UTILISATION DES ALGORITHMES GENETIQUES POUR L'ANALYSE DE SEQUENCES BIOLOGIQUES

Cédric Notredame

Doctorat en Bio-informatique

Février 1998

Université Paul Sabatier
France

Directeur De Thèse: Prof. François Amalric

Acknowledgement

I wish to thank the EMBL for their funding through an EMBL grant.

This work was carried out in the lab of Des Higgins, first at the European M
Biology Laboratory in Heidelberg,Germany and later at the EMBL outstati
European Bioinformatics Institute, in Hinxton, U.K.
Des has been a constant source of support. I wish to thank him and express
gratitude for teaching me most of what I know in bioinformatics, and so m
about being a scientist. I also wish to thank Ture Etzold who gave me a pla
team when Des had to leave for Ireland. Thanks to his expertise in the field,
had a lots of influence on my work.
A special thanks to the system managers, namely Roy Omond, Rodrigo Lo
Pettere Jokinen who have always been supportive, allowing me to overl
machines any time I needed it. Without their help, this work could not h
achieved. I also wish to thank Miguel Andrade, Inge Jonassen and Burkhard
stimulating discussions and frendship. Chris Sander and Liisa Holm have alv
available to share with me their extensive experience of the field. I wish to ex
my gratitude.

Michelle Magraine has proved an unvaluable allie for fighting against my
grammar weaknesses. I wish to thank her for that. I also wish to thank Rob I
joining me in my everyday fight against post script files. My stay at the EBI
extremly enjoyable.

This work is dedicated to my friends and family for their constant support ov
years.

RESUME DE THESE

Auteur: Cédric Notredame
Soutenance: Université Paul Sabatier, Salle des conérences du L.B.M.E. Le Lun Mars à 14 heures
Titre: Utilisation des Algorithmes Génétiques pour l'analyse de Séqu Biologiques
Directeur de Recherche: Prof. F. Amalric, L.B.M.E., Campus de l'université Pau Sabatier
Jury:

-Prof. F. Amalric
-Dr. Des Higgins
-Dr. William R. Taylor
-Dr. Manolo Gouy
-Prof. Jerzy Czaplicki

Mots Clés: Alignement de Séquences, Alignement Multiple, ARN, Protéines, SA RAGA, COFFEE.

Publications:

SAGA Sequence Alignmnet by Genetic Algorithm'
Notredame and Higgins
Nucleic Acids Research 1996, Vol 24, 8 1515-1524

RAGA: RNA Alignment by Genetic Algorithm
Notredame, O'Brien and Higgins
Nucleic Acides Research 1997 Vol 25, 22 4570-4580

COFFEE: an Objective Function for multiple sequence alignment evaluation
Notredame, Holm and Higgins
Bioinformatics, (in Press)

Optimisation of ribosomal profile alignments
O'Brien, Notredame and Higgins
Bioinformatics, (in Press)

Laboratoire d'accueil

European Institute of Bioinformatics
Genome Campus
Hinxton CB10 1RQ
United Kingdom

RESUMÉ
Une grande partie de la recherche fondamentale en biologie moléculaire repose sur l'étude et des acides nucléiques. Ces molécules extrêmement complexes résultent de la combinaiso plus simples: les acides aminés et les nucléotides. Vingt acides aminés constituent la gran des protéines, cinq nucléotides constituent la plupart des acides nucléiques. Le termes s utilisé pour désigner l'enchaînement de nucléotides constituant un acide nucléique ou l'e d'acides aminés constituant une protéine. La plupart des protéines sont codées par les gè dans l'ADN des chromosomes.

Au cours de ces dernières années, de nombreux progrès techniques ont rendu p séquençage à grande échelle du génome de plusieurs espèces bactériennes ou eucaryotes. C d'ADN sont entreposées dans des banques de données spécialisées (Swiss Prot, Gene Ba nucléotides database...) dont la croissance (en taille) est aujourd'hui exponentielle.

La bio-informatique est une sous-discipline de la biologie ayant pour objet l'analy
données par des moyens informatiques. Le principe de base d'une telle approche est l
relation entre fonction et séquence. Le but est d'extrapoler des données obtenues de façon e
sur certaines séquences à d'autres séquences pour lesquelles aucune donnée expérim
disponible.

Alors qu'il est clair que deux protéines (ou acide nucléiques) ayant la même séqu
probablement la même fonction, une corrélation devient plus difficile à établir lorsque les
présentent qu'une homologie partielle. La nécessitée d'utiliser ce type d'information est
motivation derrière le développement des méthodes de comparaison aujourd'hui utilisée
présenté dans cette thèse est essentiellement consacré à cet aspect de la bio-informatique.

L'un des moyens les plus utilisés pour la comparaison de séquences est l'aligner
alignement permet d'identifier les zones conservées entre deux séquences. Ces zone
correspondre à des motifs structuraux ou fonctionnels dont l'identification permet de faire
quant à la fonction putative des séquences analysées. De façon plus générale, un alignen
l'identification de régions sur lesquelles existent des contraintes diverses, imposant le
certaines propriétés. D'autre part, un alignement de qualité permet l'évaluation de la dista
séparant deux organismes, ou deux protéines.

Cependant, dans les cas complexes, la quantité d'information contenue dans deux
n'est pas suffisante, et il devient nécessaire d'étendre la comparaison à plusieurs séquen
l'objet des alignements de séquences multiples. Leur problématique est double.

Il s'agit tout d'abord d'un problème biologique. Etant donné un groupe de séqu
propriétés de l'alignement optimal doivent être définies. La règle la plus simple est de tent
autant d'identités que possible dans les colonnes tout en limitant le nombre d'insertio
(gaps). En pratique néanmoins, les règles utilisées sont plus complexes et peuvent prendre
nature des acides aminés alignés (protéines) ou la structure secondaire des séquences (ARN
de donner à cette liste de règles une forme mathématique associant un score à chaque ali
parle alors de fonction objective. Un nombre important de fonctions de ce types ont été
cours de ces dernières années. Globalement, elles peuvent être divisées en deux groupes: l
basées sur des matrices de substitutions et des penalitées d'insertion/délétion et les fonct
telles que les HMM (Hidden Markov Models). Une des propriétés les plus importantes d'ur
objective est sa signification biologique. De façon idéale, une fonction doit assigner à un
optimal un score traduisant l'intérêt biologique de l'information qu'il contient.

Le second aspect est purement informatique. Il ne suffit pas d'avoir une fonction ol
faut aussi être capable d'optimiser le score de cette fonction (i.e. produire l'alignement aya
score). Ce problème est loin d'être trivial. L'optimisation de la plupart des fonctions
appartient à la classe des problèmes dits NP complets. En conséquence, l'optimisation n
réalisée qu'en utilisant des méthodes dites heuristiques qui ne garantissent pas une solutio

Le travail présenté dans cette thèse englobe l'ensemble de ces problématiques. Dans l
partie, une méthode d'optimisation globale par algorithme génétique est proposée. Cette
intégrée dans un logiciel nommé SAGA (Sequence Alignment by Genetic Algorithm). Les alg
génétiques sont des stratégies d'optimisation basées sur une analogie avec le phénomène
naturelle. Cette méthode peut en théorie être appliquée à n'importe quel type de fonction o

Le second aspect du travail a consisté à définir une nouvelle fonction objective (C
Consistency based Objective Function for alignmEnt Evaluation) et à optimiser cette fo
utilisant SAGA de façon à prouver que COFFEE peut induire la création de meilleurs aligne
des méthodes alternatives.

La troisième application a été axée sur l'alignement d'ARNs ribosomiques avec de
d'une fonction objective adaptée à la prise en compte des interactions secondaires. Ce
adapté de SAGA a été nommé RAGA (RNA Alignment by Genetic Algorithm). L'une des pri
limitations de RAGA réside dans la simplicité de la fonction objective utilisée. Afin de rem
problème, un travail d'analyse a été réalisé sur des alignements de référence afin de dé
paramètres pouvant aider à la définition d'une fonction objective plus réaliste dans la pri
des contraintes à modéliser dans l'alignement. Ce travail constitue la quatrième applicati
dans cette thèse.

Dans l'ensemble, ce travail a permis d'établir l'utilité des algorithmes génétique contexte des problèmes d'alignement de séquences multiples. SAGA est à l'heure actuelle l le plus performant pour l'optimisation des fonctions objectives couramment utilisée alignements de séquences multiples. Dans le cas de séquences protéiques, SAGA est le seul capable de réaliser l'alignement global de plus de dix séquences. Pour ce qui est de l'ARN ri RAGA est le seul programme capable d'aligner des séquences ayant une longueur supérieure paires de bases, tout en prenant en compte les pseudo-noeux. D'autre part, la fonction C l'une des rares fonctions capable de permettre la génération d'alignements biologiquemer que ceux obtenus par ClustalW (ClustalW est une des méthodes d'alignement les plus popul

RESUME DES ANNEXES

Document Numéro 1
SAGA: Sequence Alignement by Genetic Algorithm

Dans cet article, une nouvelle approche est proposée pour la résolution du problème des al séquences multiples. Un algorithme génétique a été conçu et intégré dans un logiciel nom La méthode implique l'évolution d'une population d'alignements. Dans ce contexte, évolu que la qualité des alignements est graduellement améliorée au gré d'une succession (générations) contenant des étapes de modifications aléatoires (opérateurs) ainsi que d sélection basée sur le score. Le degré d'amélioration est jugé par l'évaluation du score alignement à l'aide de la fonction objective. SAGA utilise une technique de contrôle autom réguler l'utilisation simultanée de vingt opérateurs destinés à recombiner entre eux des (crossing overs), ou bien à les modifier individuellement (mutation). Afin de tester SAGA, utilisé comme référence le programme M.S.A. (Multiple Sequence Alignment) capable d'opt des fonction objectives le plus couramment utilisée (somme des paires avec pénalités délétion/insertions).

Utilisé dans ce contexte, SAGA fournit de meilleurs résultats que M.S.A. en t d'optimisation (score de l'alignement obtenu). De plus les alignements produits par S biologiquement plus exacts s'il on en juge par leur similarité avec l'alignement des même réalisés par comparaison de structures. Au total, SAGA a été teste sur treize groupes de séq lesquelles un alignement de référence basé sur les structures est disponible dans la banqu 3D_ali.

Document Numéro 2
COFFEE: A New Objective Function for Multiple Sequence Alignmnents

Dans ce travail, nous présentons un nouveau mode d'évaluation des alignements de multiples. Cette fonction est nommée COFFEE. COFFEE est une mesure du degré de con existant entre un alignement de séquences multiples et un bibliothèque de référence c mêmes séquences alignées deux par deux. Il est montré que le score COFFEE peut être eff optimisé par SAGA. La fonction a été utilisé sur onze groupes de séquences pour les alignement de référence est disponible dans la banque de données 3D_ali. Dans neuf ca SAGA, utilisé avec COFFEE, produit des alignements meilleurs que ceux obtenus avec Clust on en juge par leur similarité avec les alignements de références). Nous avons aussi montré assigné par COFFEE peut être utilisé pour évaluer la qualité d'un alignement multiple, de f ou globale. Finalement, la bibliothèque de référence peut être constituée d'alignement en par comparaison de structure (par exemple, des alignements extraits de FSSP). Dans ce cas COFFEE est capable de produire des alignements structuraux multiple de très haute qualité COFFEE devrait permettre d'appliquer aux alignements multiples n'importe quelle méthoc l'alignement de paires de séquences.

Document Numéro 3
RAGA: RNA Sequence Alignment by Genetic Algorithm.

Cet article décrit une nouvelle approche pour aligner deux séquences d'ARN homologues structure secondaire de l'une des deux molécules est connue. A cette fin, deux program développés, RAGA (RNA séquence alignment by Genetic algorithm) et PRAGA (Parallel R Ces deux programmes sont essentiellement basés sur le programme SAGA. La parallélis réalisée par la synchronisation d'un nombre défini de copies actives de RAGA. Celles-ci éch partie de leur population suivant une topologie définie comme étant un arbre à branche profondeur variable.

Cette méthode permet d'optimiser une fonction objective prenant en compte les in primaires et secondaires contenues dans les deux séquences. Une des propriétés les plus int RAGA réside dans le fait qu'il est possible de prendre en compte aussi bien les tiges boucle que les pseudo-noeux présents dans l'ARN ribosomique.

RAGA à été testé à l'aide de neuf alignements de référence extraits à partir d'alig d'experts. Ces alignements, constitués d'ARNs de la petite sous unité ribosomique, on référence. Dans chacun des cas, PRAGA est capable de surpasser en exactitude les méthodes basées sur la programmation dynamique. Ceci est vrai même lorsque la distance phyl séparant les deux séquences à aligner est très importante (comme entre l'humain et sac cerevisiae).

Document Numéro 4
Optimisation of RNA Profile Alignments

Ce projet fait pendant au projet numéro 3 et s'intègre dans un contexte plus large visant des outils nécessaires à la maintenance automatique des banques de données d'ARN ribos part le monde, plusieurs banques de données de ce type existent. Elles sont essentiellemen de façon manuelle. A long terme, la création de méthodes automatiques va devenir un absolue. Dans le document numéro 3, nous avons proposé une procédure destinée à l'al deux séquences. N'utiliser que deux séquences revient à ignorer la vaste quantité d'informa dans les alignements multiples établis par des groupes d'experts. Le but de ce projet a ét certaines des modalités d'utilisation de cette information.

Différentes méthodes de pondérations ont été testées et implémentée dans un c programmation dynamique. L'évaluation des méthodes a été réalisée en testant la qualité obtenue lorsqu'une séquence est extraite puis réintroduite dans un alignement multiple. C ne prend en compte que les contraintes primaires.

Les résultats montrent que par l'utilisation d'un mode de pondération adéquat et d'u pénalités d'insertion/délétion adapté, il est possible d'améliorer considérablement la l'alignement entre un profil et une séquence.

