-based information or pre-established collections of multiple alignments. It may also be necessary to explore a wide range of parameters given some computer package. A simple library format was designed to fit that purpose, it is shown

on Figure 2. A library is a straightforward ASCII file that contains a listing of every pair of aligned residue that needs to be described. Any knowledge-based information can easily be added manually to an automatically generated library or the other way round. This figure also shows clearly that the library can contain ambiguities and inconsistencies (i.e. two alignments possible for the first residue of Seq1 with Seq2). These ambiguities will be resolved while the alignment is being assembled, on the basis of the score given by the extended weights. The library does not need to contain a weight associated with each possible pair of residues. On the contrary, an ideal library only contains pairs that will effectively occur in the correct multiple alignment (i.e. N^2L pairs rather than N^2L^2 pairs). While this flexibility to design and assemble one's own library is a very desirable property, in practice it is also convenient to have a standard automatic protocol available. Such a protocol exists and is fully integrated within the T-Coffee package. It is ran with the default mode and does not require the user to be aware of T-Coffee underlying concepts (Library, extension, progressive alignment). This default protocol extensively described and validated in (Notredame et al., 2000) requires two distinct libraries to be compiled and combined within the primary library before the extension. The first one contains a ClustalW pair-wise alignment of each possible pair of sequence within the dataset. For that purpose, ClustalW (Thompson et al., 1994) is run using default parameters. This library is global because it is generated by aligning the sequences over their whole length (global alignments) using a linear space version of the Needleman and Wunsch algorithm (Needleman and Wunsch, 1970). The second library is local: for each possible pair of sequences, it contains the ten best non-overlapping local alignments as reported by the Lalign program (Huang and Miller, 1991) run with default parameters. In the local and the global libraries, each pair of residues found aligned is associated with a weight equal to the average level of identity within the alignment it came from. When a specific pair is found more than once, the weights associated with each occurrence are added. The main strength of

this protocol is to combine local and global information within a multiple alignment. The level of consistency within the library will depend on the nature of the sequences. For instance, if the sequences are very diverse, the requirement for long insertions/deletions will often cause the global alignments to be incorrect and inconsistent, while the local alignments will be less sensitive to that type of problems. In such a situation, the measure of consistence will enhance the local alignment signal and let it drive the multiple alignment assembly. Inversely, if the global alignments are good enough they will help removing the noise associated with the collection of local alignments (local alignments do not have any positional constraints). Overall, the current default T-Coffee protocol contains three distinct elements that lead to the collection of extended weights: the global library, the local library and the library extension that turns the sum of the two libraries into an extended library. Earlier work demonstrated that each of these components plays a significant part in improving the overall accuracy of the program. Table 1 shows that the current version of T-Coffee (Version 1.29) outperforms other popular multiple sequence alignment methods, as judged by comparison on BaliBase (Thompson et al., 1999), a database of hand made reference structural alignments that are based on structural comparison (See Table 1 legend for a description of BaliBase and the comparison protocol).

These results illustrate well the good performances of T-Coffee on the wide range of situations that occur in BaliBase. It is especially interesting to point out that T-Coffee is the only method equally well suited to situations that require a global alignment strategy (categories 1, 2 and 3) and situations that are better served with a local alignment strategy (categories 4 and 5 with long internal and terminal insertions/deletions). The other methods are either good for global alignments (like ClustalW) or for local alignments (like Dialign2 (Morgenstern et al., 1998)). It should be noted that T-Coffee still uses ClustalW 1.69 to

construct the primary global library, because this was the last 'naïve' version of ClustalW, not tuned up on BaliBase. The latest version (1.81) has been tuned on the BaliBase references (hence its improved performances over the results originally reported for ClustalW). Using this ClustalW 1.81 version when benchmarking T-Coffee would make the process circular. Nonetheless, as good as it may seem, T-Coffee still suffers from the same shortcoming as any other package available today: *it is not always the best method.* Even if on average it does better than any of its counterparts, one cannot guaranty that T-Coffee will always generate the best alignment. For instance, although Dialign2 is significantly less good, it T-Coffee outperforms on 17 test sets (11%). ClustalW is better than T-Coffee in 24% of the cases. We may conclude from this that in practice, there will always be situations where some alternative method beats T-Coffee. Furthermore, even in cases where the T-Coffee improvement over any alternative method is very significant, it may lead to an alignment much less than 100% correct. This may not be so helpful since for practical usage, it would be much more helpful to know where the correctly aligned portions lie. This is so true that a method 20% correct and a proper estimation of its reliability would be much more useful than a method more accurate 'on average'.

Several situations exist in which a biologist can make use of this reliability information. For instance, if the purpose of the alignment is to extrapolate some experimental data onto an otherwise un-characterized genomic sequence, one will need to be very careful not to deduce anything from an unreliable portion of the alignment. More generally, unreliable positions within a multiple sequence alignment should not be used for predictive purpose. For instance, when turning a multiple alignment into a profile in order to scan databases for remote homologues, it is essential to exclude regions whose alignment cannot be trusted and that may obscure some otherwise highly conserved position. Used in this fashion, reliability

information allows a significant decrease of the noise induced by locally spurious alignments. The other important application of a reliability measure is the identification of regions within a multiple alignment that are properly aligned without being highly conserved. These regions are extremely important when the alignment is used in conjunction with a predictive method that bases its analysis on mutation patterns. For instance, structure and phylogeny prediction methods require the presence of non-conserved positions to yield informative results. Any scheme that allows discriminating between positions that are degenerated but correctly aligned and positions that are simply misaligned may induce a dramatic improvement in the accuracy of these prediction methods. Furthermore a reliability measure will help identifying faulty data and provide some clues on how to correct it. In the next section, we show how consistency can be measured on a T-Coffee alignment and how that measure provides a fairly accurate reliability estimator.

3 Measuring The Consistency On A Multiple Sequence Alignment

T-Coffee is a heuristic algorithm that attempts to optimize the consistency between a multiple alignment and a list of pre-computed pair-wise alignments known as a library (Figure 2). By consistency we mean that a pair of residues described aligned in the library will also be found aligned in the multiple alignment. While the theoretical maximum for the consistency is 100%, the score of an optimal alignment will only be equal to the level of self-consistency within the library. Figure 2 shows the example of a library that is not self consistent because it

is ambiguous regarding the alignment of some of the residues it contains. Of course, the more ambiguous the library, the less consistency it will yield. For instance, given two residues *q* and *r* taken from two different sequences *S1* and *S2*, one can easily measure the consistency (CS(R_q^{S1}, R_r^{S2})) between the alignment of these two residues and all the other alignments contained in the library by comparing ES(R_q^{S1}, R_r^{S2}), the extended score of the pair *q* and *r*, with the sum of the extended scores of all the other potential pairs that involve *S1* and *S2* and either *r* or *q*.

