S2 Appendix: Partial local multiple alignment with Paloma-D

We detail here the procedure used to build the partial local multiple alignment (PLMA) \cite{1} of the ADAMTS-TSL and ADAM sequence dataset-214 contained in the fasta file \texttt{representatives.fasta} by the command \texttt{paloma-d -i \texttt{representatives.fasta} -M 20 -t 10}. An article with a complete presentation of Paloma-D is in preparation, the description will be limited here to this set of options.

Computation of candidate diagonal set. In its initial step, Paloma-D builds a set \( C \) of candidate pairwise local alignments between all the pairs of sequences to align. With the given command line options, \( C \) contains all the diagonals, i.e. gap-free pairs of sequence segments of equal length, considered for alignment by DIALIGN 2 \cite{2} of length smaller or equal to 20 residues (\(-M 20\)) and of weight greater or equal to 10 (\(-t 10\)). The weight \( w(D) \) of a diagonal \( D \) of length \( l \) and similarity score \( s \) (\( s \) is here the sum of the BLOSUM62 \cite{3} scores of the \( l \) facing residues in the diagonal) is defined in DIALIGN 2’s paper \cite{2} as the negative natural logarithm of the probability \( P_2(l,s) \) of finding any diagonal of length \( l \) of score at least as large as \( s \) somewhere within the comparison matrix of two random sequences of the same length as the original sequences. An advantage of this weight definition is that it enables to compare the similarity of diagonals of different lengths and to keep only those whose similarity is significantly higher than by chance. By default, DIALIGN 2 considers all diagonals of weight greater than 0, which is the expected weight of a random diagonal \( (P_2(l,s) = 1) \). We use here the more stringent parameters recommended by the Protomata suite for tractable and safe identification of strongly conserved domains \cite{4}. Indeed, the recommended weight threshold of 10 keeps only strongly conserved diagonals. This is the maximal weight thresholds recommended by the Protomata suite, corresponding to probabilities \( P_2(l,s) \) lower than \( \approx 5 \times 10^{-5} \). For comparison, other thresholds recommended by the Protomata suite are 1, 3 and 5 corresponding respectively to \( P_2(l,s) \) values lower than approx. \( 4 \times 10^{-1} \), \( 5 \times 10^{-2} \), \( 6 \times 10^{-3} \). The maximal length parameter is not as critical since it does not limit the size of identifiable domains since, thanks to the local extension introduced below, the diagonals can be chained to identify conservations longer than the lengths of diagonals used. We kept here the recommended value of 20 residues that was experimentally observed to provide a good signal/tractability trade-off for the identification of domains in the Protomata suite.

Selection and integration of diagonals. The core loop of Paloma-D integrates iteratively compatible diagonals from \( C \) to the set of PLMA local alignments \( \mathcal{L} \) which is here a set of diagonals. With the given options, each candidate diagonal from \( C \) is added in turn to \( \mathcal{L} \), as well as the diagonals induced by transitivity and extension of local alignments, if and only if the induced alignment is consistent \cite{5}. Otherwise, the candidate is incompatible with previously added diagonals and is discarded. Let us specify which are the induced diagonals with the help of Fig. 1. The transitive closure of the local alignments consists in adding automatically the diagonal \( T = (o_1, o_3) \) to \( \mathcal{L} \) whenever there exists \( D = (d_1, d_2) \) and \( E = (e_2, e_3) \) in \( \mathcal{L} \) such that \( d_2 \) and \( e_2 \) overlap on a segment \( o_2 \) aligned to \( o_1 \) by \( D \) and to \( o_3 \) by \( E \). This operation performs the consistent closure of pairwise local alignments to multiple sequences. Note however...
that \((o_1, o_3)\) is not required to belong to \(C\) since the experimental permissive mode of Paloma is now activated by default in Paloma-D to fasten the search. The extension of local alignment, for its part, enables longer local alignments to be built from smaller ones: it consists in replacing any pair of diagonals \(U = (u_1o_1, u_2o_2)\) and \(V = (o_1v_1, o_2v_2)\) from \(L\), with a segment overlap \(o_1\) on one sequence and \(o_2\) on the other sequence, by the diagonal \(L = (u_1o_1v_1, u_2o_2v_2)\) covering them.