TABLE OF CONTENTS

1-INTRODUCTION 9

1.1 BIOINFORMATICS AND BIOLOGY 9
1.2. COMPARING SEQUENCES 11
1.3 OUR APPROACH 12

2-THE SCOPE OF SEQUENCE ALIGNMENTS 14

2.1 WHAT IS A SEQUENCE ALIGNMENT? 14
2.2 WHAT IS THE USE OF A SEQUENCE ALIGNMENT? 14
2.3 WHAT IS A 'GOOD' ALIGNMENT? 15
2.4 HOW TO BUILD A 'GOOD' SEQUENCE ALIGNMENT? 16

3-EVALUATING ALIGNMENTS 16

3.1 PROTEIN PAIRWISE ALIGNMENTS 16
3.1.1 Substitution Matrices 16
3.1.2 Gap penalties 19
3.1.3 Database Searches 20

3.2 EVALUATING PROTEIN MULTIPLE SEQUENCE ALIGNMENTS 21
3.2.1 Using Substitution Matrices and Gap Penalties: SP Alignments 21
3.2.2 Weights 23
3.2.3 Profiles 25
3.2.4 hidden Markov models 26

3.3 RNA ALIGNMENTS: TAKING INTO ACCOUNT NON LOCAL INTERACTIONS 28

4-MAKING MULTIPLE SEQUENCE ALIGNMENT 30

4.1 COMPLEXITY OF THE PROBLEM 30

4.2 DETERMINISTIC GREEDY APPROACHES 30
4.2.1 Aligning Two Sequences 31
4.2.2 Aligning Two Alignments : Progressive Alignment Methods 32

4.3 DETERMINISTIC APPROACHES FOR NON PROGRESSIVE MULTIPLE ALIGNMENTS 34
4.3.1 The Carrillo and Lipman Algorithm 34
4.3.2 Other Approximation Techniques 34

4.4 STOCHASTIC HEURISTICS 35
4.4.1 What is a stochastic Method? 35
4.4.2 Iterative alignments and Expectation-Maximization Strategies 35
4.4.3 Simulated Annealing 36

4.5 GENETIC ALGORITHMS 37
4.5.1 What is a Genetic Algorithm? 37
4.5.2 Applications of GAs in Sequence Analysis 39

5-COMPARISON OF THE METHODS 40

6-SUMMARY OF THE CONTRIBUTIONS 41
6.1 SAGA: MAKING MULTIPLE SEQUENCE ALIGNMENT BY GENETIC ALGORITHM 41
6.2 COFFEE: IMPROVING ON EXISTING OBJECTIVE FUNCTIONS 41
6.3 RAGA: THREADING RNA SECONDARY STRUCTURES 42
6.4 OPTIMIZING RIBOSOMAL RNA PROFILE ALIGNMENTS 43
CONCLUSION 43

REFERENCES 44

APPENDIX: RESEARCH PAPERS

Annex 1 "SAGA: Sequence Alignmnent by Genetic Algorithm" Notredame and Higgins, 1996

Annex 2 "COFFEE: A new Objective Function for Multiple Sequence Alignments" Notredame, Holm and Higgins, 1998

Annex 3 "RAGA: RNA Alignmnent by Genetic Algorithm" Notredame, O'Brien and Higgins, 1997

Annex 4 "Optimisation of ribosomal RNA profile Alignments" O'Brien, Notredame and Higgins, 1998

1-INTRODUCTION

1.1 BIOINFORMATICS AND BIOLOGY

Life as we know it is a complex arrangement of biological structures designe with each other. As complex as they may appear, even the most elaborated li can be described as arrangements of smaller less complex building blocks (s which are themselves the result of the combination of even smaller basic bl metabolites, proteins, nucleic acids).

Identifying these structures and characterizing their functions is a major a Questions can be addressed at any level of organization one may wish to populations to atoms in fact). In such a top to bottom approach, molecul almost at the bottom. It deals with the biological structures at the molecular understand how these are created and interact with one another to perf functions of life.

The search for ordered systems is an important part of the biological method systems usually make it possible to establish general rules allowing a global u of otherwise disparate collections of facts. In this respect, the discovery of D and the understanding of RNA and protein synthesis have been two of the n milestones of modern molecular biology. They have allowed a deep a understanding of some of the most central cellular mechanisms. We nov proteins and RNA molecules are involved at virtually all the steps of biologic We also know that these key components of cellular life are coded in DNA sec almost universal manner. These DNA sequences are contained in the genor organisms.

At the lowest level, a genome can be described as a long string of nucleotide compared to a very long text made of four letters. As in a text, the letters are at random but organized in words. In the context of a genome, a word will be having a function. One of the difficulties when trying to identify these 'word the fact that nature uses spaces and punctuation in a very personal way. Thi only do we ignore beforehand the function of the 'words', we also do not they start and finish. Add to this the fact that there are many different clas Some allow the binding of other molecules on the DNA, others are translate that can in turn be translated into proteins. A protein or an RNA molecule motifs that will have functions (binding, catalytic site....). The genome also c very specific combinations of words such as genes (enhancer, promot exons...). Bioinformatics and molecular biology are two complementary tec similar aims: identification of biological structures and sub-structures at a n and characterization of their function.

'Function' is a very general concept. If we look at it from an experimental such as genetics, a function can be defined with respect to a gene and will of from what happens when this gene is inactivated/modified by a mutation. O enough to gain a real deep understanding of the mechanisms involved. To c have to know whether this function is performed by a protein, an RNA, or sequence. If it is a protein then the next question is 'how does the protei function?'. If it is an enzyme we will want to know where is the catalytic site like any other known site, what are the residues involved in the site and hov their function? The protein (enzyme or not) may also interact with other pr acids or metabolites. Here again we will want to know what are the portions involved in these operations and what are the potential partners. In most of information will be much easier to understand/predict when a 3D model is a protein. In a broad sense, the function of a protein (or a nucleic acid or regulatory element) is defined by the sum of all these elements.

Until recently, the only way available for gathering together these pieces o
was to use wet lab techniques. These involve genetic analysis, cloning, s
interaction experiments..... Although the results obtained that way can usua
as strong biological evidence, they suffer from a major drawback. The cos
techniques is extremely high, in terms of time and money. This means that
on the number of functions that can be thoroughly investigated through su
The problem has become especially severe now that the improvement of
techniques gives us access to far more sequences than it will ever be possibl
the wet lab. It is this situation that has promoted the massive development of
techniques over the last few years.

Bioinformatics could be regarded as an approach diametrically opposed to
experimental ones. Instead of starting from a phenotype, one will start with a
try to gather as much information as possible by comparing this sequence
which experimental evidence is available. But the difference between the two
much less acute than it seems. Bioinformatics relies on the same basic ass
classic biology. It is a method of inquiry based on a series of comparisons
classifications/predictions. The main paradigm of bioinformatics is th
conservation is correlated to function conservation. Under this framework,
extrapolate, as much as possible the information acquired experimentally
follows a traditional feed back scheme where models are built and validated
by experiments made in wet lab or in silico.

Darwinian laws of evolution, and the notion of parsimony often underlie the
approach. The assumption is that biological systems have evolved from the
constantly reusing some basic building blocks (such as metabolic pathways)
them to respond to their environmental constraints. If each time a new const
a new biological system was created from scratch, the bioinformatics app
probably be bound to fail. Fortunately, in most of the cases, this is not w
Through the cycles of mutation/selection that constitute evolution, new func
created by reusing pieces of already existing machinery, and existing fu
evolved to become more adapted to the environment in which they are r
consider this problem in terms of sequences, this means that two sequences
similar functions may be different, depending on how long they have been
how long ago the original sequence was duplicated, or how long ago the tw
containing these sequences started diverging). Nevertheless, if the distance s
is small enough, an evolutionary scenario can be reconstructed that will sho
these sequences are. Depending on what is known for one of these sequence
sequences of the same category), it will then become possible to make assur
the function. On the other hand, if the sequences are evolutionarily too far a
analysing their relationship may prove difficult by simply comparing the s
signal they contain may have to be enhanced using other techniques suc
prediction.

Sequences are only conceptual objects. As such, they have no function in a
even the distinction between RNA, proteins and DNA is artificial. As far as
concerned, all these elements only exist as complex 3D arrangement of atom
of its precise 3D structure that a molecule has the mechanical and chemica
needs to perform its function (catalytic activity, interactions...). The rela
structure and function is probably one of the oldest paradigms of molecula
also accept that broadly speaking, structures are induced by sequence alth
for a fact that very different amino acid sequences can code for similar 3D fo

This last point helps understand why proteins with different sequences ca
function and structure, since natural selection is applied on the active 3D st
than on the sequences (i.e. evolution gives more freedom to the sequenc
structure). As a consequence, relationships between proteins (or RNAs) are
to analyse when the structures are known. Unfortunately, structures ar

determine experimentally and prediction from sequence alone (ab initio folc
is still one of the main challenges of computational biology. It is true howev
tools exist that can help supplement weak sequence identity.

Developing new techniques for automatically analysing sequences is one
purposes of research in bioinformatics. It is a point of crucial importance.
major databases of nucleotide or protein sequences are growing in size at a
rate (doubling every year or so). It means that the proportion of sequen
experimental data are available is decreasing. For this reason, targeting the p
experiments are needed has become more important than ever. Such a goa
achieved by gaining some more understanding on the ways in which infor
extrapolated from one sequence (or a set of sequences) to another. This is
available for making any use of the DNA sequencing results (at least in a
amount of time). It is for this reason that sequence comparison tools are at t
bioinformatics approach.

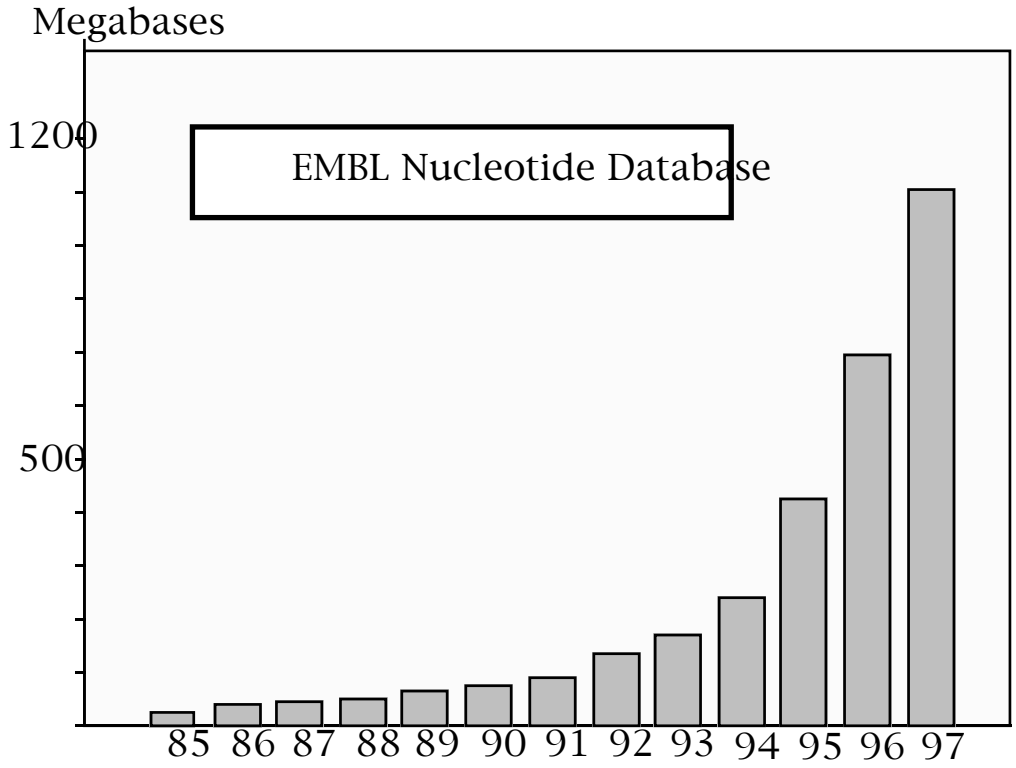


Figure 1. Growth of the EMBL Nucleotide database over the 1985-1997 period. The
release of the EMBL nucleotide sequence database (Rel. 52, October 1997) cor
1,181,167,498 nucleotides. The last release of the Swiss-Prot Database (Rel. 32 (
1996) contained 21 210 389 amino acids in 59021 entries. For comparison, the
release (proteins with known structures) of December 1997 contained a total of 673

1.2. COMPARING SEQUENCES

In most cases the problem facing the user takes the following form: a new
available and it is desirable to search the database and find out whether on
relatives of this sequence have already been reported. If so, one may wish
some of the experimental data gathered that way to the new sequence. In su
solution is to compare the sequences of interest to all the sequences con

database, keeping track of the most similar. Two very popular tools are us
basic database similarity searches: FASTA(1) and BLAST(2).

Sequence comparison can also be much more complex. For instance, by
experimental data contained in the databases and sequence analysis, one ma
if a specific motif is sufficient for a protein to bind zinc. If several types o
motifs emerge when doing such analysis, one may want to build a classifica
experimental data may also be available that allows the establishment o
differences between the zinc-binding motifs (some may be associated with
proteins and others with DNA-binding proteins). This for instance, is one
developed in the Prosite database(3). Once such results have been estal
sequences can be scanned for the known motifs they contain.

As simple and trivial as they may seem, such strategies present various v
difficulties that need to be overcome. First of all, one has to define what
means. This is typically a biological problem. The features one is interes
looking at two sequences or more will obviously depend on the aim and th
comparison. Do we want to compare two proteins for having the same seque
function, for having related functions, for being expressed in the same circu
having similar folds? Are these questions equivalent? When it comes to ma
analysis, sequence alignments appear as one of the most powerful solut
simplest case, they only involve two sequences (pairwise alignments) but
extended to a larger number (multiple sequence alignments).

Having decided that we want to use sequence alignments, the problem of def
'good' sequence alignment remains. It is a difficult question that requ
understanding of the biological information one wishes to extract from such
most of the cases, the criterion that will allow evaluation of the quality of an
take the form of a mathematical function(objective function) associating
alignment. But even so, the problem is not solved. Having a criterion for alig
is not enough. One also needs to be able to build the best scoring alignmen
this quality criterion. In most of the cases, this is far from easy. Many of th
bioinformatics and more specifically in sequence alignment are said to be
This means that the number of potential solutions rises exponentially with
sequences and their length (i.e. the solution cannot be found in polynomial t
The need to overcome such severe limitations requires the development
algorithms.

1.3 OUR APPROACH

In the work presented here, the problem of sequence alignment was appro
the two aspects mentioned above:

-defining new objective functions for sequence alignment.
-developing new ways to optimize these functions.

One of the main concerns of the approach was the fact that there is no use i
sequence comparison schemes if no tool is available to use them and allow ju
made on their potential relevance. To test an objective function, one mu
optimize it and compare the quality of the alignments it provides with other

The new optimization scheme proposed here is a genetic algorithm named S
1) for Sequence Alignment by Genetic Algorithm. This algorithm was used to
biological relevance of COFFEE (Consistency Based Objective Function for a
Evaluation) ,a new objective function designed for protein multiple
alignment(Annex 2). SAGA was also adapted to RNA alignments (RAGA
Alignment by Genetic Algorithm) using an objective function that takes i

secondary structure interactions in RNA (Annex 3). In order to improve RAG function, a new function was designed for aligning an RNA sequence to a la RNA sequence alignment (Annex 4). This function was only tested using a optimization method (Dynamic Programming).

The following sections will deal with the three main concepts associated w alignment: what it is useful for, how to define a sequence alignment and f build a sequence alignment. The last section will put the four contributions i context.

2-THE SCOPE OF SEQUENCE ALIGNMENTS

2.1 WHAT IS A SEQUENCE ALIGNMENT ?

A sequence alignment is the representation of two sequences in a way their relationship. If the alignment is correct, two residues aligned with or homologous. The definition of homology depends on the criterion used for For instance if the aim is to identify the relationship between two structures will be aligned because they are equivalent in the 3D structures. If the alignm to reflect phylogenetic relationships, two residues will be aligned when they the same residue in the common ancestor. The definition of a pairwise alig extended to multiple sequence alignments. In this case, several sequence together and each column contains homologous residues. However, homolo do not necessarily exist in each sequence for each position of the alignme sequence lacks one residue, a gap will be inserted in its place at the correspo Gaps usually take the form of strings of nulls. In an evolutionary context means that a residue was inserted in one of the sequences or deleted in the sequences were diverging from their common ancestor.

There are two types of alignments: global and local. In a local alignment, the that are aligned are those which are clearly homologous. The rest of the ignored. In a global alignment, whole sequences are aligned, regardless of the similarity. The scope of global and local alignment is usually different. Loca are more appropriate when the sequences analysed are remotely related and a few domains. Global alignments are mostly designed to analyse sequen known to be homologous to one another. In this thesis, I will mostly con sequence alignments.