$$\mathrm{CS}\left(R_q^{S1}, R_r^{S2}\right) = \mathrm{ES}\left(R_q^{S1}, R_r^{S2}\right) / \left\{ \sum_{z=S1} \mathrm{ES}\left(R_z^{S1}, R_r^{S2}\right) + \sum_{z=S2} \mathrm{ES}\left(R_q^{S1}, R_z^{S2}\right) \right\} \quad (1)$$

If we want to use it as a quality factor, this measure suffers from two major drawbacks. Firstly it is expensive to compute: given a multiple alignment of N sequences and of length L, each pair of residues found in the multiple alignment needs O(L) operations of extension that require a minimum of O(N) operations each. "O(L)" is a standard notation called *big-O notation*, meaning that the computation time is proportional to L, up to a constant factor. Since there are $L*N^2$ pairs of residues in a multiple alignment, this leads to $O(L^2N^3)$ operations for an estimate of the CS of every pair. This cubic complexity becomes problematic with large numbers of sequences. The second limitation of this measure is that with sequences rich in repeats, the summation factor can become artificially high and cause a dramatic decrease of the consistency score. In practice, we found it much more effective to use the extended score of the best scoring pair contained in the alignment as a normalization factor. This defines the aCS (approximate Consistency measure).

$$\mathrm{aCS}\left(R_q^{S1}, R_r^{S2}\right) = \mathrm{ES}\left(R_q^{S1}, R_r^{S2}\right) / Max\left\{ES\left(R_m^{Sx}, R_n^{Sy}\right)\right\} \quad (2)$$

With R_m^{Sx}, R_n^{Sy} any two residues found aligned in the multiple alignment.

Our measurements on the BaliBase dataset indicate that the CS and the aCS are well correlated.

An important criteria, when using the aCS as a reliability measure, is its ability to discriminate between correct and incorrect alignments within the so-called twilight zone (Sander and Schneider, 1991). Given two sequences, the twilight zone is a range of percent identity (between 0 and 30%) that has been shown to be non-informative regarding the relationship that exist among two sequences. Two sequences whose alignment yields less than 30% identity can either be unrelated or related and incorrectly aligned or related and perfectly aligned. To check how good the aCS is when used as an accuracy measure, every 142 BaliBase dataset was aligned using T-Coffee 1.29 and the similarity of each pair of sequence was measured within the obtained alignments. Pairs of sequences with less than 30% identity (5088) were extracted and the accuracy of their alignment was assessed by comparison with their counterparts in the reference BaliBase alignment, the aCS score was also assessed on each pair of aligned residues and averaged along the sequences. Figure 3a shows the scattered graph Identity Vs Accuracy (See Figure legend for the definitions). Despite a weak correlation between these two measurements, the percent identity is a poor predictor of the alignment accuracy. For 75% of the sequence pairs (identity lower than 25%) the accuracy indication given by the percent identity falls in a 40% range (i.e. the average identity indicates the average accuracy +/- 20%). On the other hand, when the accuracy is plotted against the aCS score (Figure 3b) the correlation is improved and for pairs of sequences having an aCS higher than 20 (this is true for 60% of the 5088 pairs) this measure is a much better alignment accuracy predictor than the percent identity. While they do not solve the twilight zone

problem, these results indicate that the aCS measure provides us with a powerful mean of assessing an alignment reliability within the twilight zone. Nonetheless, from a practical point of view, the aCS measure is not so useful since one is often more concerned by the overall quality (i.e. is residue r of sequence S correctly aligned to the rest of the sequences?) than by pair-wise relationships. In order to answer this type of questions, the aCS measure was used to derive three very useful non pair-wise indexes.

The Consistency of the Overall Residue Evaluation (CORE) is obtained by averaging the scores of each of the aligned pairs involving a residue within a column.

$$\mathrm{CORE}\left(R_q^{Sx}\right) = \sum_{y=1, y\neq x}^{N} aCS\left(R_q^{Sx}, R_r^{Sy}\right) / (N-1) \quad (3)$$

Where *q* and *r* are two residues found aligned in the same column.

The CORE index and equivalent approaches have been shown in the literature to be good indicators of the local quality of a multiple sequence alignment (Heringa, 1999; Notredame et al., 1998), as judged by comparison with reference biological alignments. In the T-Coffee package, an option makes it possible to output multiple alignments with the CORE index (a rounded value between 0 and 9) replacing each residue. It is also possible to produce a colorized version (pdf, postscript or html) of that same multiple alignment where residues receive a background coloration proportional to their CORE index (blue/green for low scoring residues and orange/red for the high scoring ones). Such an output is shown on Figure 5 and 6.

The CORE index described in equation (3) is merely an average aCS measure, and whether that measure provides some indication on the multiple alignment quality is a key question. We tested that hypothesis on the complete BaliBase dataset. Given each T-Coffee alignment, residues were divided in 4 categories: (i) *true positives* (TP) are correctly aligned residues rightfully predicted to be so, (ii) *true negative* (TN) are incorrectly aligned residues rightfully predicted to be so, (iii) *false positive* (FP) are residues predicted to be well aligned when they are not, (iv) *false negative* (FN) are residues wrongly predicted to be misaligned. Following previously described definitions (Notredame et al., 1998), a residue is said to be correctly aligned if at least 50% of the residues to which it was aligned in the reference alignment are found in the same column in the T-Coffee alignment. Each of the 10 CORE indexes (0 to 9) was used in turn as threshold to discriminate correctly and non-correctly aligned residues on the T-Coffee alignments. The BaliBase reference alignments were then used to evaluate the TP, TN, FP and FN. Sensitivity and the specificity were then computed according to Sneath and Sokal (Sneath and Sokal, 1973) and plotted on a graph (Figure 4). Our results indicate that the best trade off between sensitivity and specificity is obtained when CORE=3 is used as a threshold (i.e. every residue with a score higher or equal to 3 is considered to be properly aligned). In that case the specificity is 84% and the sensitivity is 82%. These high figures partly reflect the overall quality of the T-Coffee alignments in which 80.5% of the residues are correctly aligned according to the criteria used here. It is therefore more interesting to note that when the CORE index reaches 7, the specificity is 97.7% and the sensitivity is close to 50%. This means that thanks to the CORE index, half of the residues properly aligned in a multiple alignment can unambiguously be identified (e.g. more than 40 % of all the residues contained in BaliBase). In the next section we will see that this proper identification sometimes occurs in cases that are far from being trivial, even for an expert eye. Similar results were observed when applying the CORE index on multiple alignments obtained using

another method (i.e. ClustalW alignments evaluated with a standard T-Coffee library). This suggests that the CORE measure may be used to evaluate the local quality of a multiple alignment produced by any source. However, one should be well aware that the relevance of the CORE measure regarding the reliability of an alignment is entirely dependant on the way in which the library was derived. All the conclusions drawn here only apply to libraries derived using the standard T-Coffee protocol.

