Dynamic Programming Algorithms for Biobjective Sequence Alignment

M. Abassi[1], L. Paquete[1], M. Pinheiro[2] and P. Matias[1]
1-CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
maryam/paquete@dei.uc.pt; pamatias@student.dei.uc.pt
2-School of Medicine, University of St Andrews, UK; mmp2@st-andrews.ac.uk

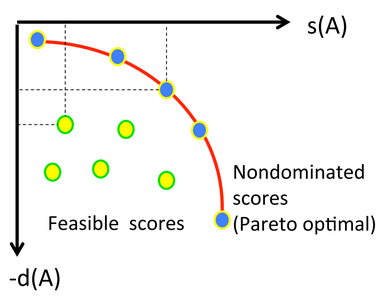
Introduction

Given an alignment A :

- s(A) : the substitution score of A obtained by a substitution matrix (e.g., BLOSUM or PAM).
- d(A) : the number of indels in A
- g(A) : the number of gaps in A

The goal

- To find all the alignments A that are "maximal" with respect to these score vector functions [1]:
  - VSD := (s(A) , -d(A))
  - VSG := (s(A) , -g(A))
- Alignment A dominates A′ (VSD(A) >VSD(A′)) iff
  - s (A) ≥ s (A′)
  - d (A) ≤ d (A′)
  - VSD(A) ≠ VSD(A′)



- An alignment A is Pareto optimal (PO) if there exists no other alignment A′ such that VSD(A′) > VSD(A)

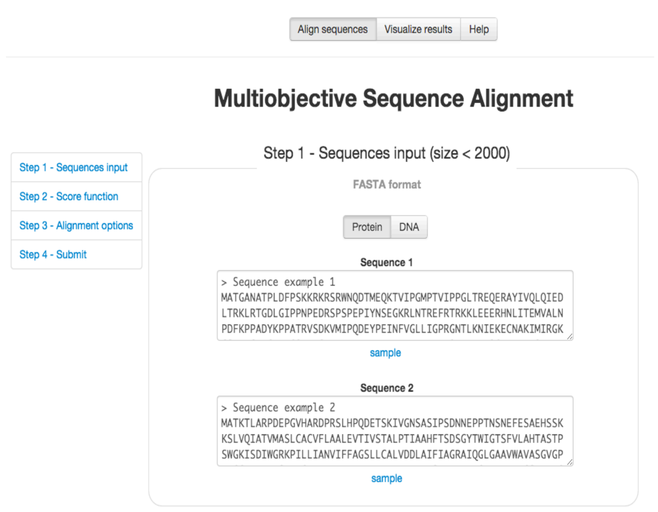
Our interest is to design algorithms to find the set of all PO alignments.

MOSAL – Software Tools for Multiobjective Sequence Alignment

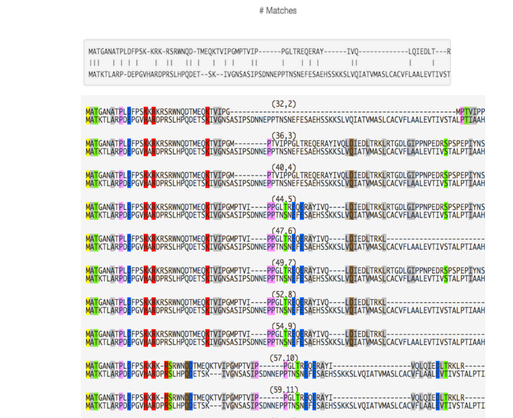
MOSAL is a software tool that provides the open-source implementation and on-line application for multiobjective pairwise sequence alignment. The web-server is available at http://mosal.dei.uc.pt. The implementation can be setup for several multiobjective score functions such as maximization of the number of matches or substitution score and minimization of gaps or indels. Speed-up techniques (pruning) are also implemented and the number of bounds are adjustable [2].

To produce the set of Pareto optimal alignments, these four steps are needed :

Step 1 : Insertion of each sequence in FASTA format.



Step 2: choose score function.
Step 3: alignment options.
Step 3: submission.



Results: An iterative plot will show all the score function values and the user can zoom and choose a given point to see the corresponding alignment. Furthermore, he can compare all the alignments and download all the solutions and values.

Approaches for the Problem VSD

Multiobjective Dynamic Programming

A two-dimensional matrix P is constructed where each entry stores the set of all states corresponding to PO partial alignments of strings A := (a_1,…, a_i) and B:= (b_1,… , b_j).

$$P[0,0] := \{(0,0)\}$$
$$P[i,0] := \{(0,-i)\}$$
$$P[0,j] := \{(0,-j)\}$$
$$\text{for } 1 \le i \le n$$
$$\text{for } 1 \le j \le m$$
$$P[i,j] := vmax \begin{cases} \{p + (s(a_i,b_j),0) \ : \ p \in P[i-1,j-1]\} \\ \{p + (0,-1) \ : \ p \in P[i-1,j]\} \\ \{p + (0,-1) \ : \ p \in P[i,j-1]\} \end{cases}$$

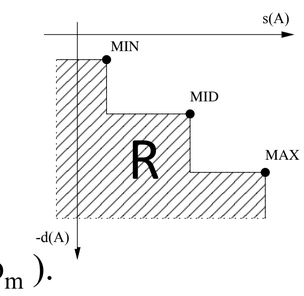
where $s(a_i, b_j)$ is the substitution score

The time and space-complexity of the algorithm is O(m·n·(n+m)).