It is also important to realize that, given a set of sequences, there are lots alignments. For instance, given two sequences of 1000 residues each. the 10^{764} different possible alignments. This rules out any naive enumeratio identifying the correct one! Instead, we will see that several strategies have b that allow more or less efficient computation in polynomial time.

2.2 WHAT IS THE USE OF A SEQUENCE ALIGNMENT ?

Quantitatively, the most widely used application of sequence align database searching. When doing so, the aim is to find for a given query seq related sequences contained in a database. The principle is very straightforw sequence is aligned in turn to all the members of the database and resul according to some similarity criterion. FASTA(1) and BLAST (2)are the most p of this type. They rely on local alignments rather than global.

However important the results obtained in this way, there is a clear limit on the alignments that can be deduced from these searches. Firstly, the seque algorithms implemented in these programs are only crude approximations sequence alignment algorithm. This is necessary in order to search very large reasonable amount of time. Secondly, in most of the cases, the searches a pairwise alignments. This means that they only contain a limited amount of i

Although database searching is at the heart of many approaches, pushing further may require important refinements. Such refinements can invo prediction, identification of new motifs or domains, generalization of the fam (i.e. combining the information contained in the known sequences in ord

distant members), phylogenetic analysis. For all these applications, pairwise a
of limited use. A way to simultaneously combine the information contain
sequences is needed. Such a need is the main motivation for building mult
alignments.

Multiple alignments are very important for phylogenetic analyses because t
way to compute evolutionary distances and phylogenetic trees. Trees are c
sets of pairwise distances using some clustering algorithms such as the nei
method(4). When computing a tree, it is very important to have accurate pai
hence the use of a multiple alignment in which the pairwise alignment of t
depends on the information contained in all the sequences of the set.

Another fundamental application of multiple sequence alignments is the id
motifs or domains. In a multiple alignment, these elements often appear a
which constraints exist that limit divergence. If some of the sequences are e
characterized, these motifs can be used for function prediction. This is, for in
the aims of the Prosite(3) or the ProDom databases(5). The information c
multiple sequence alignment can also be generalized in order to produce a
hidden Markov model (7)that can be used for identifying new family membe

The other important use of multiple alignments is structure prediction. In a
residues do not evolve in the same fashion, depending on their role in
(buried/exposed, helix/beta strand/loop...). It is very hard to extract such in
a sequence alone while it can be accessed through analysis of multiple al
looking at the distribution of the substitutions. Using multiple sequence alig
of sequences alone has had a dramatic effect on this area of sequence analys
accuracy of protein secondary structure predictions from 55%(8) (9)to 75%
One can also go further and try to identify correlated mutations in mult
alignments. This has been done on many occasions in proteins with limite
13). On the contrary, in RNA analysis, the identification of correlated mutat
of great help, allowing accurate prediction for secondary structure analysis a
structures(14-16).

Finally, a less challenging but very important application of multiple sequenc
the localization of highly conserved area for the design of efficient PCR prime
clone new members of a family. All these examples reflect the importance
sequence alignments in the domain of sequence analysis. We will show here
good multiple sequence alignments is a multi-step task.

2.3 WHAT IS A 'GOOD' ALIGNMENT ?

A scoring function associates a score to an alignment. Ideally, the bett
the more biologically accurate the alignment. An alignment with the best pc
said to be optimal, whether it is biologically relevant or not. An optimal alig
exists. Being able to distinguish between biologically relevant and non releva
is an important issue, especially when analyzing databases. Powerful statisti
been developed for this purpose, allowing the discrimination of hits for t
biological relevance(2). This problem remains when aligning sequences k
related. For instance, two homologous domains may surround a loop that is
two structures. In this case, any alignment of the residues contained by thes
meaningless even if a mathematical optimum exists.

Another important problem, common to many areas of computational biol
with the choice of parameters. Most of the objective functions come with co
parameters. In many cases, one has to rely on empirical values, known to la
(i.e. small changes of the parameter values may induce very different align
may lead to inaccurate alignments. This explains the fact that a large amou
dedicated to objective function definition focused on parameter elimination

much work has been done in this field that the choice of a scoring scheme regarded as one more parameter requiring optimization. Among the countl existing methods, we will only describe some of the most important scheme: those related to the work carried out for this thesis.

There are two types of alignments: sequence and structure alignment: alignments do not require any non local interactions to be taken into ac therefore less complex (algorithmically speaking) than structural alignm problem of sequence alignments we will mostly talk about protein sequen problem of structure alignment will be addressed through the example of R structures. The problem of protein structure alignment will not be analyse complexity and the amount of literature available on this subject put it beyo this dissertation which is mostly oriented toward sequence analysis rather th

2.4 HOW TO BUILD A 'GOOD' SEQUENCE ALIGNMENT

In many cases, building a sequence alignment takes the form of a c between biological relevance, mathematical optimality and efficiency. Given function, it may be very hard to produce the mathematically optimal alignm mentioned earlier that done in a naive way through enumeration, sequenc beyond the scope of any computer. For this reason, trade-offs need to be side. Objective functions need to be defined in such a way that they fit alr optimization techniques, and optimization techniques need to be improv accommodate the complexity of the problem.

For instance, in its more general form, the problem of aligning two sequ complete(17). However, if formulated under certain constraints, it can be technique known as dynamic programming(18) but becomes NP complete a number of sequences (i.e. when trying to align more than two s simultaneously)(19). Structure alignments (i.e. sequence alignments taking non local interactions) are also NP complete, even for two sequences, and addressed in a simplified form(i.e. using an objective function that does not known constraints)(20).

Because of this NP completeness most of the algorithms developed in the sequence alignments are heuristics(21-24). It means that they do not mathematically optimal solution, but rather a good approximation. In many off is reasonable and allows the computation of multiple sequence alignments manner. In this thesis, I will describe some of the optimization methods cu with a special emphasis on the genetic algorithms.

3 EVALUATING ALIGNMENTS

3.1 PROTEIN PAIRWISE ALIGNMENTS

3.1.1 Substitution Matrices

The twenty amino acids commonly found in proteins have very specific phy properties such as size, charge and hydrophobicity. The role of a residue mostly depends on these properties. For this reason, substitutions do not oc but in a way that reflects physico-chemical constraints in the 3D structure. I very intuitive idea to try to associate with each possible substitution a cost d probability. This information can be stored in what is known as a substitutio by 20 table giving the relative cost or the probability of each possible substitution. Although many types of matrices have been proposed, the more those derived empirically (25, 26). The principle often involves statistical

large set of alignments. Interestingly, these matrices tend to be in general a
what would be expected, knowing the physico-chemical properties of the r
substitutions conserving charge, size or hydrophobicity have lower costs). W
here the most popular of these matrices and their relative strengths/weaknes

The simplest possible substitution matrices are those only rewarding ide
alignment. Considering their simplicity, they do remarkably well in a vari
probably owing to the fact that they put a very drastic threshold on backgro
allowing the identification of very strong signals(27). On the other hand, t
disregard a large part of the information and this proves a big disadvanta
pairwise alignments between sequences with a low level of identity but a cle
The need to show that two sequences with a low level of identity can still be
similar has been one of the main motivations behind the development of r
substitution matrices that take into account similarity as well as identity.

The Dayhoff matrices(25), also known as PAM matrices, are among the most
The principle on which they are built is quite straightforward. Alignments of
proteins are made (more than 85% identity). When such a high level of ident
the sequences, alignments are usually straightforward and accurate. In the
the frequency of each possible substitution is measured. The table of freque
in this way is turned into a probability model (log-odds matrix). This model
define weight matrices, appropriate for comparing sequences of any degree
The distance is measured in PAM, Point Accepted Mutations per 100 residue
have matrices from 1 PAM up to 500 PAM.

```
C 11.5
S  0.1  2.2
T -0.5  1.5  2.5
P -3.1  0.4  0.1  7.6
A  0.5  1.1  0.6  0.3  2.4
G -2.0  0.4 -1.1 -1.6  0.5  6.6
N -1.8  0.9  0.5 -0.9 -0.3  0.4  3.8
D -3.2  0.5  0.0 -0.7 -0.3  0.1  2.2  4.7
E -3.0  0.2 -0.1 -0.5  0.0 -0.8  0.9  2.7  3.6
Q -2.4  0.2  0.0 -0.2 -0.2 -1.0  0.7  0.9  1.7  2.7
H -1.3 -0.2 -0.3 -1.1 -0.8 -1.4  1.2  0.4  0.4  1.2  6.0
R -2.2 -0.2 -0.2 -0.9 -0.6 -1.0  0.3 -0.3  0.4  1.5  0.6  4.7
K -2.8  0.1  0.1 -0.6 -0.4 -1.1  0.8  0.5  1.2  1.5  0.6  2.7  3.2
M -0.9 -1.4 -0.6 -2.4 -0.7 -3.5 -2.2 -3.0 -2.0 -1.0 -1.3 -1.7 -1.4  4.3
I -1.1 -1.8 -0.6 -2.6 -0.8 -4.5 -2.8 -3.8 -2.7 -1.9 -2.2 -2.4 -2.1  2.5  4.0
L -1.5 -2.1 -1.3 -2.3 -1.2 -4.4 -3.0 -4.0 -2.8 -1.6 -1.9 -2.2 -2.1  2.8  2.8  4.0
V -0.0 -1.0  0.0 -1.8  0.1 -3.3 -2.2 -2.9 -1.9 -1.5 -2.0 -2.0 -1.7  1.6  3.1  1.8  3.4
F -0.8 -2.8 -2.2 -3.8 -2.3 -5.2 -3.1 -4.5 -3.9 -2.6 -0.1 -3.2 -3.3  1.6  1.0  2.0  0.1  7.0
Y -0.5 -1.9 -1.9 -3.1 -2.2 -4.0 -1.4 -2.8 -2.7 -1.7  2.2 -1.8 -2.1 -0.2 -0.7  0.0 -1.1  5.1  7.8
W -1.0 -3.3 -3.5 -5.0 -3.6 -4.0 -3.6 -5.2 -4.3 -2.7 -0.8 -1.6 -3.5 -1.0 -1.8 -0.7 -2.6 -3.6  4.1 14.2
...C   S   T   P   A   G   N   D   E   Q   H   R   K   M   I   L   V   F   Y   W
```

Figure 2. A log odds matrix computed from mutation data. This is a PAM 250 m
extrapolated from the original PAM 15. Each entry indicates the cost for aligni
residues. The worst substitution costs are usually associated with the tryptophan (W

Originally, the Dayhoff matrices were established using 71 sets of aligned pr
pairs with 1572 point mutations (amino acid substitutions). The main limit c
is the fact that the content of information in alignments having 85% seque
low. Therefore, it appears risky to extrapolate such limited informati
evolutionary distances such as PAM 250. Furthermore, the extrapolation
model to any PAM distance is based on the assumption that mutations are
events. This hypothesis was challenged by several alternative methods.

The most popular alternative to PAM based scoring functions is the BLOSU
(26). It is based on a library of blocks, extracted from sequences of relate
block is a local multiple alignment that does not contain any gaps. About 20
used for establishing the matrices. Given a set of substitution frequencies, F

PAM matrices are computed in a similar fashion. The main difference is that the frequencies are measured in a way that takes into account sequence ider using the collection of blocks, several sets of matrices can be generated with do any extrapolation. They range from 80 to 45 % average identity. BLOSU have been shown on various occasions to outperform PAM or other matrices(

The main reproach that can be made to these two types of matrices, is that be general while only relying on alignments with a low information conter more than 85% identical), or domains that can be aligned without gaps (question of the existence of bias in these matrices has often been discussed let us consider alpha helices and beta strands. The type of substitutions obs two types of structural elements are known to be slightly different (this is th efficient secondary structure prediction algorithms(29)). As a consequence built using a data set that contains more helices than beta sheets, it will be l perform poorly when aligning portions of sequences coding for the beta-s other hand, if the data set is equilibrated for the two types of structures, t simply be an average, and will not be as good as it could be in either of the t attempt to compensate for this type of potential problem, several alternative built for helices(30), beta strands(30) or transmembrane domains(31). The the way in which such matrices should be used is far from being obvious. problem often amounts to solving structure prediction, unless a structure i some of the homologous sequences in which one is interested.

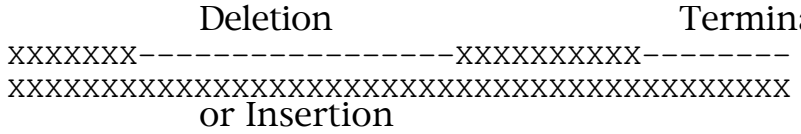
Overall, about 40 different substitution matrices have been proposed, using different methods with different training sets in most cases. Recently, two g have been made in an attempt to understand the fundamental differences schemes(27, 32). The main motivation in the work of Vogt et al. was to behavior of these matrices for the accuracy of the alignments they induce u programming (See Section 4.2.1). Correctness was judged using the structur contained in 3D_ali(33). Their work shows that there is very little differenc best matrices (Gonnet(34), BLOSUM(26) and Benner(35)). They also conc matrices which are able to identify remote homologues in database searches leading to the more correct alignments. Finally, from an algorithmic point 4.2.1) they concluded that these matrices are more suited for global alignm alignments. The second study(32) focused on understanding the way these m amino acids properties. The authors found that PAM units are significantly volume and hydrophobicity while other matrices are much more biased to Interestingly their results indicate that despite the different methods and di used for their construction, most of the substitution matrices based on seq are highly correlated with the Dayhoff's. When trying to group these matrice their level of correlation, the global matrices fall into the same cluster (i.e. ar while structure-specific matrices fall into separate clusters. This is furth supporting the idea that matrices should be applied in a way that takes structural information. We will see later that the Dirichlet mixtures(36) interesting alternative to this problem (see section 3.2.4).

The matrix comparison problem was also addressed in a different contex smaller set of matrices, by Henikoff et al. who compared matrices for the discriminate sequences in a database search using FASTA or BLAST (i alignments). The results obtained that way confirm that matrices based on tl distantly related sequences (such as BLOSUM) or structures (37) perform mu PAM matrices.

Finally, a point on which all these studies agree is the necessity of using ap penalties when scoring an alignment with a matrix. Most of the results ob substitution matrices can be dramatically affected, depending on the set of p score gaps. In the next section, we review some of the concepts underlying and the scoring of gaps when aligning two sequences.

3.1.2 Gap penalties

Substitutions are not the only events affecting sequences while they diverge. deletions also occur. This means that two sequences may contain unrelatec should not be aligned. Such an event is represented by a gap in the sequen receive the insertion, or where the deletion occurred.



It is obvious that a cost should be given to these events when scoring a Unfortunately the choice of a scoring scheme often implies some evolutiona we still lack a good understanding of the underlying biology of insertions/ instance, unless a very reliable phylogenetic scenario is available, it is ofter distinguish between insertion and deletion(38, 39), hence the word indel t position where an insertion OR a deletion occurred.