The sequence CORE (sCORE) is obtained by averaging the CORE scores over all the residues contained within one sequence.

$$\text{sCORE}(Sx) = \sum_{q=1}^{L} \left(ROC\left(R_q^{Sx}\right) \right) / L \quad (4)$$

That measure can be helpful for identifying among the sequences an outlier, a sequence that should not be part of the set either because it is not homologous or because it is too distantly related to the other members to yield an informative alignment.

The alignment CORE (alCORE) may be obtained by averaging the sCOREs over all the sequences. Our analysis suggest that this index may be a reasonable indicator of the alignment overall accuracy. Yet, to be fully informative, it requires the sequence set to be homogenous (i.e. the standard deviation of the sCOREs should be as low as possible).

4 Using the CORE Measure To Assess Local Alignment Quality.

The driving force behind the development of the CORE index is the identification of correctly aligned blocks of residues within a multiple sequence alignment. It is common practice to identify these blocks by scanning the multiple alignment and marking highly conserved regions as potentially meaningful. ClustalW and ClustalX provide a measure of conservation that may help the user when carrying out this task. Unfortunately, situations exist where it is difficult to make a decision regarding the correct alignment of some residues within an alignment. Such an example is provided in Figure 5 with the BaliBase alignment known as 1pamA_ref1, made of 6 alpha-amylases.

This set is difficult to align because it contains highly divergent sequences. Not only have these sequences accumulated mutations while they were diverging but they have also undergone many insertion/deletions that make it difficult to reconstruct their relationships with accuracy. The average level of identity measured on the BaliBase reference is 18%, the two closest sequences being less than 20% identical. As such, 1pamA_ref1 constitutes a classic example of a test set deceptive for most multiple sequence alignment methods. The fact that less than one third of the 1pam_ref1 reference alignment is annotated as trustable in BaliBase confirms that suspicion. When ran on this test-set, existing alignment programs generate different results, Prrp finds 37% of the columns annotated as trustable in BaliBase, ClustalW (1.81) 40%, T-Coffee 54% and Dialign2 56%. Regardless of the methods used, such an alignment is completely useless unless correctly aligned portions can be identified. It is exactly the information that the CORE index provides us with. An alignment colorized according to its CORE indexes is shown on Figure 5.

The results are in good agreement with those reported in Figure 4. Out of the 905 correctly aligned residues (42% of the total), 267 have a score higher than 7. No incorrectly aligned residue has a score higher or equal to 7. Using 7 as a prediction threshold gives a sensitivity of 29% and a specificity of 100%. Residues with a CORE index of 3 or higher (yellow pale) yield a sensitivity of 65% and a specificity of 79%. In this alignment, the main features are the red/dark-orange blocks: they are 100% correct. These blocks could be fed as they are to any suitable method (structure prediction, phylogeny….). They are not very well conserved at the sequence level and are therefore very informative for structural and phylogenetic analysis. For instance, the block II in Figure 5 is perfectly aligned although within that block, the average pair-wise identity is lower than 30% (41 % for the two most closely related sequences). The measure of consistency can also help questioning positions that may seem unambiguous from a sequence point of view. In the column annotated as I, the position marked with a "*" could easily be mistaken to be correct: it is within a block, aromatic positions are usually fairly well conserved and owing to their relative rarity, unlikely to occur by chance. Yet the green color code indicates that this position may be incorrectly aligned (the green tyrosine has a CORE index of 1). This is confirmed by comparison with the reference that shows the correct alignment to incorporate another tyrosine at this position.

When analyzing these patterns, one should always keep in mind that the consistency information only has a positive value. In other words, inconsistent regions are those where the library does not support the alignment. This does not mean they are incorrectly aligned but rather that no information is at hand to support or disprove the observed alignment.

5 Identifying Faulty Gene Predictions

Another possible application of the T-coffee CORE index is to reveal and help resolving sequence ambiguities in predicted genes. In the structural genomic era, many projects involve hypothetical proteins, for which an accurate prediction of the start and stop codon is needed to properly express the gene product. Since over-predicted N or C-terminal are rarely conserved at the amino acid level, sequence comparison provides us with a very powerful mean of identifying this type of problems. A simple procedure consists of multiply aligning the most conserved members of a protein family before measuring the T-Coffee CORE index on the resulting alignment. Inspection of the CORE patterns offers a diagnostic regarding the correctness of the data. This approach can also be applied to frame-shift detection where the identification of abnormally low scoring segments may lead to their correction. Such an alignment will make it possible to decide if the abnormal length of a coding region could be due to a sequencing error (and the resulting frame-shift). At least the CORE measure will indicate that a thorough examination is needed. Of course, one could also detect these frame-shifts using standard pair-wise comparison methods such as Gene-wise (Birney and Durbin, 2000), but the advantage of using a multiple sequence alignment is that the simultaneous comparison of several sequences can strengthen the evidence that the frame-shift is real. Furthermore, thanks to the multiple alignment, one may be able to detect mistakes in sequences that lack a very close homologue.

To illustrate this potential usage of T-Coffee, we chose the example of an *Escherichi coli K12* gene (Accession # U00096) predicted to encode a protein of unknown function, yifB. Orthologous genes were found in complete genomes using BLAST (Altschul et al., 1997) and the four most conserved sequences (identity >70% relative to the *Escherichia coli K12* gene,

see figure for ID numbers) were retrieved along with their flanking regions (80 nucleotides on the N-terminus side) in order to check whether these supposedly non coding regions did not contain any coding information. The 'elongated' sequences were translated in the same frame as their core coding region, their multiple alignment was carried out using T-Coffee and the CORE indexes were measured. The resulting alignment is displayed on Figure 6 with the CORE indexes color-coded (low CORE in blue and green, high CORE in orange and red). The main feature on the N-terminus is an abrupt transition (II) from low to high CORE indexes. This position is also a conserved methionine. The combination of these two observations suggests that the starting point of these five sequences is probably where the transition occurs, ruling out other methionines as potential starting points in the first sequence (I). Another discrepancy occurs in this alignment that is also emphasized by the CORE analysis: the sequence yifB_SALTY_1 yields a very low N-terminal CORE index, relatively to the other family members. The CORE score of this sequence becomes abruptly in phase with the other sequences at the position marked III. This pattern is a clear indication of a frame-shift: a protein highly similar to the other members of its family but locally unrelated. To verify that hypothesis, we used some data provided by SwissProt (Bairoch and Boeckmann, 1992) and found that in the corresponding entry, the nucleotide sequence has been corrected to remove the frame-shift we observed (entry P57015). The corrected sequence has been added to the bottom of the alignment on Figure 6 (non-colored sequence). The position where yifB_SALTY_1 and its corrected version start agreeing is also the position where the CORE score changes abruptly from a value of 2 (yellow) to a value of 7 (orange). That position also turns out to be the one where the frame-shift occurs in the genomic sequence.