Pruning Technique : Reduce the number of states by comparing an upper bound with a pre-computed lower bound set.

Lower Bound: Compute alignments with minimum number of indels (**MIN**), maximum score of substitution (**MAX**) and by using Needleman-Wunsch algorithm (**MID**)

Upper Bound: An upper bound of a state s in P[i, j] is given by the maximum score of substitution **u** and minimum number of indels **v** that can be achieved from entry P[i, j] to entry P[n, m].



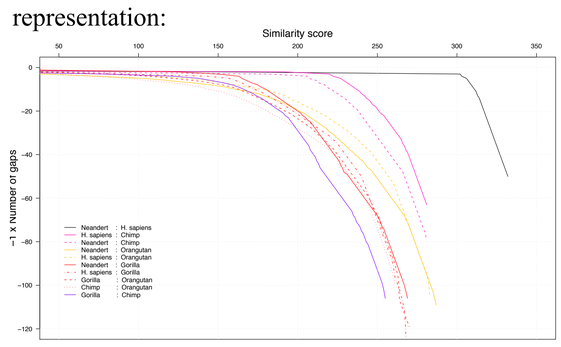
u : Size of LCS for subsequences (a_i ,…, a_n) and (b_j ,… , b_m).
v : Difference between the size of these subsequences.
A state (s,-d) is removed if the upper bound (s + **u**,-d- **v**) is inside of R.

These approaches can be easily extended for the VSG problem [1].

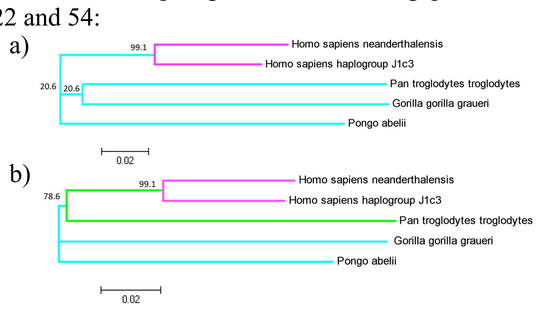
Phylogenetic Tree Construction

The non-dominated score set can provide further information for constructing phylogenetic trees [1].
Consider an example of comparison between primates: *Homo sapiens haplogroup J1c3, Homo sapiens neanderthalensis, Gorilla gorilla graueri, Pan troglodytes troglodytes* and *Pongo abelii species*. By using our biobjective method we found two trees topologies that are different slightly in relationship to the *Pan traglodytes traglodytes* with the remaining species. The relative branch frequencies suggest that the tree of plot (b) may be more reliable. Interestingly, this is also confirmed by the tree obtained with the ML method.

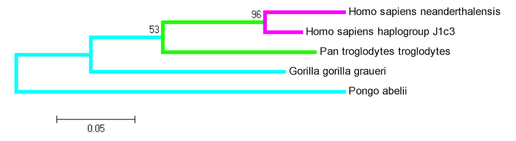
Non-dominated score set with staircase line representation:



The two trees topologies obtained for a gap value of 22 and 54:



Evolutionary tree using Maximum Likelihood (ML):



Future Work

These techniques can be extended for multiobjective multiple sequence alignment within some heuristic approach [3].

Acknowledgements:

Co-funded by Fundação para a Ciência e Tecnologia, project PTDC/EIA-CCO/098674/2008 and FEDER, by the "Programa Operacional Factores de Competitividade do QREN", COMPETE no. FCOMP-01-0124-FEDER-010024.

Governo da República Portuguesa
FCT Fundação para a Ciência e a Tecnologia — MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
UNIÃO EUROPEIA — Fundo Europeu de Desenvolvimento Regional
QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL PORTUGAL 2007.2013
COMPETE — PROGRAMA OPERACIONAL FACTORES DE COMPETITIVIDADE

References:

[1] M.Abbasi, L. Paquete, A.Liefooghe, M. Pinheiro and P.Matias. Improvements on bicriteria pairwise sequence alignment: algorithms and applications. Bioinformatics, 29(8):996-1003, 2013.
[2] L. Paquete, P. Matias, M. Abbasi and M. Pinheiro. MOSAL: Software tools for multiobjective sequence alignment. Source Code for Biology and Medicine, 9(2), 2014. (http://mosal.dei.uc.pt/)
[3] M.Abbasi, L.Paquete, F.Pereira and S.Schenker. Local search for bicriteria multiple sequence alignment. German Conference on Bioinformatics, 102, 2013.