Some useful information can be gathered about indels by looking at alignments(39). As one would expect, indels do not happen at random, concentrate on the portions of the structure with less steric constraints such theory possible to extrapolate this information to sequences with unkn structure. For instance, Chou and Fassman(40) proposed using seconda propensity in order to derive some local gap penalties. Another method Pascarella and Argos (39) involves measuring, on reference structural ali probability of having a gap after a given residue. This scoring schem implemented in ClustalW(21) where a gap can be given a cost depending c after which it occurs. Similarly, some heuristics have been implemented in attempt to locate areas more prone to gap insertion such as stretches o residues, usually exposed in the loops. However, these methods are both e very general, they do not take into account the specificity of the sequences o in aligning (for instance due to the structural constraints, some positions conserved than others...). We will later discuss the way information der multiple sequence alignment can be used in order to established more reli propensity (cf. the Profile section, 3.2.3).

The question "where do gaps occur?" is only one part of the problem. When one also has to ask: "Is this gap long enough?". There is clear evidence that penalized according to their sizes. Analyses made on mammals(41) sug logarithmic scheme could be quite appropriate, with a gap opening penalty a penalty that is a function of the length of the gap (i.e. a penalty per residue the length of the gap). These results confirm those obtained by Benner e suggest that linear gap penalties are less than realistic (penalty cost increasi the gap length). Results also suggest that insertion and deletion should be from one another(38, 41). However, even if there is a wide agreement on the linear schemes would probably be biologically more relevant(42), alterna suffer from a major drawback: their implementation poses significant algorit when it comes to optimizing alignments(43, 44). For this reason, in practice are usually optimized under a simplified form known as "affine gap penal formalised as follow:

$$\text{cost} = \text{Gap Opening Penalty} + \text{Gap Extension Penalty} * \text{Length} \quad \text{(eq. 1)}$$

This gives penalties as a linear function of gap length and an efficient a optimizing this was introduced by Gotoh(45). There is no real justification type of model apart from the fact that it performs reasonably well, especially are small (less than 20 residues). Since the size of indels is known to be a fu evolutionary distance(38), this means that linear gaps will be acceptable w closely related sequences. However, since affine gap penalties are not empi they raise the problem of defining the values of the two parameters: the gap o (GOP) and the gap extension penalty (GEP). There is no guaranteed way to values so that they fit the sequences one is interested in. A popular practice opening penalty a value equal to the average of the values contained in th matrix used for comparison (excluding the main diagonal). The extension p set to a tenth of the opening value. It is also common practice not to penalize at least for opening.

When making an alignment there is a competition between gap insertion an To some extent, gap penalties can be regarded as thresholds used to deci stretch of residues has a homologue or not in the other sequences. In this co sense to modulate the penalties with some local information (seconda propensity, profile information....). But even so, a major problem remai sequences aligned are only remotely related, the gap penalties lack robust made by Vingron and Waterman showed that slight variations of values for the GEP can induce very different optimal alignments(46). Under such conc be hard to decide which alignment is biologically the most relevant. An atter the robustness of the penalty parameters has been proposed by Taylor: "Sco Enhancement"(47). It originates from the observation that in biologica alignments, gaps are usually clustered in a few parts of the sequences and long uninterrupted blocks. The technique proposed by Taylor involves enha of long ungapped portions in order to avoid them being interrupted by ga shows that under this scoring scheme, a correct guess for the penalty values less critical than previously reported.

There is little doubt that the correct treatment of gaps and a deep underst biology are critical parameters when making accurate sequence alignments shown here, the problem is mostly algorithmic. It is possible to define gap describe sequence relations in a realistic way, it is simply not practical to o The problem of practicality becomes even more acute when it comes to scan containing hundreds of thousands of sequences.

3.1.3 Database Searches

An exhaustive treatment of database search methods is beyond the scope of most commonly used methods will be briefly described here as respect to th involve specific scoring schemes designed for evaluating the statistical signi alignment. The principle of a database search is quite straightforward. A que aligned in turn with all the other sequences of the database. Depending o alignments are kept as relevant or discarded.

BLAST(2) is probably one of the most popular methods for database searche proteins, the method involves finding the High Scoring Pairs of residue stretches of aligned residues, with no gap, which have high scores). The evaluated using a substitution matrix (PAM120 for instance). These segments looking for words of a specified size(48), and by extending these words. Sinc does not allow gaps, it will in many cases be restricted to more or less small such a context, the score alone will not be informative enough to decide wh high scoring pair is significant. These scores will need to be normalized in or comparable from a statistical point of view. This normalization takes into ac of the database and the size of the sequences and gives the probability that by chance. This score is called the E value and is used to rank the hits.

The other popular tool for database searches is FASTA(49). As in BLAST, FA
by looking for high scoring segments using the Wilbur and Lipman meth
segments the method is interested in are those having a high proportion of
are scored by using the main diagonal of a substitution matrix such as PAM
considering identities, but giving them a score that depends on the amino
best diagonals found that way are then re-scored using a full substitutio
second step, non collinear segments are joined by dynamic programming, u
joining penalty (analogous to a gap penalty). The resulting scores are then
used to rank the matches. In order to prevent this chaining step from decrea
it is only applied when the best scoring segment is above some empirical t
mean and the standard deviation of this distribution are then evaluated an
on a final threshold that is used to separate spurious hits from real ones.

Of course, both these methods lead to false negatives and false positives, p
they do not use the most accurate method for local alignments (50) that hav
significantly outperform FASTA(51), and also because some background noi
to avoid especially when dealing with very large databases. The reason why
behind FASTA are less evolved than for BLAST has to do with the fact that a
statistical significance of a gapped alignment is much harder than for an ung
as in BLAST. This may change soon. Vingron and Waterman have recently
scheme that allows the estimation of probabilities for gapped alignments(52
very recently, a new version of BLAST, gBLAST(53), has been publish
incorporates some of these results and allows statistical ranking of gapped al
scanning a database. Other statistical scoring systems include the one
Bucher(54).

3.2 EVALUATING PROTEIN MULTIPLE SEQUENCE ALIGNMENT

3.2.1 Using Substitution Matrices and Gap Penalties: SP Alignments

Extending the definition of pairwise costs to multiple sequence alignments
complicated. To keep within the framework used for pairwise alignments
multiple alignment cost which is the sum of the substitution costs (co
substitution matrices and gap penalties). Nonetheless, based on different
scenario.

Because genetic mutations are binary events which change one protein or
sequence into another, substitution costs for multiple alignments are gener
terms of those for pairwise alignments(42). Two different approaches have b
The first is to define the substitution cost for a set of elements as the sum of
cost for all pairs of elements chosen from the set(55-57). Thus for three seq
k the cost of the alignment i, j, k will be equal to the sum of the cost of e
alignment induced by the multiple alignment (cost (i,j)+cost(i,k)+cost (j,k)).
pairwise alignments are also called pairwise projections. An alignment defin
called SP for "sum-of-pairs" alignment. In such a context the evaluation
alignment amounts to measuring the dissimilarities within a set of letters.
approach has the advantage of being straightforward and intuitive, its main
that it has no clear foundation in the theory of molecular evolution. Sankof
an approach in closer agreement with biological intuition. In his model, an ev
is assumed where each sequence is a leaf. The nodes are occupied by re
sequences. If the tree has k nodes, then substitution costs are defined on k-tu
the sum of the pairwise substitution costs associated with each edge of
alignment defined that way is a tree alignment. It must not be confused wi
alignments (see Section 4.2.2) which often rely on estimated phylogenetic
computation of an approximated SP alignment. In the cases where the tree
only one central node, it is named star alignment, being based on star phylo

Despite the fact that tree alignments are biologically more realistic than SP al
have not become very popular, mostly because of the fact that their constru
serious algorithmic difficulties. A majority of the multiple alignment meth
pairwise substitutions attempt to produce optimal SP alignments. In an
defining gap costs is not necessarily straightforward, as can be gathered from
alternative schemes (55, 56, 59, 60).

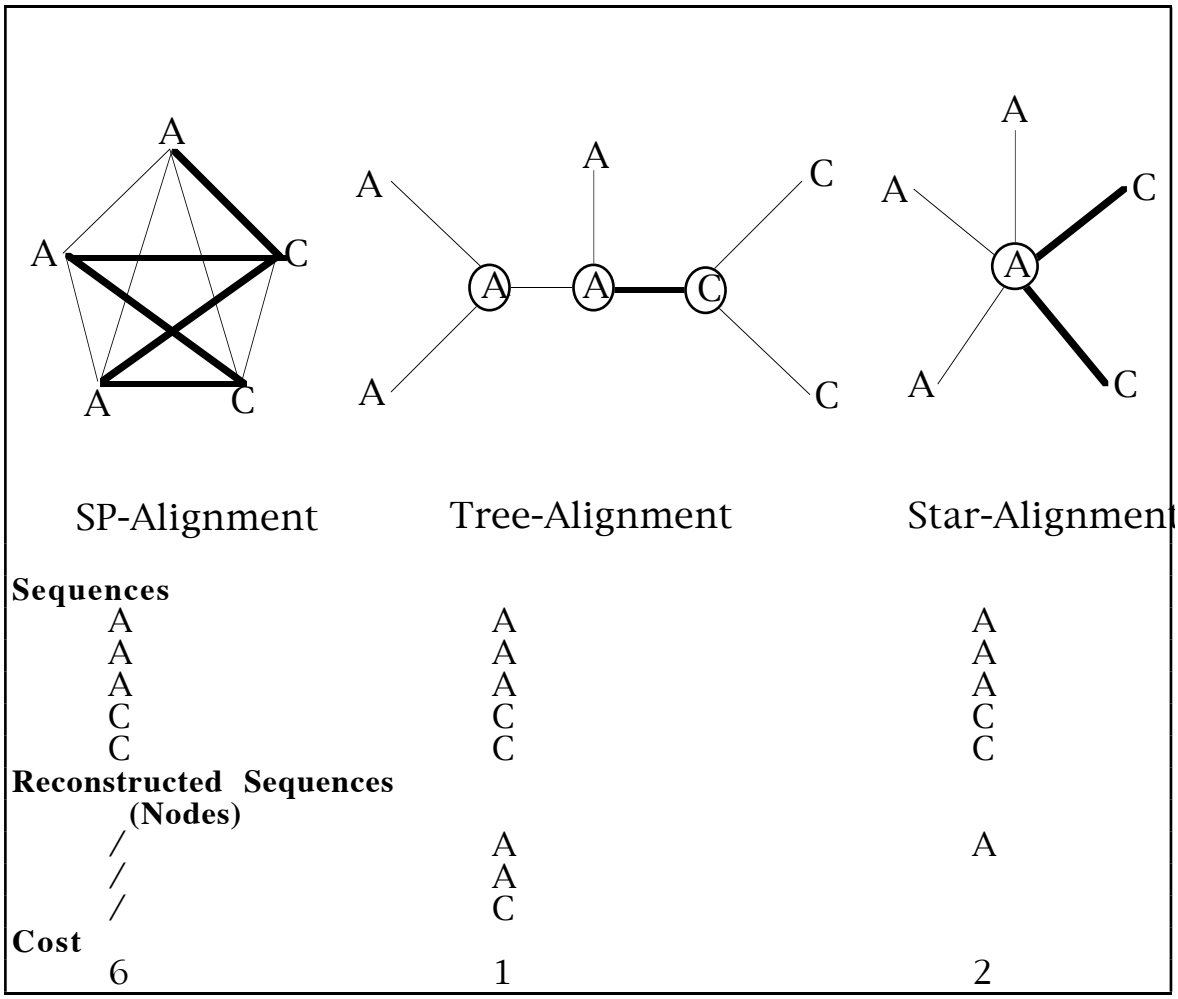


Figure 3. SP, Tree and Star alignment substitution costs for five 1 letter sequences
Altschul, (42)). The reconstructed sequences are indicated by circles at the tree nod
lines indicate substitution cost of 0 while plain lines indicate a cost of 1.

The simple implementation of pairwise costs in SP evaluation is known as
costs'. The idea is to consider the costs of the gaps in each pairwise projectio
alignment (any column of nulls is removed from the pairwise projection). A
gap costs seem to be the obvious companions of an SP alignment, Altschul
one to formally propose them(42). Most of the alternative schemes have be
for aligning simultaneously three sequences. For instance, Gotoh(56) propos
gap as a set of columns having a null in identical positions. For Murata(55),
of adjacent columns, each containing at least a null. The main motivation
schemes was algorithmic. These approximations were made in order to m
evaluation easier when computing an SP alignment. It is for this same reason
proposed a simplified version of the natural gap cost named 'quasi natural g

Quasi natural gap costs are very similar to natural ones. The main difference
pairwise projection is considered, columns of nulls are not removed, and a

counted as opening when a null run in one sequence starts after and finish
run in the other sequence, such as:

```
Sequence 1 XXXXXX----------XXXXXXX
Sequence 2 XXXXXXXX-----XXXXXXXXXX
```

that will lead to considering two gaps when in practice there is only one be
sequences. The motivations behind this approximation are purely algorithm
do with efficiency requirements when implementing the Carillo and Lipman
multiple sequence alignment(57) (see Section 4.3.1). This scheme induce
favors similar gaps in aligned sequences, Nonetheless, this approximation
because indels seem to be rare events that tend to be kept unchanged throug
As a consequence, in most multiple alignments, the number of times whe
natural scheme induces alignments different from the natural one should b
Natural and semi-natural gap costs can also be applied to tree or star alignme

This type of gap penalties in the context of an SP alignment constitutes on
widely used objective function for multiple sequence alignments. It is often
"sums of pairs with affine gap penalties". Some of the drawbacks of this r
already been discussed in the previous section. The main ones stem from
substitution matrices are general descriptions that do not take into account l
this is true as well of gap penalties that should incorporate local information
do so has been made in ClustalW where penalties are locally reassessed, tryir
information from the other sequences being aligned(21). Another weakn
function has to do with the fact that sequences are considered in pairs on
makes more sense to consider a column in a multiple alignment as a distrib
acids generated by evolution. In the next sections about hidden Markov mod
(Sections 3.2.3 and 3.2.4), we will present some methods that attempt to tak
account when scoring multiple sequence alignments.

Finally, a potential weakness inherent in any scoring scheme is the prob
representative information. The sequences used for building an alignment ra
representative set. They are often biased by the composition of the data
collecting these sequences. In such a case, the alignment of the sequences
isolated minority will suffer from the fact that the information they contain
among the rest. We will see that several weighting schemes have been design
overcome this problem.

3.2.2 Weights

With most of the multiple sequence scoring systems, weighting of the s
necessary. Weights are designed in order to correct for unequal representati
of sequences. For instance, when aligning together globin sequences found
database like Swiss-Prot(61), large numbers of sequences will be identical,
will be quite different from the rest of the set and will have no close relativ
we want our alignment to be representative of the globin family in general, it
avoid a complete domination by the vertebrate myoglobin and hemoglob
simply because the database contains far more of them. Such an alignment, r
each sequence the same weight would be biased. Weights are used to avoi
bias. Several methods have been proposed that can be separated into two g
they depend on an alignment or an estimated phylogenetic tree.

Alignment based weighting methods do not require the sequences to be
Therefore, complex issues of tree topology and root placement are avoided.
can be based on pairwise distances (62) or on the distances from some avera
sequence (63). Whatever the method, the general trend is similar and res
weighting of the sequences which are poorly related to the rest of the set wh

which, on average, are more similar to the other sequences have their weigh
lowered.