6 Conclusion

In this chapter, we introduced an extension of the T-Coffee multiple sequence alignment method: the CORE index. The CORE index is a mean of assessing the local reliability of a multiple sequence alignment. Using the CORE index, correct blocks within a multiple sequence alignment can be identified. This measure also makes it possible to detect potential errors in genomic data, and to correct them. The CORE index is a relatively add hoc measure and even if it may prove extremely useful from a practical point of view, it still needs to be attached to a more theoretical framework. One would really need to be able to turn the consistency estimation into some sort of P-Value. For instance, to assess efficiently the local value of an alignment, one would like to ask questions of the following kind: what is the probability that library X was generated using dataset Y? What is the probability that alignment A yields p% consistency with library X? Altogether these questions may open more venues to the automatic processing of multiple alignments. That issue may prove crucial for the maintenance of resources that rely on a large scale usage of multiple sequence alignments such as Hobacgene (Perriere et al., 2000), Hovergene (Duret et al., 1994)or Prodom (Corpet et al., 2000).

Figure Legends

Figure 1) T-Coffee Outline

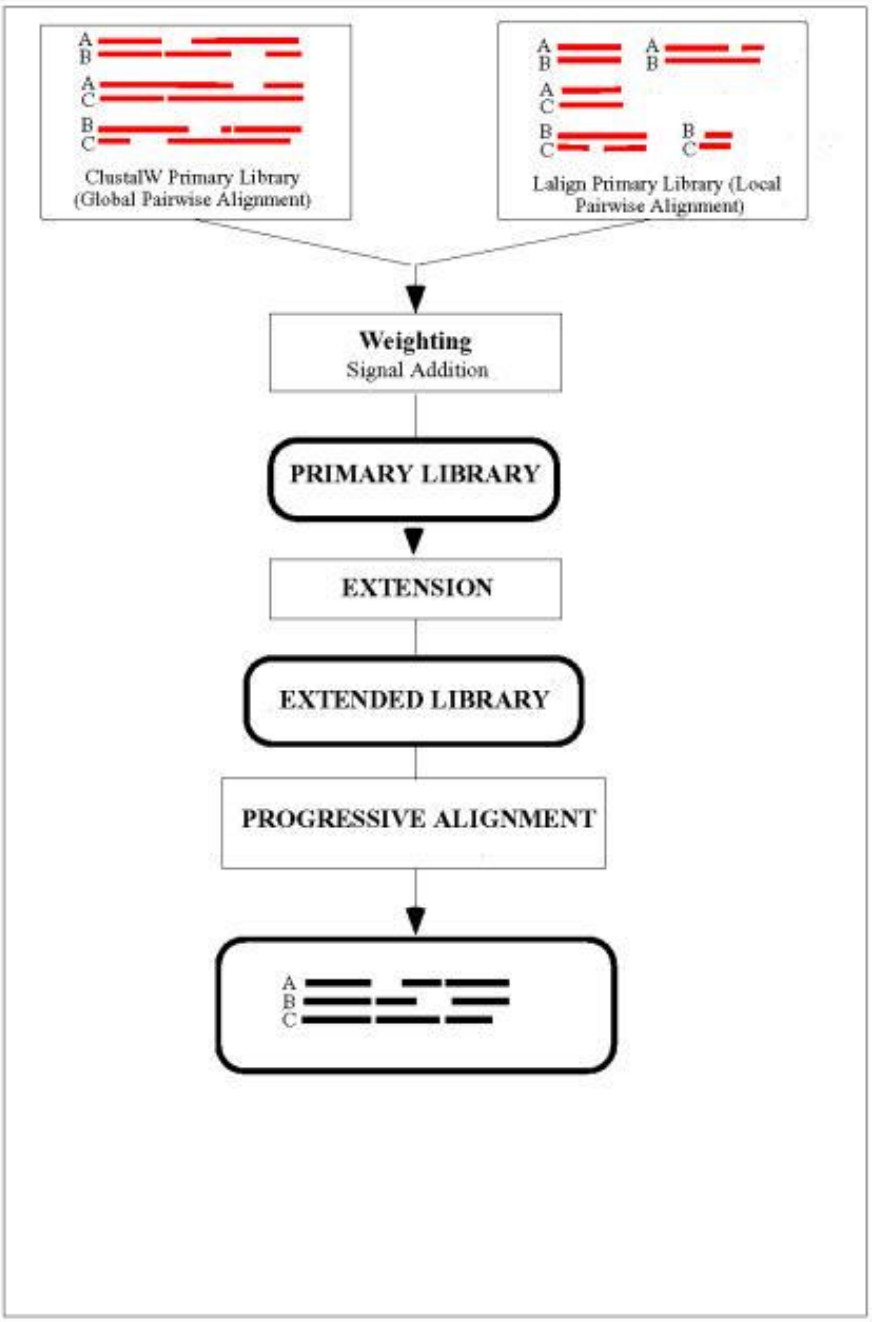

Figure 1

Layout of the T-Coffee algorithm

This figure indicates the chain of events that lead from unaligned sequences to a multiple sequence alignment using the T-Coffee algorithm. Data processing steps are boxed while data structures are indicated by rounded boxes.