In this greedy approach, the ordering of the candidates is important to establish a preference between them. In Paloma-D, the default ordering aims at favoring diagonals whose segments are conserved at most in most of the sequences. It can also be seen as a heuristic aiming at maximizing the final alignment score, defined here as the sum of weights of the diagonals in \(L\). More precisely, let us denote by \(X\) the set of sequences to align and by \(\text{Segments}(S)\) the set of segments of a sequence \(S\). For each diagonal \(D\) in \(C\) aligning two segments \(d_1\) and \(d_2\), we define for each sequence \(S\):

\[
d_{\text{max}}^S(d_1, d_2) = \arg\max_{\{d \in \text{Segments}(S) \setminus \{d_1, d_2\} : (d_1, d) \in C, (d_2, d) \in C\}} \left( w(d_1, d) + w(d_2, d) \right).
\]

If the set \(d_{\text{max}}^S(d_1, d_2)\) is not empty, its elements are each a segment of the sequence \(S\) that would maximize the sum of the weights if aligned to \(d_1\) and \(d_2\) with diagonals from \(C\). We denote by \(d_{\text{max}}^S(d_1, d_2)[1]\) the first element of this set with respect to an arbitrary order. Considering all the sequences, we can now compute the maximal sum of weights contribution of \((d_1, d_2)\) to the score of a local alignment extending it optimally to the other sequences with diagonals pairs of \(C\) by:

\[
W_{\text{max}}(d_1, d_2) = w(d_1, d_2) + \sum_{S \in X : d_{\text{max}}^S(d_1, d_2) \neq \emptyset} w(d_1, d_{\text{max}}^S(d_1, d_2)[1]) + w(d_2, d_{\text{max}}^S(d_1, d_2)[1]).
\]

\(W_{\text{max}}(d_1, d_2)\) is an estimation based on \(C\) of the contribution that \((d_1, d_2)\) can make to the global score of \(L\). It is the default heuristic score used in Paloma-D to sort by decreasing values the diagonals from \(C\) considered for addition in \(L\).

**Alignment blocks.** Once the set of diagonals \(L\) is built, the final step consists in identifying the set of the alignment blocks \(B\) which are the maximal multiple gapless alignment units induced by \(L\), aligning each a set of segments conserved in, and only in, a specific subset of sequences from \(X\). An alignment block can be formally defined here as a set of aligned segments such that 1) the block aligns each pair of its segments on their full length and this alignment is a restriction to these segments of a diagonal from \(L\) 2) there exists no diagonal in \(L\) aligning a position of a segment of the block to another position than those to which the position is aligned in the block 3) the set of aligned segments is maximal in the sense that there exists no other set of aligned segments including it that satisfies the two previous conditions.

The different steps of this alignment procedure are summarized by Algorithm 1 and an example of partial local multiple alignment graph showing the alignment blocks built from a set of candidate diagonals on the sequences is presented in Fig 2.
Fig 1. Induced diagonals. Left: If diagonals $D$ and $E$ overlap on a segment $o_2$ which is aligned to a segment $o_1$ by $D$ and to a segment $o_3$ by $E$, then the diagonal $T = (o_1, o_3)$ is added to $\mathcal{L}$ by the transitive closure of diagonals. Right: If diagonals $U$ and $V$ overlap on the segments $o_1$ and $o_2$, it implies a local alignment $L$ covering both diagonals which is added to $\mathcal{L}$ in replacement of $U$ and $V$ by the local extension of diagonals.

**Algorithm 1** Build partial local multiple alignment

Input: set of sequences $\mathcal{X}$

Segments $\leftarrow \{(S, b, e) : S \in \mathcal{X}, 1 \leq b \leq e \leq |S|, e - b + 1 \leq 20\}$  \hspace{1cm} \triangleright 20 is the chosen maximal length

$\mathcal{C} \leftarrow \{(d_1, d_2) \in \text{Segments} \times \text{Segments} : w(d_1, d_2) \geq 10\}$  \hspace{1cm} \triangleright 10 is the chosen weight threshold

$\mathcal{L} \leftarrow \emptyset$

for $(d_1, d_2) \in \mathcal{C}$ by decreasing value of $W_{\text{max}}(d_1, d_2)$ do

if $\mathcal{L} \cup \{(d_1, d_2)\}$ induces a consistent alignment then

$\mathcal{L} \leftarrow $ transitive closure and local extension of $\mathcal{L} \cup \{(d_1, d_2)\}$  \hspace{1cm} \triangleright$\triangleright$ Compatibility test

end if

end for

$\mathcal{B} \leftarrow$ alignment blocks defined by $\mathcal{L}$ on $\mathcal{X}$

return $(\mathcal{L}, \mathcal{B})$

Fig 2. Partial local multiple alignment built from a set of diagonals. In this schematic example, $S_1 . . . S_5$ are the sequences to align, $C_1 . . . C_7$ are the candidate diagonals successfully added by Paloma-D (examples of discarded diagonals $C_8$ and $C_9$, respectively incompatible with $C_1$ and with $\{C_3, C_7\}$, are also shown), $T_1$ and $T_2$ are the diagonals added by transitive closure of respectively $\{C_2, C_4\}$ and $\{C_3, C_7\}$, $L_1$ is the diagonal replacing $C_5$ and $C_6$ by local extension and $B_1 . . . B_6$ are the resulting alignment blocks. We see here that blocks can be shorter than the candidate diagonals (see for instance $B_5$) or longer (see for instance $B_2$).
References