The tree-based weights assume that sequences are related through evoluti
reasonably correct tree can be deduced from pairwise distances (64). Two s
type have been proposed: branch-length proportional weights (65).and the
Lipman (ACL) weights that are based on a statistical analysis of the tree topol

In the ACL scheme, a sequence receives a low weight if it is far from the tre
has close neighbors in the tree. ACL weights have the advantage of correcti
information without biasing the alignment toward very divergent seq
underlying assumption is that although a very divergent sequence cont
information, it is hard to exploit such information without bringing in too m
The main weakness of the ACL method is that when the topology of the tr
establish, mistakes can be made regarding the position of the root. The wei
by Thompson et al. (65) provide an alternative solution. They also rely on
tree, but under the scheme, sequences only get down weighted for having
neighbors.

In an SP context, applying pairwise or sequence weights to score a multiple
straightforward. The weighted sums of pairs alignment score can be formula
Given N sequences $S_1..S_n$ a weight can be estimated for each possible pai
sequence S_i, S_j. This pairwise weight will be obtained directly through com
will result from the combination of individual sequence weights. Each pairw
$A_{i,j}$ of the sequences Si and Sj in the alignment has a cost value $COST(A_{i,j})$ evaluated using a
substitution matrix and a set of gap penalties. Given these definitions, the g
SP score is equal to:

$$SCORE = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} W_{i,j} * COST(A_{i,j}) \qquad \text{(eq. 2)}$$

As with matrices, an important issue for weights is to decide which schem
applied. Each of these weights have desirable properties and unwanted
Vingron and Sibbald proposed a systematic way of comparing five different
Their conclusion was that when sequences are related through a robust phy
the ACL do better than alignment based methods. On the other hand, when t
between the sequences is harder to estimate, leading to an inaccurate phylog
Sibbald and Argos(63) or related methods(67) are preferable. Similar resu
recently established by Henikoff and Henikoff(68) using an empirical evalu
These authors found that for phylogenetically related sequences, tree base
preferable and that the Thompson's scheme slightly outperforms ACL. It
however, to what extent these findings are method dependent, especially if
that most of these weights are used in different heuristic alignment making s

Gotoh pointed out the fact that weighting schemes are very likely to
dependent(69). For instance, an important difference between the ACL we
Thompson's ones is that the ACL method produces pairwise weights while th
individual sequence weights. As a consequence, the ACL weights cont
information since the pairwise weights they depend on are not necessarily
there is not always a set of sequence weights corresponding to a set of pairw
On the other hand, as we will see later, these two types of weights are u
different contexts. Thompson's weights are mostly used for progressive alig
they are probably very appropriate since remotely related sequences usually
on the overall multiple alignment(21). ACL weights are used in the program M

Sequence Alignment(22)) which does global simultaneous alignments where
is given a chance to affect the overall alignment.

We will now see that weights can also be useful for the construction and
sequence profiles (i.e. generalized alignments used to describe a protein
domain).

3.2.3 Profiles

Multiple sequence alignments can be used to provide position specific sco
known as profiles(6). The procedure of turning an alignment into a pro
straightforward. It involves counting the residue frequencies in each column
alignment, and deducing from these measures, a table of substitution c
position of the profile. A local cost is also evaluated for gap insertion and e
term profile refers to the collection of costs associated with each position of
A profile can be treated as a single sequence and aligned to any other sequer
using the profile substitution costs and penalties instead of a single matrix.

		A	C	D	E	F	G	H	I	K	L	M	P	Q	...	Gap Penalty
POS	ALN															
1	EGVL	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	...	9
2	LLSP	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	...	9
3	VVVV	2	2	-2	-2	2	2	-3	11	-2	1	-2	-2	0	...	9
.		.	.	.	.	.	.	.	.	.	.	.	.	.		.
.		.	.	.	.	.	.	.	.	.	.	.	.	.		.
.		.	.	.	.	.	.	.	.	.	.	.	.	.		.
21	SS-D	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	...	4
22	S--S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	...	4
.		.	.	.	.	.	.	.	.	.	.	.	.	.		.
.		.	.	.	.	.	.	.	.	.	.	.	.	.		.
.		.	.	.	.	.	.	.	.	.	.	.	.	.		.
49	SSNY	2	5	2	1	1	2	1	0	1	-2	-2	5	1	...	9

Figure 4. Example of profile (adapted from Gribskov et al.(6)). For each position (P
a multiple alignment (ALN, presented in a vertical format with each line correspon
column) a substitution cost is calculated for any amino acid that would be align
position. A gap penalty is also evaluated. Note that at positions 21 and 22 of the p
gap penalty is lowered because the alignment used for the profile contains gap
position.

A profile is specific for a family (or a domain). One of the main usage of p
search databases for new members of a family. In such a context the desira
are sensitivity and selectivity (i.e. recognize very remote homologues and dis
positives). Such a result can only be achieved if the profile induces very goo
This in turns depends on the quality of the profile itself and can be affec
factors: (i) choice of the sequences, (ii) method used for building the multi
(iii) method used to turn the multiple alignment into a profile (especially the
gaps and the method used to describe background frequencies).

In many cases, the choice of the sequences is directly imposed by the datab
the best way to remove this type of bias is to use a weighting scheme wher
sequences (see previous section). However if one wishes to build some v
profile, it is also possible to select the appropriate sequences, using techniq
one described by Neuwald et al. (70). Weights also need to be applied when
is turned into a profile. On various occasions, it has been shown that many
used for sequence alignments do as well when used to build profiles, and h
the level of generalization(65, 71). Accumulation of gaps is another side effe

when many sequences are used to build a profile. Because the number of increase with the number of sequences, schemes have to be used for down effect of their occurrence(65, 72, 73).

Profiles are not only important for database searches, their computation is step for some multiple sequence alignment strategies based on a progressive 74, 75). In this context, to build the full multiple sequence alignment, pa alignments need to be aligned in intermediate steps. The best way to do so i these alignments into profiles, and to align these profiles with one another doing so have been extensively described by Higgins(76, 77) and Gotoh(69,

Finally, since profiles are involved in database searches, some significant w done on establishing the statistical meaning of alignment scores(78, 79). T aspect of the problem has received much more attention in the context of the models based approaches that will now briefly be reviewed.

3.2.4 hidden Markov models

A hidden Markov model (HMM) describes a series of observations generated b stochastic process (a Markov process)(80). They have been used extensive recognition. HMMs designed for sequence alignments are related to profiles to provide a statistical model representative of a given family of proteins(7). of the main advantages of HMMs (as opposed to profiles) is that they prov estimate a model directly from unaligned sequences. However, in practice methods available for HMM optimization require the computation of multip Nevertheless, HMMs have some interesting features that distinguish them fro

A key concept in HMMs is the notion of states. A HMM is a chain of elem different possible states. The number of possible states is arbitrarily define for instance, there are three states: align, insert and delete. When going thr probabilities are given to each possible transition. The values of thes probabilities are evaluated by training the model using known members of a Sequences can be aligned to a trained model using a variant of dynamic pro known as the Viterby algorithm(80). An alignment between a sequence an called a path, in the sense that it joins different states in order to produce th highest probability. To a large extent, aligning a sequence to a model can b equivalent to aligning a sequence to a profile. There is however a fundamen difference. A new sequence is not 'aligned' to a HMM. What is measured is th for a given HMM to generate the optimally aligned sequence (i.e. the seque right pattern of gaps/unaligned residues/aligned residues).

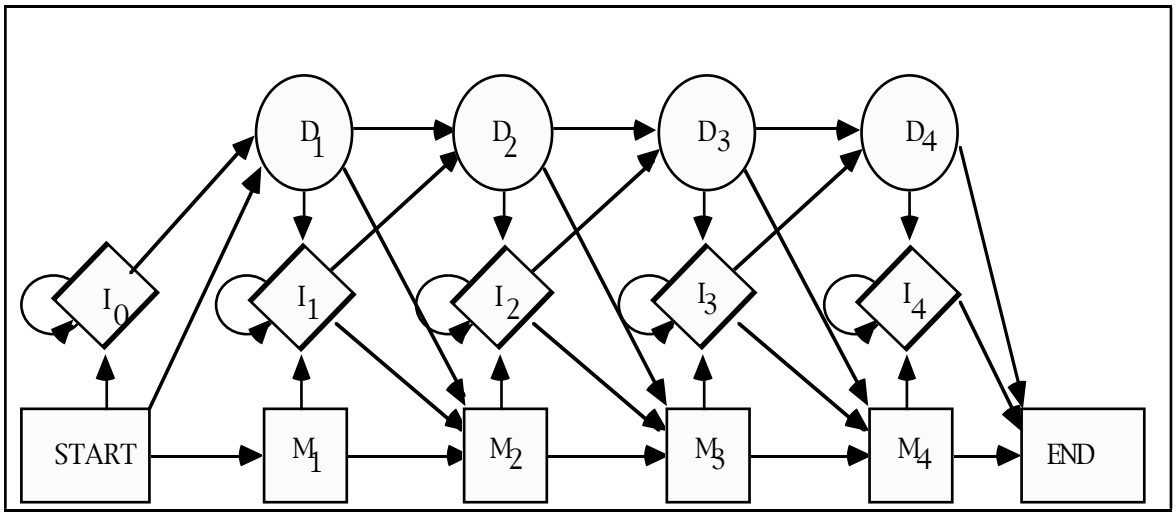


Figure 5. A linear hidden Markov model (from Hughey and Krogh, (82). This mod three different states (M, I, D). Each state is connected to the others by a tr probability (arrows). Assigning the weights to each transition is the purpose of the

The number of sequences, and their range of identities are critical factors tha the model. In their simplest expression, HMMs do not require any prior in opposed to profiles that required a multiple alignment made using a substitu HMMs, residues are described as 'letters' and the training only relies on establish the parameters. If there are enough sequences in the training set, t a sensible model since the substitution constraints will be 'discovered' by t position per position basis. The accuracy and the sensitivity of a model tra will be highly dependent on the number of sequences and their informati overcome possible lack of information (missing data), pseudo counts are in order to simulate background frequencies (82, 83). Their influence on the m with the amount of information present in the sequences.

The actual values of the pseudocounts can be estimated using various met measure the probability of each amino acid in the training set, or othe probabilities from a standard substitution matrix. Generally speaking, th number of sequences used for the training, the more critical the values of the In this context, Dirichlet mixtures(36) proved extremely useful. A Dirichlet mathematical tool that, given an observed amino acid distribution and a s distributions allows the computation of a probability for the observed dis hidden Markov model context, these mixtures can be regarded as the eq substitution matrix. They have been shown to be more sensitive to sequenc or variation than traditional substitution matrices(36).

As with profiles, HMMs can be used to generate multiple sequence alignn scanning databases. Scoring can be made by combining the probabilities of a alignments of a sequence to a model, which is equivalent to calculating the t of a sequence given the model. This can be done efficiently using the Viter (80). Such a score is called the NLL score for Negative Log Likelihood score scores measures how far a sequence is from its model (in other words, the st forcing a given model to produce an aligned sequence). The problem with that they depend on the size of both the sequence and the model. One way t is to measure the Z score, the number of standard deviations a NLL score is a average NLL score of unrelated sequences of the same length. A complete alg computation of Z scores is described in (82).

This study of multiple sequence alignment scoring systems is far from exhau variety of alternative methods have been described that roughly fall int

categories: those relying on SP evaluation and those (like HMMs) tha distributions of amino acids rather than pairs. Although it seems that distributions is a more realistic approach, the main reason why SP scheme been more popular has mostly to do with algorithmic problems associated w based methods.

Both methods only deal with one aspect of the problem: the use of local inf know that since proteins fold into active 3D structures, there must be more the sequences (i.e. tertiary structure interactions...) than what we discussed in an appropriate manner there is no doubt it could help to improve the alignments. Few methods have been proposed to deal with these non-local There are two good reasons for that: this type of signal is usually very weak i the algorithmic problem is even more complex than when taking into acc sequences. We will now see that the problem is different with RNA hence the complex types of objective functions.

3.3 RNA ALIGNMENTS: TAKING INTO ACCOUNT NON LOCAL INTERACTIONS

When reliable non local interaction information is available, it makes sense t into the scoring scheme. This is rarely the case with proteins, hence the encountered when doing structure threading (threading a protein sequenc structure)(84). Fortunately, in the case of RNA the problem is different and govern the formation of non local interactions are better understood(85, 86) the secondary structures encountered in RNAs are due to Watson and Crick The existence of mathematical models that describe these interactions on a t basis make RNAs good candidates for 'ab-initio' folding predictions. Howev parameters influence the structure of an RNA molecule (local conditions proteins, unknown tertiary interactions...) that accurate prediction thermodynamic models are still very imperfect despite some recent improve This does not mean that the thermodynamic approach is wrong, but it may information it uses is not sufficiently detailed considering our knowledge folding process. However when combined with phylogenetic information (s taken from multiple sequence alignments) predictions can be realized to a accuracy(16).

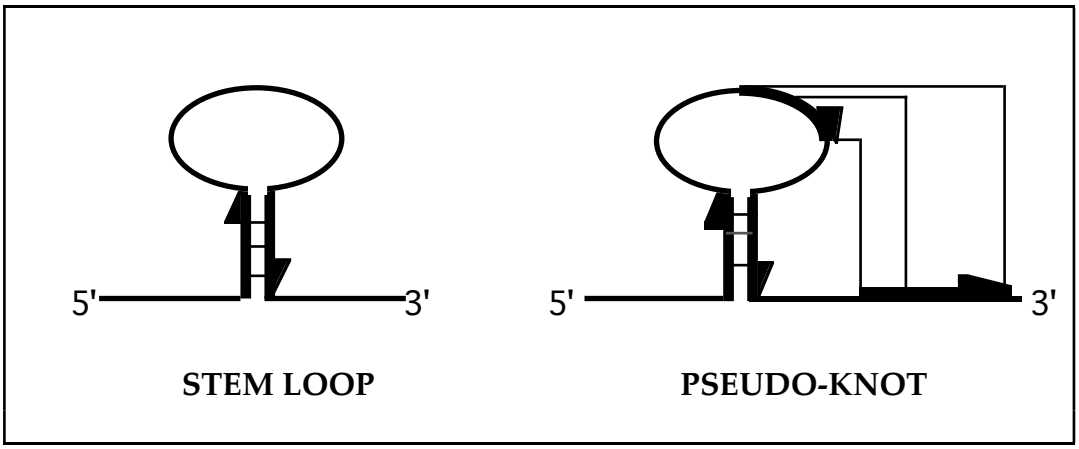


Figure 6. Some of the motifs commonly found in RNA secondary structure. Base pa are usually made through Watson and Crick interaction although non canonical p often been reported.

Multiple sequence alignments have the advantage of revealing constraints wi of hypotheses on their origin. Such analysis, performed on ribosomal RNA se

made it possible to predict the secondary structures of these molecules wit
outperforming traditional energy minimization schemes like Zuker(89). U
these types of RNA alignments are difficult to build automatically due
algorithmic problems.