Figure 2) Library Format

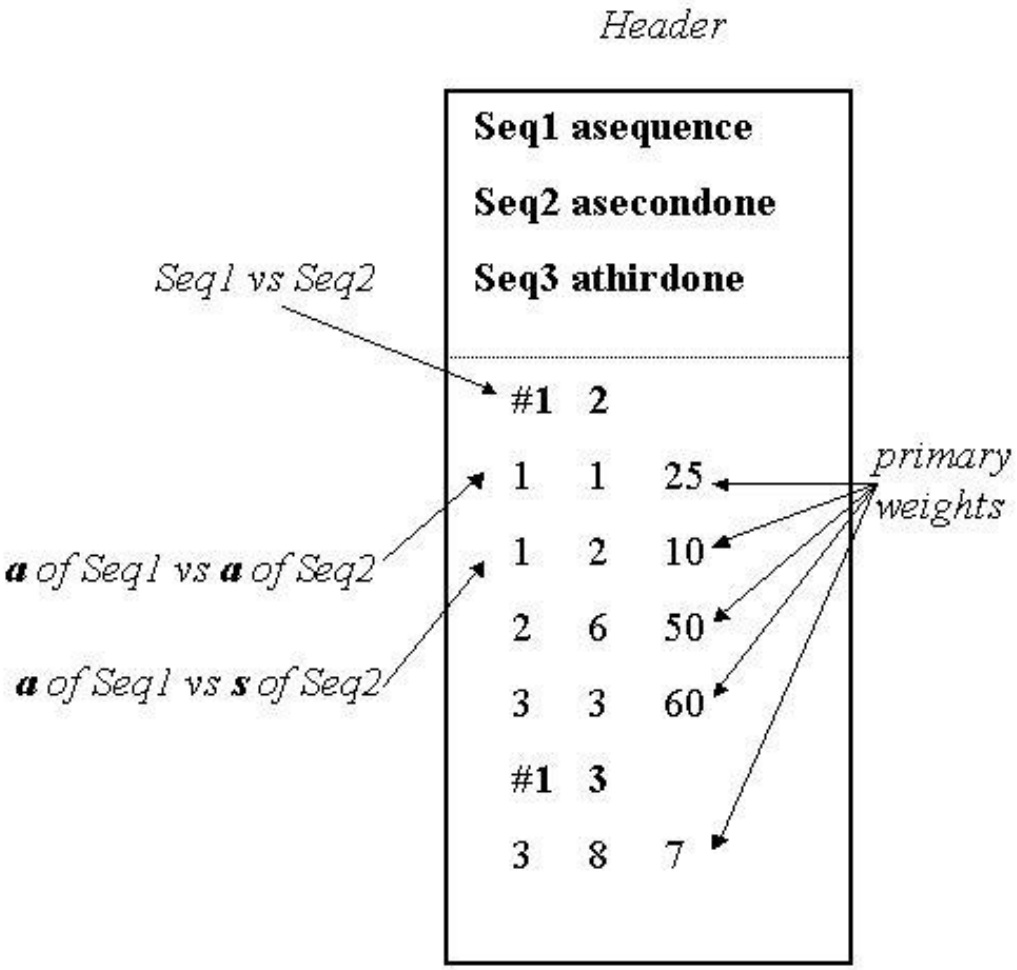

Figure 2

Library Format

An example of a library used by T-Coffee. The header contains the sequences and their names. ‘# 1 2’ indicates that the following pairs of residues will come from sequences 1 and 2. Each pair of aligned residues contains three values: the index of residue 1, the index of residue 2 and the weight associated with the alignment of these two residues. No order or consistency is expected within the library.

Figure 3a) Percentage identity Vs Accuracy in the twilight zone

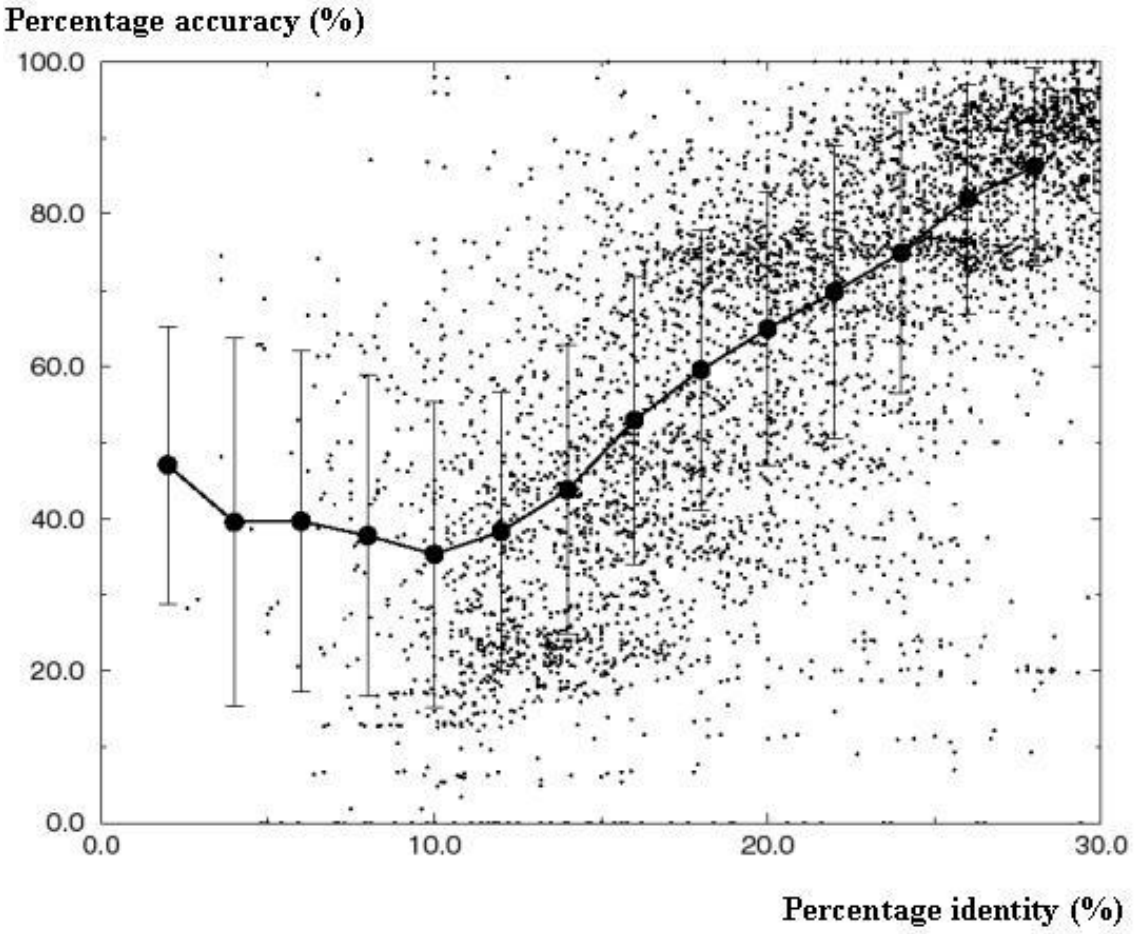

Figure 3

a) *Percentage identity Vs Accuracy in the twilight zone*: the 5088 pairs of sequences that have less than 30% identity in the BaliBase reference alignments were extracted. The accuracy of

their alignment was measured by comparison with the reference, and the resulting graph was plotted.

b) *Approximate Consistency Score (aCS) Vs Accuracy in the twilight zone*: the aCS was measured on the 5088 pairs of sequences previously considered and was plotted against the average accuracy previously reported. The vertical line indicates aCS=25 and separates the pairs for which the aCS is informative from those whose aCS seems to be non-informative.

Figure 3b) Approximate Consistency Score Vs Accuracy

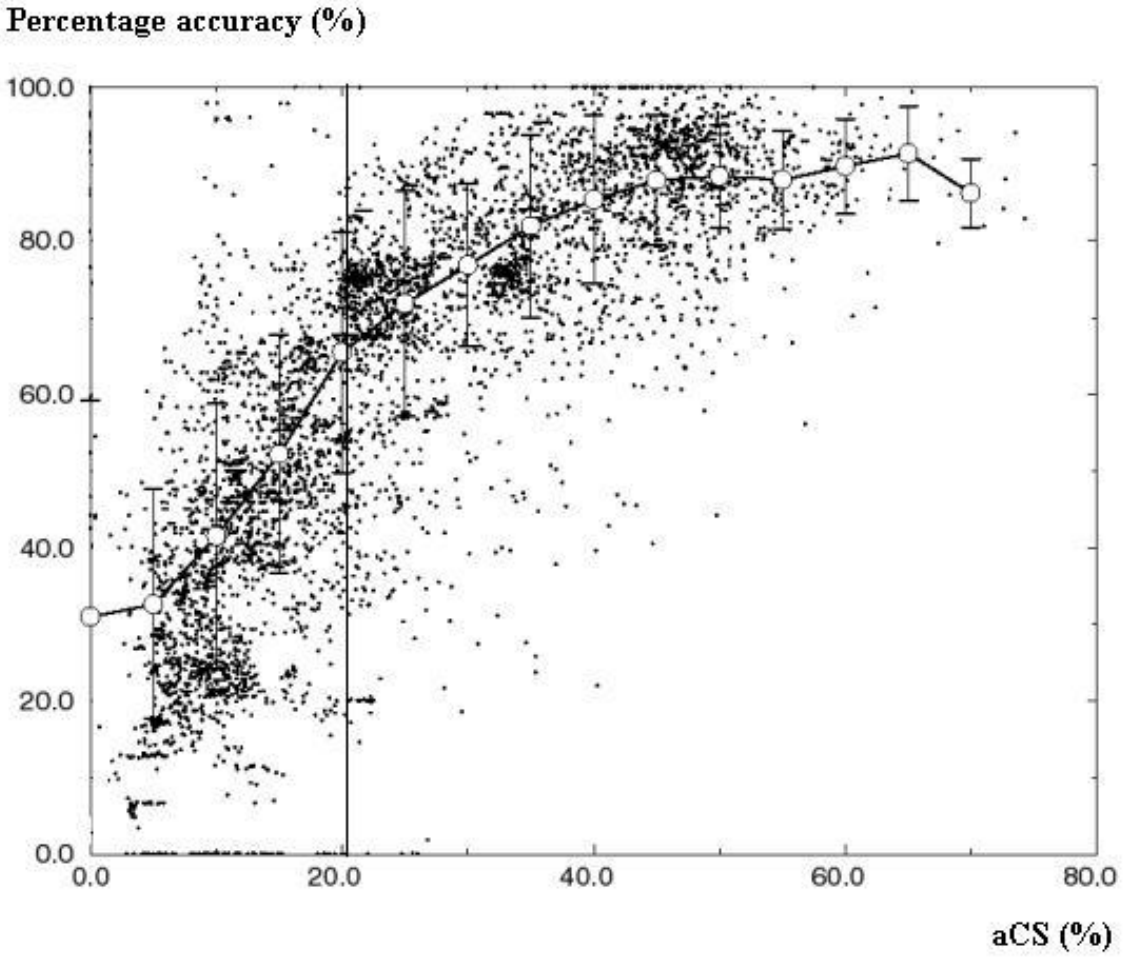

Figure 4) Specificity and Sensitivity of the CORE index

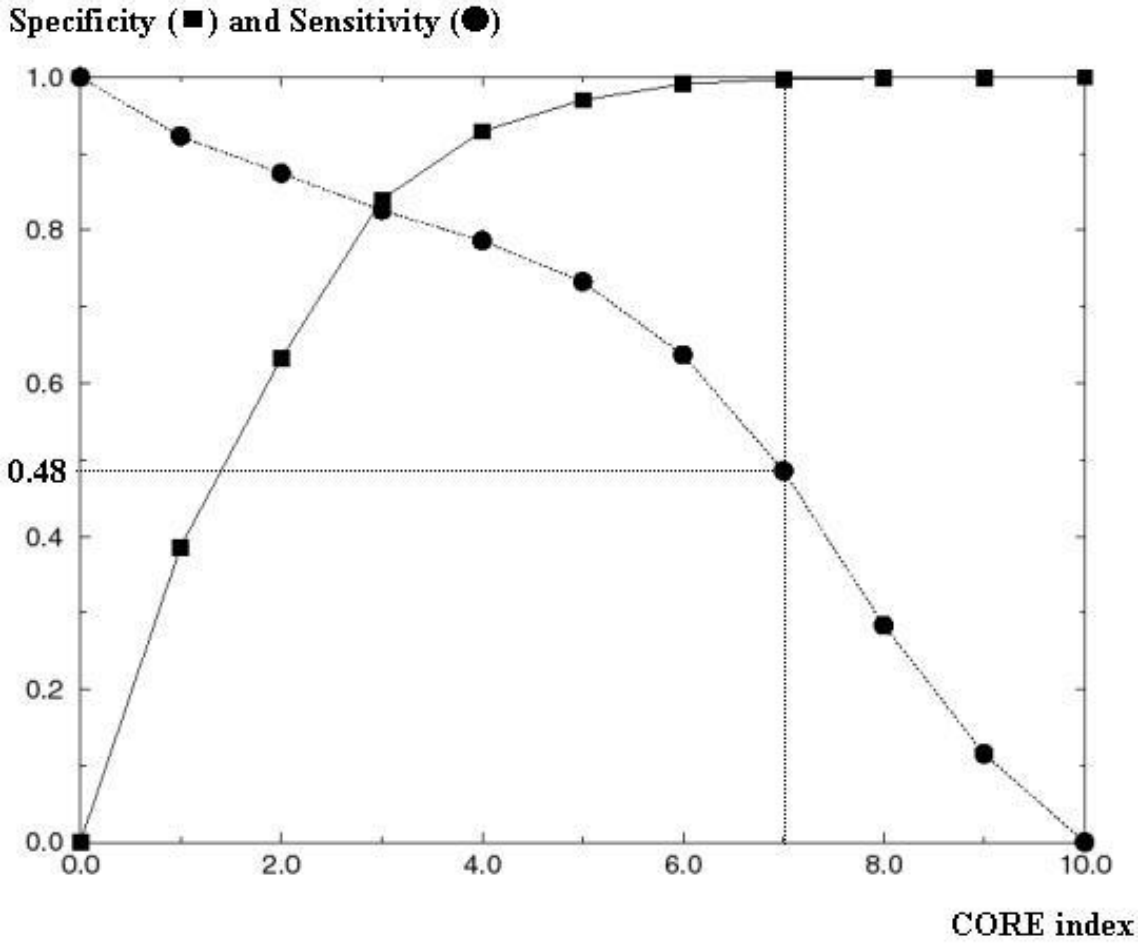

Figure 4

Specificity and Sensitivity of the CORE measure

The sensitivity and the specificity of the CORE index used as an alignment quality predictor were evaluated on the BaliBase test-sets. Measures were carried out on the entire BaliBase dataset. The sensitivity (●) and the specificity (■) were measured on the T-Coffee alignments after considering that every residue with a CORE index higher than x was properly aligned (see text for details).