Several functions have been proposed that incorporate RNA secondar
information into their evaluation of multiple sequence alignment quality(9
approach, Kim et al. (90) proposed a function that computes the probabil
potential secondary structures contained in a multiple sequence alignment t
account. It is a scheme that has the advantage of being very flexible (for inst
pseudo-knots) and not requiring the actual computation of a secondary stru
are not assessed for compatibility). Its main drawback is that evaluation tim
with the length of the alignment. This can prove a severe limit when dealing
sequences (like ribosomal RNAs) while using an iterative or stochastic optimiz
such as simulated annealing. Eddy and Durbin(91) proposed a different typ
In their method, the secondary structure is expressed as a binary tree in wh
stands for a column in a multiple sequence alignment. This tree can be s
through a generalized HMM (HMM with bifurcation) named Covariance Mo
CV can be trained like an HMM. Once this has been done, the sequences are t
the model in order to produce the multiple sequence alignment. This app
similar to the stochastic context free grammar (SCFG) methods(93, 94), whe
to express the structure using a special type of regular expression. The align
by parsing the sequences through the proper expression that has been obtai
on the sequences. CV and SCFG methods suffer from the same drawback. Th
allow nested structures and cannot take into account pseudoknots(95, 96) as
method proposed by Kim. Furthermore, their computation is very expensiv
these methods to small sequences (<200 nucleotides). An option for de
complexity of the problem is to do threading: assume that some master stru
and thread the sequences onto it. In several important cases, like ribosomal
realistic assumption.

This approach can of course be taken using a SCFG based objective
Alternatively, one can consider a simpler function such as the one described
Michot(92). In this case, the evaluation of the alignment is split in two steps:
the primary sequence alignment (ii) evaluating the quality of the fold in
sequence of known structure onto the sequence of unknown structure. To
overall score, these two terms are combined with one another. Although this
real theoretical justification as opposed to those previously described, it ha
being conceptually simple. It can also accommodate a range of interactions su
knots and other non nested structures. In this case, the alignment problem
NP-complete(17). It is in order to provide a reasonable heuristic solutio
sequences (>1000 nucleotides) that we developed the package RAGA(97) (Se
and Annex 3).

4 MAKING MULTIPLE SEQUENCE ALIGNMENT

4.1 COMPLEXITY OF THE PROBLEM

So far, we have focused on reviewing some of the objective functions evaluating the quality of sequence alignments. However, as pointed out earli one side of the coin. The other one, that constitutes in fact the main b optimization. In other words, given an objective function, is it possible to c producing the best scoring alignment? There are at least two good reason efficient and accurate optimization strategies. The first one is obvious: alignments that are needed for whatever purpose.... The second reason is le extreme importance. The evaluation schemes described above are only theore using phylogenetic or structural. criteria. As such, they do not constitute must therefore be validated through empirical analysis (i.e. how well do they

The optimization methods required for these two reasons do not necessar equivalent. One can, for instance, use a very robust but expensive (compu memory) method to compare and validate alternative scoring schemes. If schemes proves useful it may later become appropriate to develop a very sp method that approximates the optimization reasonably well while being effici production purpose. Needless to say, whatever direction one wishes to take, an optimization technique will always prove to be a very demanding problem

We already mentioned that even for two sequences of moderate length, a nai alignment computation can lead to impractical enumeration problems. Fo situation does not have to be that bad and will depend on the scoring schem optimize. In many cases, there are short cuts that allow efficient compu alignment, given some specific objective functions. We will see in section dynamic programming(18) is one of these techniques that allows the co pairwise alignments in time proportional to the product of the length of the This essential technique constitutes the core of many alignment methods. I not restricted to two sequences, but since its complexity is a function of the length of the sequences to align, it can hardly be used for more than three time(55). This does not mean multiple sequence alignments cannot b automatically with dynamic programming, but it means that to do so, one heuristic algorithms.

Heuristic methods do not guarantee an optimal solution but may perfo sometimes even guarantee the solution to be inside given boundaries. Gene multiple sequence alignment algorithms can be divided into two classes: (algorithms that usually rely on sequence clustering algorithms and dynamic for making progressive alignments (ii)the non-progressive algorithms tha simultaneously align all the sequences. These non progressive algorithms th into two distinct sub-categories:

-deterministic heuristics.
-stochastic heuristics.

In the following sections, the underlying principles of these algorithms ar differences will be briefly explained. More emphasis will be given to the gene techniques on which the SAGA package is based(98)(see section 5.1 and ann

4.2 DETERMINISTIC GREEDY APPROACHES

4.2.1 Aligning Two Sequences

The main algorithm for aligning two sequences, often referred to as the Ne Wunsch(18) or dynamic programming (DP) algorithm, is one of the olde important tools in bioinformatics. Over the last 30 years, it has been used u another in most of the methods developed for sequence comparison.

When applying dynamic programming to two sequences, it is possible to cor scoring alignment between these sequences using an amount of memo proportional to the product of the lengths of the two sequences to align. T dramatic improvement over the naive approach that would require enum possible alignments. An important advantage of dynamic programming is t general scheme. Given substitution costs (e.g. matrix or profile...) and a scoring gaps, the algorithm can compute the alignment with the best score. can accommodate any context-independent scoring scheme.

The algorithm is based on extending recursively the best scoring alignmen residues of each sequence have been aligned. In practice, this means findin through a matrix constructed from the scores of all pairs of elements bet sequences. Let us consider two sequences, A of length m and B of length n, assigns a score $d_{i,j}$ to the substitution of residue i in A by the residue j in penalty g. Computation of the optimal score is achieved by incrementally e path with a locally optimal step. For example; element $d_{i,j}$ can extend any pa in the preceding row (i-1, m:m<j) and/or the preceding column (n, j-1:n-1) penalty. If the extension that produces the highest new path score is alwa applying the condition repeatedly (to every element in the sequence matrix sequence matrix of pairwise match-scores into a sequence matrix of path-sc can be expressed more formally by the recursive expression:

$$S_{ij}=d_{ij}+\max \{ (S_{i-1,j-1}), (S_{i-1,m} -g, (m<j-1)), S_{n,j-n} -g, (n<i-1)) \} \quad \text{(eq. 3)}$$

Once the global score S_{ij} has been found, the optimal alignment can be retrie traceback from the cell with the highest score. This algorithm is slightly more the original Needleman and Wunsch that did which used simpler gap Computation is quadratic in time and space for the score. Gotoh propose version of this algorithm that allows computation of score in linear space affine gap penalties(45). Finally, a recursive algorithm originally de Hirschberg(99) and refined by Myers and Millers(100) allows computation and alignment in linear space, using affine gap penalties. A vast number of v been proposed around the original algorithm. For instance, since it is kn mathematical optimal is rarely biologically optimal methods have been descr computation of sub optimal alignments(101-103). Such computations can when trying to assess the reliability of a sequence alignment(104). Anoth proposed by Taylor, involves biasing the DP toward motifs, in order to minir of the gap penalty choice(47).

DP is not restricted to global alignments but can also be used for finding th sequences shared between two otherwise unrelated sequences. This algo extremely important for database searches. It is mostly an adaptation of the Wunsch algorithm and involves extending a path as long as the alignment improving (hence the 0 in the following equation):

$$Sij=\max \{ (S_{i-1,j-1})+dij, (S_{i-1,m} -g, (m<j-1))+d_{ij}, S_{n,j-n} -g, (n<i-1))+d_{ij}, 0 \}$$

The score of a path is set to 0 if this path gets a negative score (the subst needs to give negative values to unfavored substitutions). Best segments are

finding the best scores in the score matrix and making tracebacks, using 0 for the path.

The use of DP is not restricted to sequences. For instance, the Viterby algori aligning a sequence to a HMM(80) is an adaptation of the classic DP algorith modified versions of DP can be used to align structures. This was shown b Michot who described a DP strategy for aligning two RNA molecules taking i potential secondary structures in one of the sequences while knowing the s other(92). Taylor and Orengo(105) also described a heuristic based programming designed to align two protein structures (double dynamic p Finally, the algorithm used to align an RNA molecule to a covariance model on DP and is very similar to the one described by Zuker for predicting RN Most of these algorithms are computationally very demanding (higher co regular DP). For this reason, alternative techniques (like stochastic optimizati can be sensible alternatives or even necessities when dynamic programmi used.

Nonetheless, in the case of pairwise sequence alignments DP may provi efficient and accurate way to compute a solution, as long as traditional gap p are used. We will now see that how this pairwise method can be extende sequences.

4.2.2 Aligning Two Alignments : Progressive Alignment Methods

Although it is hard to align more than two sequences at a time by regu programming, it is possible to align two alignments (i.e. two sets of prealign with a pairwise algorithm. This is due to the fact that an alignment (pairwis can be regarded as a generalized sequence where instead of having one position, one has a residue vector. Such a definition allows the alignm alignments, treating them like normal sequences(69). It makes sense when d each alignment as a profile.

These methods for aligning multiple alignments are very important because to one of the most popular multiple alignment strategies: the progressiv algorithms described by Feng and Dolittle(74) and Taylor(75). The pri progressive alignment is quite straightforward. Since it is impossible to pr multiple sequence alignment by DP, using all the sequences at the same ti begin by aligning pairs of closely related sequences. Such pairwise alignm with very similar sequences are likely to be correct if the sequences are simi the second stage, pairs of closely related pairwise alignments will then be a two, and so on until the whole multiple sequence alignment has been produc order in which the sequences are aligned depends on their phylogenetic rel phylogenetic tree) and the procedure will follow the topology of some es Several methods exist that allow the computation of such trees using vari algorithms that rely on an estimates of the pairwise distances between the 106).

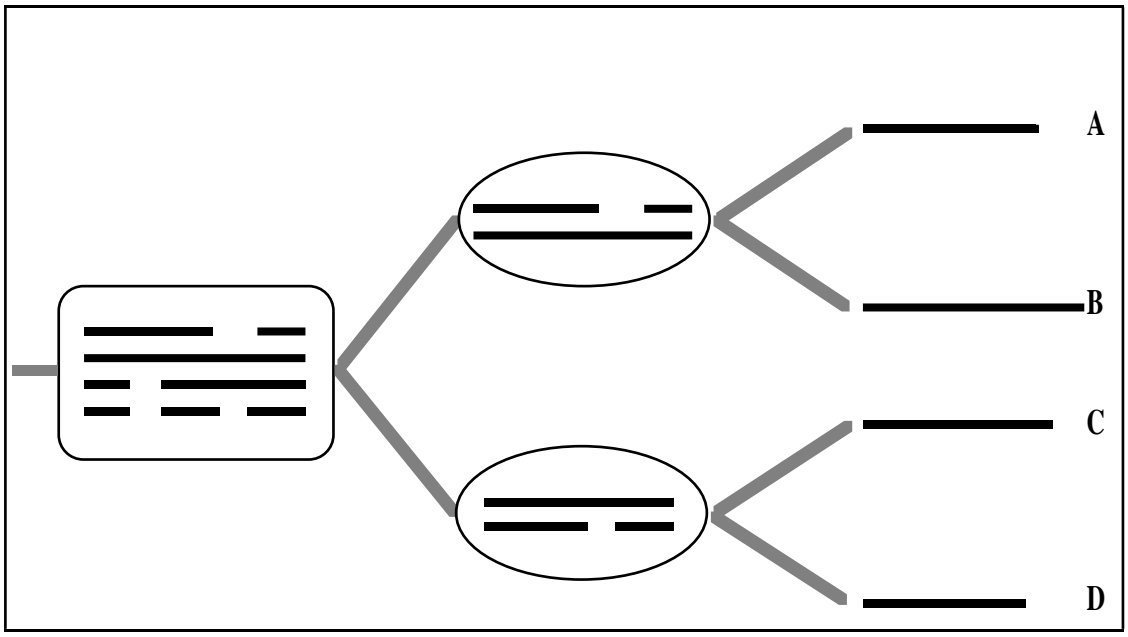


Figure 7. Progressive alignment strategy(74, 75). A phylogenetic tree is estimated f four sequences A, B, C and D (gray lines). Its topology is used to decide the order fo the sequences by dynamic programming.

The advantage of progressive alignments is that they provide an efficient overcoming the problem of computational complexity posed by global techniques. However, if the aim is, for instance, to optimize the sums of substitution matrix and a pair of gap penalties, one has to be aware that t method does not offer any guarantee. Its performances may be unpredict depend on the quality of the intermediate alignments. These will depend relations between the sequences and the accuracy and the density of the tre alignments are not necessarily correct (in terms of optimization) because t using only a subset of the information available. This may have serious conse early mistakes will never be corrected ('once a gap always a gap') and may a by inducing more mistakes in further intermediate alignments. This prob referred to as a 'local minimum problem', in the sense that the method by i pushed toward satisfying short term constraints that may lead to a non globa

Nevertheless, in many cases, this progressive strategy leads to convincing al low computational cost. There are two good reasons for that. First of all, in the sequences are well fitted to a progressive approach, and secondly becaus optimality does not necessarily mean better biological alignments. This expla of the most popular multiple sequence alignment packages are based on th They include Clustal V(77), Multal(75),ClustalW(21),Barton and Stenberg(1 wide variety of cases these methods have been shown to do well when their are assessed using sets of reference alignments. Many of the techniques used multiple sequence alignments such as position specific gap penalties, tree b secondary or tertiary structure information can be implemented in a progre strategy as shown in ClustalW(21) and in the work of Barton and Stenberg(1

One of the problems of progressive alignments is that their accuracy will clea the relation between the sequences aligned. If these are closely related, alignments are likely to be accurate and to induce a correct overall alignmer hand, if the sequences are only remotely related, it may be necessary to simu all the information they contain to build a realistic model (i.e. most of sequences needed to build the intermediate alignments will not contain enou for inducing a correct pairwise alignment). The only way to improve the rel

the sequences in a set is to incorporate some more sequences into this set. cases this is not an option, simply because the sequences are not present ir and may not even exist. Simultaneous sequence alignment methods are there We will now describe methods that attempt to produce this type of multi alignment.

4.3 DETERMINISTIC APPROACHES FOR NON PROGRESSIVE MULTIPLE ALIGNMENTS

4.3.1 The Carrillo and Lipman Algorithm

The basic DP algorithm is hard to apply to more than two sequences a the computation of an SP multiple alignment using DP has been shown complete(19). In order to compensate for these limitations,. an algorithm wa Carrillo and Lipman(57) to perform such operations on a larger number of se 10). It relies on the idea that the whole alignment space does not necessar explored when aligning sequences. Bounds can be derived that guarante optimal alignment inside a portion of the hyper space defined by the seque Several methods have been described for deriving guaranteed boundari Unfortunately, these boundaries are often too loose to be practical (i.e. they that is too large to be explored). To be of any use, the Carrillo and Lipman al to rely on tighter heuristic boundaries that do not guarantee optimalit implemented in the package MSA(22, 109). Their main property is that they level of similarity of the sequences. For very closely related sequences, the b be extremely tight and define a very small portion of the hyper space. If similarity decreases, the boundaries become looser, until the computation o requires the exploration of a space beyond computational ability.

The limit on the number of sequences and on the level of similarity sha sequences constitutes the major drawback of MSA. It means that in most ca alignments that are within the scope of the program are those for which progressive alignment can be made because the sequences are related enou important drawback of MSA is that it is restricted to a very specific type function. This is especially drastic with regard to the type of gap penalties M use ('semi natural gap penalties' instead of 'natural gap penalties', see secti

4.3.2 Other Approximation Techniques

More recently, non-dynamic programming based methods have been propos to generate multiple sequence alignments by solving graph theory problems techniques is known as the MWT (Maximum Weight Trace) formulation, in 1993 by Kececioglu(110), (111). Like dynamic programming, the solving known to be NP complete. But even so, using a branch and bound algorit based algorithm has been described(111) that manages to align up to 15 s reasonable amount of time. Unfortunately, this approach suffers from the s as MSA: it requires tight boundaries to be established, which is not alwa especially when the sequences have very low similarity.