Figure 5) Identifying correct blocks with the CORE index

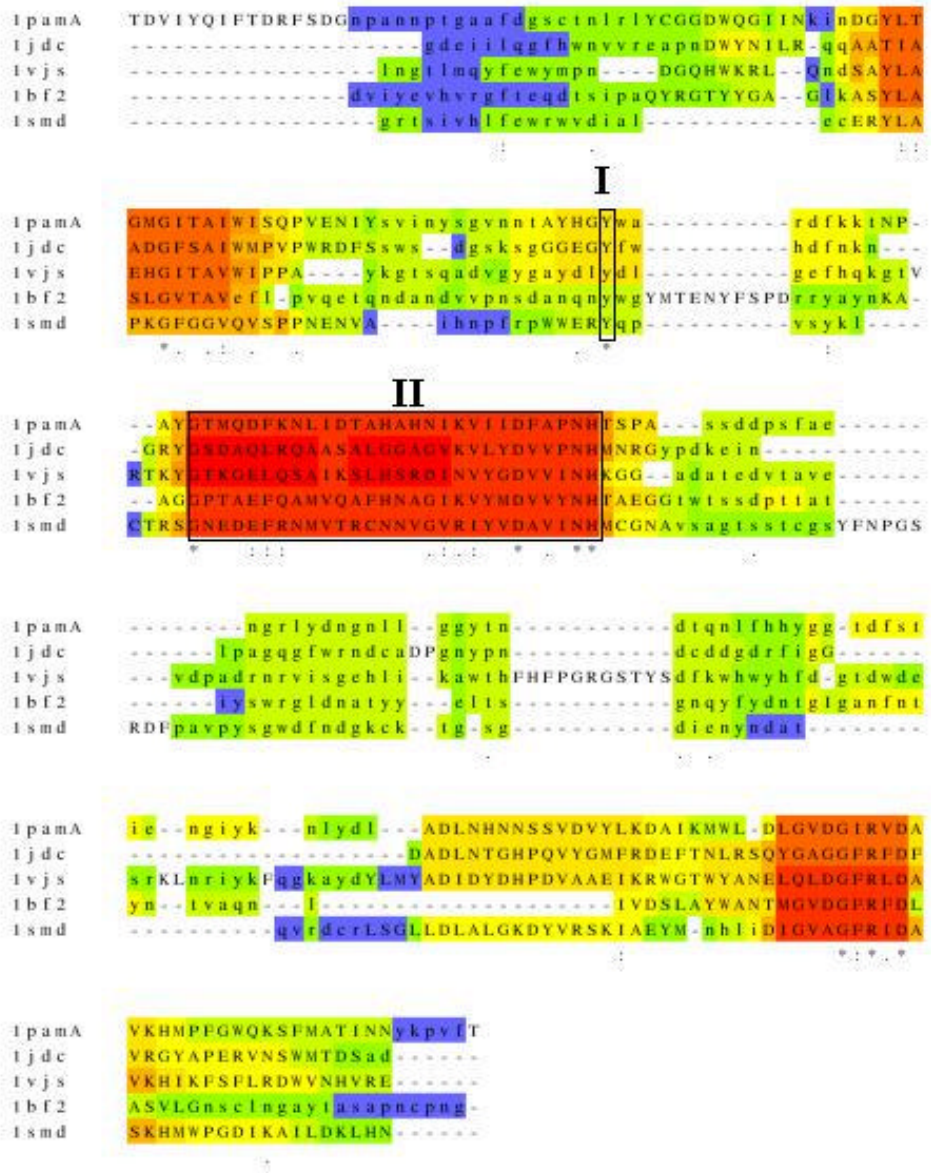

Figure 5

Identifying correct blocks with the CORE measure

An example of the T-Coffee output on a BaliBase test set (1pamA_ref1) that contains five alpha amylases. This alignment was produced using T-Coffee 1.29 with default parameters and requesting the score_pdf output option. The color scale goes from blue (CORE=0, bad) to red (CORE=9, good). The residues in capital are correctly aligned (as judged by comparison with the BaliBase reference). Those in lower case are improperly aligned. Box I indicates a conserved position that is not properly aligned, box II indicates a block of distantly related segments that is correctly aligned by T-Coffee.

Figure 6) Identifying frameshifts and start codons

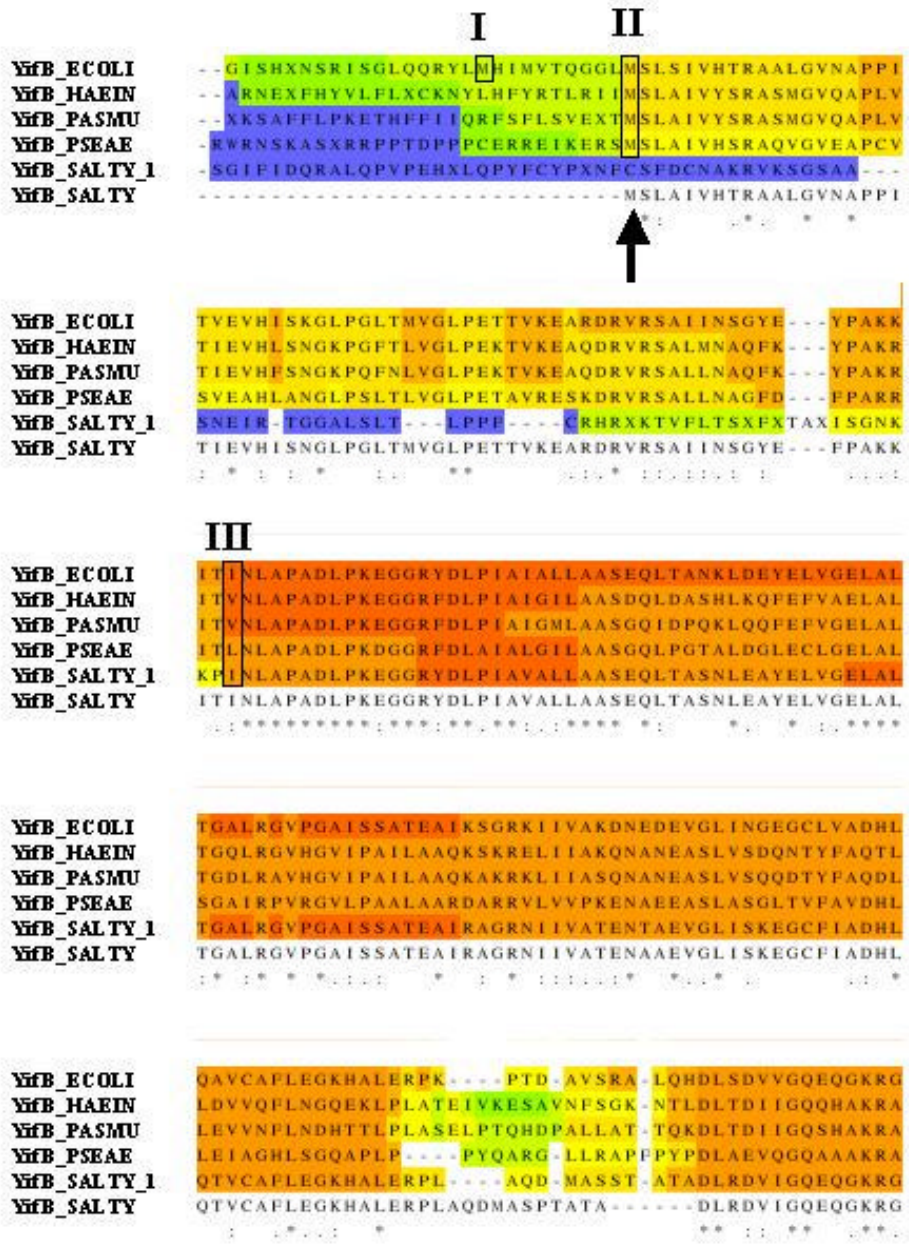

Figure 6

Identifying frame shifts and start codons

The chosen sequences came are YifB_ECOLIA *(Escherichia coli*, accession # AE005174), YifB_SALTY_1 (*Salmonella tiphy*, C18 chromosome, Sanger Center), YifB_HAIN (*Haemophilus influenzae,* Accession # L42023), YifB_PASMU (*Pasteurella multocida*, Accession # AE004439) and YifB_PSEAE (*Pseudomonas aeruginosa*, Accession # AE004091), they were aligned using the standard T-Coffee alignment procedure and requesting the score_pdf output option. The corrected sequence of *Salmonella tiphy* YifB protein sequence was later added for further reference (YifB_SALTY, SP: P57015) but it was not used for coloring the residues (Non colored sequence) or improving the multiple alignment. The figure only shows the N-terminal portion of the alignment, and the arrow indicates the positions annotated as starting codons in SwissProt (except for salmonella tiphy). Box I indicates a putative starting codon in YifB_ECOLIA, Box II indicates the true starting codon in most sequences, and Box III indicates the position where the frame-shift occurs in YifB_SALTY_1.

Table 1

	cat 1	cat 2	cat 3	cat 4	cat 5	avg 1	avg 2
cw	79.53	32.91	48.72	74.02	67.84	67.89	61.82
prrp	78.62	32.45	50.14	51.12	82.72	66.45	60.25
dialign2	70.99	25.21	35.12	74.66	80.38	61.54	57.99
T-Coffee	**80.67**	**36.15**	**53.20**	**83.41**	**91.69**	**72.18**	**69.55**

To produce this table each dataset contained in BaliBase was aligned using one of the methods (cw: ClustalW 1.81 (Thompson et al., 1994), Prrp (Gotoh, 1996), dialign2 (Morgenstern et al., 1998) and T-Coffee 1.29 (Notredame et al., 2000). In BaliBase, reference alignments are classified in 5 categories: category 1 contains closely related sequences, category 2 contains a group of closely related sequences and an outsider category 3 contains two groups of sequences that are distantly related, category 4 contains families with long internal indels, Category 5 contains sequences with long terminal indels. The resulting alignments were then compared to their reference counterpart in BaliBase, only using the regions annotated as trustable in BaliBase. Under this scheme we define the accuracy of an alignment to be the percentage of columns that are found totally conserved in the reference divided by the total number of columns within that reference. The comparison is restricted to the portions annotated as trustworthy in the reference alignment. *avg 1* is the average of the results obtained on each of the 142 test cases, *avg 2* is the average of the results obtained in each category. Bold numbers indicate the best performing method.

Bibliography

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. nucleic acids res. 25: 2289-3402.

Bairoch, A. and Boeckmann, B., 1992. The SWISS-PROT protein sequence data bank. Nucleic Acids Res: 2019-2022.

Birney, E. and Durbin, R., 2000. Using GeneWise in the Drosophila annotation experiment. Genome Res 10: 547-548.

Corpet, F., Servant, F., Gouzy, J. and Kahn, D., 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. nucleic acids res 28: 267-269.

Dayhoff, M.O., Schwarz, R.M. and Orcutt, B.C., 1979. A model of evolutionary change in proteins. Detecting distant relationships: computer methods and results. In: M.O. Dayhoff (Editor), Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, D.C., pp. 353-358.

Duret, L. and Abdeddaim, S., 2000. Multiple Alignment for Structural, Functional, or phylogenetic analyses of Homologous Sequences. In: D. Higgins and W. Taylor (Editors), Bioinformatics, Sequence, structure and databanks. Oxford University Press, Oxford.

Duret, L., Mouchiroud, D. and Gouy, M., 1994. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res. 22: 2360-2365.

Feng, D.-F. and Doolittle, R.F., 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of Molecular Evolution 25: 351-360.

Gotoh, O., 1996. Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinements as Assessed by Reference to Structural Alignments. J. Mol. Biol. 264: 823-838.

Gribskov, M., McLachlan, M. and Eisenberg, D., 1987. Profile analysis: Detection of distantly related proteins. Proceedings of the National Academy of Sciences 84: 4355-5358.

Henikoff, S. and Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. 89: 10915-10919.

Heringa, J., 1999. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. Computers and Chemistry 23: 341-364.

Huang, X. and Miller, W., 1991. A time-efficient, linear-space local similarity algorithm. Adv. Appl. Math. 12: 337-357.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195-202.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K. and Haussler, D., 1994. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. J. Mol. Biol. 235: 1501-1531.

Morgenstern, B., Frech, K., Dress, A. and Werner, T., 1998. DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics 14: 290-294.

Needleman, S.B. and Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48: 443-453.

Notredame, C., Higgins, D.G. and Heringa, J., 2000. T-Coffee: A novel algorithm for multiple sequence alignment. J. Mol. Biol. 302: 205-217.

Notredame, C., Holm, L. and Higgins, D.G., 1998. COFFEE: an objective function for multiple sequence alignments. Bioinformatics 14: 407-422.

Perriere, G., Duret, L. and Gouy, M., 2000. HOBACGEN: database system for comparative genomics in bacteria. Genome Research 10: 379-385.

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

Sander, C. and Schneider, R., 1991. Database of homology-derived structures and the structurally meaning of sequence alignment. Proteins: Structure, Function, and Genetics 9: 56-68.

Sneath, P.H.A. and Sokal, R.R., 1973. Numerical Taxonomy. Freeman, W.H., San Francisco.

Taylor, W.R., 1988. A flexible method to align large numbers of biological sequences. Journal of Molecular Evolution 28: 161-169.

Thompson, J., Higgins, D. and Gibson, T., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4690.

Thompson, J.D., Plewniak, F. and Poch, O., 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27: 2682-2690.