Considering all this evidence, it is clear that in the years to come, our ab multiple sequence alignments will increase along with computer power. Th these methods will be able to handle more and more sequences. On the ot problem, that in most cases has been shown to be NP complete is likely unsolvable in polynomial time. This implies that the number of sequences a strategy is able to handle will always remain severely limited. Another seriou the fact that in many cases these algorithms are very specific for a given obj and difficult to adapt. We will now focus this review on stochastic techniqu proven to be much more flexible and versatile tools.

4.4 STOCHASTIC HEURISTICS

4.4.1 What is a stochastic Method ?

A stochastic optimization strategy (also known as Monte Carlo simulation) finding a solution to a problem through some form of random sampling. which the search is random mostly depends on the heuristic one uses. The randomness is that there is a non null probability for any potential solution regardless of the solution space size. Of course, the randomness also impli solutions may not be looked at (including the global optimum). In order to problem, a large number of heuristics have been designed that attempt to l which the solution space is sampled. The aim is to improve the chances of optimal solution. As a consequence, most stochastic strategies can be regarde between greediness and randomness.

This can best be explained considering a solution landscape with hills and va The optimal solution is the top of the highest hill (or the bottom of the dee there is only one hill, one can find this point by simply climbing along gradient. Such a method does not have to be stochastic and is known as However, if there are several small hills and a high one, the outcome of hill be restricted by the point where it was started. The action of climbing the necessarily lead to the optimum, but to the first local optimum. This is the p the excessive greediness of a hill climbing strategy that always goes up. I element of randomness that decreases the greediness of the strategy, it may possible to take alternative routes (going down, for instance, which in turn search to encounter the highest hill). This less greedy technique is known as Climbing. Ultimately, its success (as for any stochastic strategy) will mostly d fitness landscape. The more complex this landscape, the hardest it is to fi optimum.

Stochastic methods are quite well suited for solving sequence alignment pr the landscape is very complex(113) and the number of solutions much to exhaustively analysed(19). Many of these techniques are inspired from physi Annealing), biological (Genetic Algorithms, Evolutionary strategies), or even s search) phenomena, others are simply based on iterative statistical analyses sampling (stochastic expectation-maximization strategy). In this section, w some of the applications of stochastic heuristics to the problem of sequen Since one of these techniques (Genetic Algorithm) is central to the work pre will be given a more detailed presentation in the next section.

A very important advantage of many Monte Carlo strategies over more tradit is the fact than they often allow conceptual separation between optimizatior function. Ideally, the optimization strategy will play the role of a black box i objective function can theoretically be plugged. The main drawback of thes that they rarely give any indication of how complete the optimization is. In f of these methods are iterative, it is usual practice to stop them if the search specific number of iterations. The general lack of guaranteed criteria for stop is usually the main problem. It can sometimes be overcome if the score obt objective function contains some information regarding the complete optimization.

4.4.2 Iterative alignments and Expectation-Maximization Strategies

Iterative strategies are heuristic approaches, very similar in spirit to E (Expectation maximization). Expectation-Maximization is a Hill climbing st involves two steps: (i) given a model and some data, estimation of new para

model, (ii) given the new parameters and the ancient model, estimation of
The procedure is carried on iteratively until the model and the parameters st
between the model and the parameters). Depending on the implementat
incorporate some stochastic steps as do iterative alignment strategies. The
strategies is quite large. The optimization of RNA covariance models is r
EM(91), as well as the training of HMMs in SAM(81). Several multiple
alignments strategies have also been described based on EM(114, 115)

To perform an iterative multiple sequence alignment, two components are n
prealigned sequences and an algorithm for aligning pairs of sequences
alignments) such as DP using Gotoh's operators for aligning profiles and al
73). An iterative alignment algorithm usually takes the following steps:

1-split the multiple sequence alignment into two groups of prealigned s
2-apply the DP algorithm to the two multiple alignments
3-if the score is not stabilized, go to 1

Since the number of possible partitioning in step 1 increases factorially with
sequences, exhaustive analysis cannot be done, and one has to use random
other types of partitions that use for instance the phylogenetic tree. Seve
schemes of this type have been described (for a review, see (116)) These
better on average than progressive strategy (in terms of scores) but are als
(100 times on average). Recently, Gotoh presented such a technique named
randomized iterative strategy(117). He uses it to optimize the weighted sums
reevaluation of the weights after each iteration. According to the author, a
reference structural alignments, this technique is significantly more ac
progressive alignments strategies.

Gibbs sampling is a more sophisticated example of stochastic EM(118). In Gil
applied to multiple sequence alignment(83), a random solution is generate
guide the generation of the next solution (ungapped multiple alignmen
probabilities to all the neighboring solutions, and randomly choosing one of
exactly, choosing a new solution according to their probabilities, given the c
The search is iterative and has been successfully implemented for identify
motifs conserved among a set of sequences. In such a context, given a set o
sequences, the aim is to generate the more informative model (the less likel
obtained by chance). The objective function described in that context by Law
purely based on statistical analysis of the data and can be made fully indep
prior information. In many respects it is very similar to HMMs.

4.4.3 Simulated Annealing

One of the first stochastic techniques described was simulated annealing (S
SA relies on an analogy with physics. The idea is to compare the solving of ar
problem to the cooling of a metal (i.e. finding the best position of each atom
metal is equivalent to solving a multi constraints problem). The princ
straightforward. Given a function, a random solution will be generated, and
chosen. Every iteration, a new solution is created based on the previous on
random modification algorithm). If the new solution is better than the old
accepted. Otherwise, it will be accepted/rejected with a probability that de
temperature and the level of disimprovement (i.e. the higher the tempera
likely a solution is to be accepted, regardless of its quality). Temperature wil
cooled down. This means that at the beginning, almost any new solution is a
in the last phases (low temperature) only the solutions bringing an imp
accepted. The SA algorithm can be summarized as follow given that Sc is
solution at the n^{th} iteration:

1) Generate an initial solution Sc=So and an initial Temperature (To), T
2) Modify the solution Sc into Sn
3) IF F(Sn) is better than F(Sc) then keep the new solution: Sc=Sn
ELSE
keep Sc with a probability $P=e^{-(|F(Sn) - F(Sc)|)/Tc}$
4) Change the value of Tc according to the annealing schedule
5) Go to 2 as long as the termination criterion has not been met.

The strategy depends on three parameters: the initial temperature, the cooli the acceptance function. The cooling and the acceptance function define th SA. The most common are those using Boltzman distributions, but one c Gaussian or other ad-hoc distributions(121). A remarkable property of SA i cooling function slow enough, the search is guaranteed to reach the optim course this argument remains theoretical since in most of the cases an ade schedule would require too much time to be of any practical use (hence th SA).

Despite this intrinsic limitation, SA has been applied to multiple sequence several occasions(122-124). The conclusion of these studies has mostly been it does reasonably well, because of its slowness, SA has to be restricted alignment improver method rather than a full alignment method. In all thes applied to sequence alignment in a similar fashion: a multiple alignmer (randomly or using some heuristic) and modified using specific sub-routir gaps around or insert them. We will later see that these 'modification proc many ways similar to the mutations used in a genetic algorithm strategy.

As opposed to the iterative methods previously discussed, SA is not restrict SP multiple alignments using DP. On the contrary, the method displays properties of a black box. If it was not so slow, SA would possibly be the idea any optimization problem. It has successfully been applied to the alignm molecules using potential secondary structure information(90). Similarl methods(125) have been described for predicting the fold of very long RNA are beyond the scope of traditional methods like Zuker's. Finally, a variant c maximization based on SA is used for the training of HMMs(126). In the wo here (SAGA and RAGA), SA has been an important source of inspiration, be similarities that exist between SA and GAs.

GAs are probably some of the most interesting stochastic optimization to today. They can best be described as a very flexible framework in which available methods (deterministic or not) can be integrated in order to cons tool. One of the reason why GAs have received so little attention in the conte sequence alignment is probably due to the fact that the implementation algorithm specialised for multiple alignment is much less straightforwa simulated annealing. In other areas of computational biology, GAs have established as powerful tools. These include protein 3D structure predic secondary structure prediction

4.5 GENETIC ALGORITHMS

4.5.1 What is a Genetic Algorithm?

Genetic algorithms are based on a loose analogy with the phenomenc selection. They have existed in one form or another for quite a while, but introduced by Holland in 1975(127). The principle is quite straightforw problem, a set of potential solutions (population) compete with one anoth

These solutions can be modified (mutations), or combined with one anothe
The idea is that acting together, selection and evolution will lead to an overal
of the population. Most of the ideas developed here about GAs are taken fror

There are two essential concepts at the heart of the GA strategy. The first or
Selection is established in order to lead the search toward improvement. It
best solutions to the problem (as judged with the objective function one is int
be selected according to their quality. If one was to do so in a determini
instance, by only selecting the best solution every generation, the search
rapidly converge toward the first local optimum it encounters. In such a cas
mostly behave like Hill Climbing. To avoid this major problem (called
convergence), the selection procedure in a GA is not absolute but statist
statistical selection is a very straightforward process. In a first step, each
evaluated using the objective function. The score obtained that way is turn
fitness measure. In the selection round, an imaginary wheel is spun where e
has a number of slots proportional to its fitness. This means that individua
high score are the most likely to be selected, while those with a low score a
However, everything is possible and it is this uncertainty that prevents
converging prematurely to the closest local minimum.

The second key concept in a GA is the concept of evolution. The aim of the c
create new solutions based on those present in the population. By analogy to
are two sets of operators: the crossover (combining two individuals) and t
Since crossovers can be regarded as a very primitive interpretation of sexual
it is current practice to apply selection when choosing the two individuals th
be combined. When doing so, one hopes that the child produced that way wi
qualities of both parents.

In practice, a genetic algorithm follows a series of cycles known as generat
each cycle, individuals are evaluated and used to create the next genera
selection and operations. The variations one can put inside such a scheme
infinite. Different schemes can be used for turning a score into a fitness mea
can be made stronger or weaker. Many of the effects of such parameters hav
in detail. However there is no solid argument for one model of GA to out
others. In fact the choice of a model seems to be mostly problem dependent
our experiments that if the problem is suited to combinatorial optimization
GA does reasonably well. However, some do better than others and the appr
needs to be done in what is mostly a trial-and error-strategy.

There is very little theory around the reasons why GAs perform efficient opti
instance, as compared to SA, there is no formal proof that a GA will reach
solution given enough time. The most popular concept to explain the perfor
is the notion of building blocks. The concept is best understood if we consid
a coding system that consists of binary strings (i.e. each individual is a chron
each gene can have two allelic values 0 or 1). In such a context, a block is d
stretch of 0s and 1s present in a chromosome. The GA strategy through
selection and combination makes it possible for important blocks to be
increase the number of their copies in the population). This in turn will incre
for these blocks to be extended through mutations and crossovers. The more
(i.e. the better the score it induces), the more likely it is to spread in the p
would be the equivalent of the fixation of a group of alleles in an animal popu
context mutations help create new blocks, or restore blocks that may have
fact that the number of existing blocks is much larger than the population i
phenomenon known as implicit parallelism.

This GA procedure can also be looked at through another angle. If we co
solution as a vector in a hyperspace, with each piece of the chromosome be
coordinates, we can view the search as a form of multivariate analysis. In

building blocks are sets of coordinates that restrict the search to some smalle
hyperspace that has a higher concentration of good solutions than the rest. V
goes on, this hyperspace becomes smaller and smaller until it cannot be redu

The theory of the building blocks was initially proposed for the simple Gene
developed by Goldberg(128). For this reason, it is restricted to binary ch
where each gene only has two allelic values (1 or 0). It is hard to extrapolate
the more complex representations required by sequence alignment proble
empirical evidences suggest that the building block theory constitutes a
approximation when analysing the behavior of more complex representati
should also be stressed that although it provides an elegant way to answer t
how a GA works, the theory of the building blocks is useless for predicting
type of problem the GA approach will fail or succeed.

Deceptivity is an area of GA research that has received some significant atte
last few years. It amounts to defining the conditions in which a GA will fail a
correct solution. What these studies have shown is that like any optimizatio
GA is sensitive to the fitness landscape of the function one is interested in.
established what would intuitively be assumed. When there is very little co
function one is interested in, the optimization becomes much harder or imp
other hand, given a complex problem, there are often alternative ways o
landscape. In fact, the landscape will depend on the representation of a solu
means one uses to walk along the landscape (operators in our case). When
coding, methods that attempt to reshape the fitness landscape in order to
continuous are known as gray coding. More generally, whatever the prob
representation, efforts should be made so that neighboring solutions
comparable fitness. Defining the neighborhood of two solutions is a complex
depends on two factors: the representation system and the operators used.
having a representation that induces a very smooth fitness landscape if the o
as well designed to sample close neighbors...

For instance, in the case of a multiple sequence alignment, there are many
define such neighbors. The neighbor of a given alignment will be anothe
created when inserting a gap or when shifting an already existing gap. The co
landscape explored during the search will depend on the shape of the fitn
these operators define (i.e. do they induce any gradient that can be followe
that a mutation that inserts a gap at a random position (thus potentially gene
a whole sequence) does not explore the same neighborhood as a mutation th
a gap by one residue. This is a serious problem, and in practice defining th
operators constitutes the main difficulty when designing a Genetic Algorithm

4.5.2 Applications of GAs in Sequence Analysis

Although the concept of GA is relatively new in the context of sequence alig
themselves have been applied to sequence analysis on several occasions, m
structure predictions problems for DNA, RNA and proteins. This makes
structure predictions problems fit the concept of GAs very well. A protein st
described as a list of amino acids associated with some conformation va
conformation values may be angles of chemical bonds. In such a context, the
representation can be the list of the bonds that need to be characterized (
allelic values being the actual angles in a given conformation. These problem
ones, mostly because it is hard to create an objective function that accurate
protein folding process. Several attempts have been made that suggest GAs a
best suited optimization strategies for ab initio folding(130-136). Recently
been proposed for docking analysis using a GA(137). Finally, RNA foldin
received some attention(87, 88, 138, 139). Generally speaking, the use
algorithms in sequence and structure analysis is rapidly expanding. For inst

records about 110 publications describing GA based methods for sequence a
period of 8 years [1989-1997]. Out of these 110 publications, 45 appeared
alone.

In terms of sequence alignment, the only work we are aware of is an attem
iterative alignments using a genetic algorithm in order to speed up the proc
from SAGA, no method has been described which is able to directly perform
the alignment and to accommodate non-standard objective functions. We w
review the motivations behind SAGA as well as the work done to validate o
and extend its scope to a larger variety of objective functions and problems.

5 COMPARISON OF THE METHODS

We are only aware of two general studies made to compare systematica
alignment methods(117, 140) .The work by McClure et al. was made by co
performances of seven different methods(22, 23, 74, 77, 107, 110, 141) v
four different protein families (Hemoglobin, RibonucleaseH, Kinases ar
proteases). The methods were evaluated for their ability to correctly align s
known structural or functional importance in each family. The conclusio
proteins with more than 50% identity can usually be correctly aligned, reg
methods. When identity decreases, all the methods become affected by th
sequences (more sequences do not necessarily improve the results). T
conclusion is that progressive methods doing global alignments are usua
However, the test sets were not large enough to really allow strong distinctio
between the methods ability and the non progressive methods tested were s
development stage.

A main drawback of the approach taken by McClure was the limited set
alignments on which their comparison was based, and the fact that these alig
been established by structural analysis. It has now become standard proced
new methods by comparing their output with reference alignments available
structural databases (33, 142-144). Gotoh recently presented such a
(117)based on structural databases(33, 142). In his study, he compared fou
based on progressive alignments (21) , two iterative alignment strategies
previously described(73, 145) and a new one that involves iterative ali
dynamic reevaluation of the weights and the phylogenetic tree for optimizin
sums of pairs with affine gap penalties. The methods were applied to 54 pr
The conclusion was that iterative alignments methods do on average a bi
ClustalW(21). It should be noted that the study did not make any use of the
Previous results obtained with the MSA program(22) indicate that MSA
outperform ClustalW(98),(Barton, personal communication). This suggests th
the quality of the sums of pair optimization provided by the iterative alignm
(i.e. global optimization) the improvement is probably due to the dynam
scheme used by Gotoh that allows a better use of the information. The weigh
strategy are similar to those described for the MSA program(66). In MSA, th
are computed from an initial progressive alignment which is probably less
those obtained by Gotoh on optimized alignments and used for reevaluatio
may also be that the iterative strategy used by Gotoh does not lead to n
optimality, but to sub-optimal alignments that are biologically more accurate.

6 SUMMARY OF THE CONTRIBUTIONS

6.1 SAGA: MAKING MULTIPLE SEQUENCE ALIGNMENT BY GENETIC ALGORITHM

SAGA is a package designed to perform multiple sequence alignments usi
algorithm. The method involves evolving a population of alignments
evolutionary manner and gradually improving the fitness of the population
objective function which measures multiple alignment quality. SAGA uses a
scheduling scheme to control the usage of 22 different operators for combin
or mutating them between generations. When used to optimize the sums of
function, SAGA performs better than some of the widely used alternative p
MSA. This is seen with respect to the ability to achieve an optimal solution.
attraction of the approach is the ability to optimize any objective functio
invent.

This last point was the main motivation behind the development of SAG
alternative packages exist that can provide reasonable optimizations of
alignments, for instance. However, it is known that in cases where the seque
remotely related, these methods fail, not necessarily because of an optimiza
but also because traditional SP objective functions may not be well adap
alignments. Another potential application for a good quality optimization
hidden Markov model training. The EM based algorithms used at present ar
very sensitive to local minimum problems. This problem could easily be
using a GA based training scheme.

The validation of SAGA as an optimization tool was made using the prog
Remarkably, starting from unaligned sequences, SAGA was always able to
MSA results or to outperform them. As far as we know, SAGA is the only o
method that can achieve this result without using dynamic programming.
fair to say that SAGA is not yet ready for systematic use and does not
challenger to programs like ClustalW or other progressive alignment pack
other hand, SAGA is a powerful tool for analysing new objective functions a
the quality of faster heuristic methods. In the long term, we hope that SAGA
much more practical, thanks to the increasing power of the computers
improvements made in the algorithm (parallelisation, better seeding, optimiz

The original version of SAGA has been available for just over
(http://www.ebi.ac.uk/~cedric/saga_hp.html). The package now regroups a
known users of about 30 people. The software is supported and a new releas
for the beginning of 1998.

6.2 COFFEE: IMPROVING ON EXISTING OBJECTIVE FUNCTIONS

COFFEE is a natural extension of SAGA. It is an attempt to design a new type
function for evaluating the quality of multiple sequence alignments. Ther
number of objective functions described for evaluating multiple sequence ali
all have qualities and drawbacks. With COFFEE, the aim is not to add one m
function to the list, but to propose a scheme that makes it possible to combi
scoring schemes. The COFFEE score reflects the level of consistency betwee
sequence alignment and a library containing pairwise alignments of the sa
We show that multiple sequence alignments can be optimized for their COFF
the genetic algorithm package SAGA. The function is tested on 11 test ca
structural alignments extracted from 3D_ali(33). On 9 of these test cases, S

COFFEE function is able to outperform ClustalW (progressive multiple alignments) as judged by comparison with the structural references. We alsc COFFEE score can be used as a reliability index on multiple sequence alignme we show that given a library of structure based pairwise sequence alignme from FSSP, SAGA can produce high quality multiple sequence alignments.

An important issue in COFFEE is the validation of an objective function. T should not be confused with the evaluation of an optimization strategy, as SAGA. An objective function may be optimized correctly but lead to biologica alignments. Verifying the correctness of an alignment is a complex process, the use of biologically correct references. Since the closest thing to a biolog alignment is a structural alignment, we used such alignments to validate approach (3D_ali).

COFFEE is powerful mostly because of its openness. It can accommodate vi type of modification, including position specific weights and different type weights. Furthermore any alignment making technique (pairwise, multiple, can be integrated into the COFFEE framework and several otherwise in techniques can be combined together.

6.3 RAGA: THREADING RNA SECONDARY STRUCTURES

The distinction between sequence and structure alignments has already b here. It seemed like an interesting question to ask whether SAGA would accommodate a structure based objective function. The fitness landscape of s is likely to be very different from those used for protein alignments. We technique to RNA secondary structure because this is a much simpler proble threading. Even if the algorithmic complexity is the same, the problem ca formulated in a fairly accurate way, thanks to the fact that RNA secondary mostly based on Watson and Crick base pairing.

In RAGA, we describe a new approach for accurately aligning two homolc sequences, when the secondary structure of one of them is known. To do so two software packages called RAGA and PRAGA which use a genetic algorithr to optimize the alignments. RAGA is mainly an extension of SAGA. In PRAG genetic algorithms run in parallel and exchange individual solutions. This me to optimize an objective function that describes the quality of a RNA pairwi taking into account both primary and secondary structure, including pse report the results obtained using PRAGA on nine test cases of pairs of euk subunit ribosomal RNA sequence (nuclear and mitochondrial).

The parallel implementation is described in detail in the corresponding p involves a set of synchronized RAGA processes running on different machine set of sequences. Every N generations (typically 5), the processes exchange fittest individuals. The main originality of this implementation is its relative only involves a population sharing scheme between several GAs and does n low level modification of the GAs. Of course, the parallel program is very di the original GAs (i.e. it induces a very different population structure). Despi parallel version has properties very close to those that would have been e RAGA parallelized at a lower level (speed and accuracy). The parallelization made very general. It can be used with SAGA and does not depend in any objective function. We plan to maintain this module so that it remains compa developments made on SAGA or RAGA. This type of parallelisation (know parallelisation) is not completely new, but we are not aware of any previous a model strictly identical to ours. Although no systematic work has yet b accurate characterization, we found that on the RAGA objective function, tl much more significant than when the parallel module is applied to SAGA. Thi do with the fact that our parallelisation makes it easier for a GA to analyse

fitness landscape, such as the one required by a structure based objective fu
and PRAGA have been made available over the W
(http://www.ebi.ac.uk/~cedric/raga_hp.html)

In its present form, RAGA is mostly a prototype. There is little interest in a
align two ribosomal RNA sequences at a time. Large accurate alignments exis
be used in order to guide the characterization of a new sequence. This has
motivation for the last project presented here: analysing large multiple aligr
to optimally use the information they contain. However, out of a ribosomal
the requirement of knowing the guide structure of the sequences to align is
many interesting cases, such prior information will not be available, and will
extract when sequence identity is very low. For this reason, we plan to
algorithm in order to use an objective function similar to the one described ir
that allows the alignment of two RNA sequences using primary and potent
structure. Finally, applying RAGA to the problem of Protein Threading is
natural continuations of this work. We plan to do so in the context of t
phDthreader(20).

6.4 OPTIMIZING RIBOSOMAL RNA PROFILE ALIGNMENTS

The purpose of this project is to show that the use of the information contai
can be optimized when trying to introduce (align) a new sequence. The mair
decide how each sequence of the alignment should contribute to this new alig

It is quite obvious that sequences closely related to the one of interest sho
higher weight. For this reason, we developed a weighting scheme based on t
the distances between the new sequence and the rest of the sequences. This c
pairwise weighting scheme (cf. section 3.2.3) quite similar in fact to the one
COFFEE. On the other hand, there is also a need to avoid bias due to
representation of different taxa. This effect is achieved by using tree based
attempt to correct for unequal representation.

We found that the ideal weighting scheme, in a dynamic programming conte
be a combination of these two. Attempts were also made to use position
penalties. In a DP context , it is hard to take into account secondary structur
be done in a second step that will involve implementing a profile based objec
the RAGA/PRAGA framework. In the long term the alignment method is mear
available over the WWW through a JAVA-based server.

CONCLUSION

This work provides further evidence for the usefulness of genetic algorithms
analysis context. GAs mimic very efficiently the human approach that is ma
error and of a continual attempt to combine partial results in order to glob
solution. The work done here is only preliminary, and there is still a lot to b
SAGA or RAGA come into everyday use. However, this may happen if the a
improved in terms of speed and if the developments of COFFEE keep the pr
preliminary results.

Putting aside the slowness that constitutes their main drawback, GAs have a l
properties. They can accommodate a large variety of problems and are eas
with other alternative methods. They also allow a conceptual barrier to be
biological and computational problems. Optimizing a function is purely a fo
that does not teach us much about biology. On the other hand, designing
objective function can be extremely informative about the problem analyse
understanding the type of constraints that occur during evolution.

It is along these lines that I plan to extend the genetic algorithm strategy to
range of problems, namely genome alignments and motif discovery. In a per
complete genome sequences are published every month or so, there is a sur
number of tools allowing one to compare these genomes. It is unfortunate
comparisons can be extremely informative in understanding the way genom
way functionally related sequences get clustered or not. In terms of comput
difficult problem, because of the nature of the events that occur duri
(inversion, transpositions, deletions, insertion) that are almost impossible t
traditional techniques. I believe a GA approach is very likely to provide a
solution to such a problem.

REFERENCES

1. Lipman, D. J. and Pearson, W. R., *Rapid and sensitive protein similarity searches.* Science, 1985. 227: p. 1435-1441.

2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., *local alignment search tool.* Journal of Molecular Biology, 1990. 215: p. 403-410.

3. Bairoch, A., Bucher, P., and Hofmann, K., *The PROSITE database, its status i 1997.* Nucleic Acids Research, 1997. 25: p. 217-221.

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

5. Sonnhammer, E. L. L. and Kahn, D., *Modular Arrangement of Proteins as Infered from Analysis of Homology.* Protein Science, 1994. 3: p. 482-492.

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

7. Haussler, D., Krogh, A., Mian, I. S., and Sjölander, K. *Protein Modeling u Hidden Markov Models: Analysis of Globins*. in *Proceedings for the 26th Hawaii International Conference on Systems Sciences*. 1993. Wailea, HI, U.S.A.: Los Alamitos, CA: IEEE Computer Society Press.

8. Garnier, J., Gibrat, J.-F., and Robson, B., *GOR method for predicting protei secondary structure from amino acid sequence.* Methods in Enzymology, 1996. 266: 540-553.

9. Garnier, J. and Robson, B., *The GOR method for predicting secondary structure in proteins,* in *Prediction of protein structure and the principles of protein conformation,* F.G. D., Editor. 1989, Plenum Press: New York. p. 417-465.

10. Rost, B., Sander, C., and Schneider, R., *PHD - an automatic server for protei secondary structure prediction.* CABIOS, 1994. 10: p. 53-60.

11. Göbel, U., Sander, C., Schneider, R., and Valencia, A., *Correlated mutations residue contacts in proteins.* Proteins, 1994. in press: p. 000-000.

12. Neher, E., *How frequent are correlated changes in families of protein sequences?* Proceedings of the National Academy of Sciences, 1994. 91: p. 98-102.

13. Shindyalov, I. N., Kolchanov, N. A., and Sander, C., *Can three-dimensi contacts in protein structures be predicted by analysis of correlated mutations?* Protein Engineering, 1994. 7: p. 349-358.

14. Woese, C. R., Gutell, R., Gupta, R., and Noller, H. F., *Detailed analysis of higher-order of 16S-like ribosomal ribonucleic acids.* Microbiol. Rev., 1983. 47: p. 621-669.

15. Gautheret, D. and Gutell, R. R., *Inferring the conformation of RNA base pairs and triples from patterns of sequence variation.* Nucleic Acids Res, 1997. 25(8): p. 1559-64.

16. Gutell, R. R., Weiser, B., Woese, C. R., and Noller, H. F., *Comparative anato of 16S-like ribosomal RNA.* Prog. Nucleic Acid Res. Mol. Biol., 1985. 32: p. 155-21

17. Lathrop, R. H., *The protein threading problem with sequence amino acid interaction preferences is NP-complete.* Protein Engineering, 1994. 7: p. 1059-1068.

18. Needleman, S. B. and Wunsch, C. D., *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J. Mol. Biol., 1970. 48: p. 443-53.

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

20. Rost, B., Schneider, R., and Sander, C., *Protein fold recognition by prediction based threading.* Journal of Molecular Biology, 1996. : p. in press.

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

22. Lipman, D. J., Altschul, S. F., and Kececioglu, J. D., *A tool for multiple seque alignment.* Proc. Natl. Acad. Sci. USA, 1989. 86: p. 4412-4415.

23. Taylor, W. R., *Multiple sequence alignment by a pairwise algorithm.* Computer Applications in Biological Science, 1987. 3: p. 81-87.

24. Corpet, F., *Multiple sequence alignment with hierarchical clustering.* Nucleic Acids Res., 1988. 16: p. 10881-10890.

25. Dayhoff, M. O., Schwarz, R. M., and Orcutt, B. C., *A model of evolutionc change in proteins. Detecting distant relationships: computer methods and results,* in *Atlas of Protein Sequence and Structure,* M.O. Dayhoff, Editor. 1979, National Biomedi Research Foundation: Washington, D.C. p. 353-358.

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

27. Vogt, G., Etzold, T., and Argos, P., *An assessement of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited.* J. Mol. Biol. 1995, 1995. 299(4): p. 816-831.

28. Henikoff, S. and Henikoff, J. G., *Performance evaluation of amino acid substitution matrices.* Proteins: Structure, Function, and Genetics, 1993. 17: p. 49

29. Rost, B. and Sander, C., *Prediction of protein secondary structure at better than 70% accuracy.* Journal of Molecular Biology, 1993. 232: p. 584-599.

30. Lüthy, R., McLachlan, A. D., and Eisenberg, D., *Secondary structure-bas profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities.* Proteins: Structure, Function, and Genetics, 1991. 10: 239.

31. Jones, D. T., Taylor, W. R., and Thornton, J. M., *A mutation data matrix transmembrane proteins.* FEBS Letter, 1994. 339: p. 269-275.

32. Tomii, K. and Kanehisa, M., *Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins.* Protein Engineering, 1996. 9: p. 27-36.

33. Pascarella, S. and Argos, P., *A data bank merging related protein structures and sequences.* Protein Eng., 1992. 5: p. 121-137.

34. Gonnet, G. H., Cohen, M. A., and Benner, S. A., *Exhaustive matching of t entire protein sequence database.* Science, 1992. 256: p. 1443-1445.

35. Benner, S. A., Cohen, M. A., and Gonnet, G. H., *amino acid substitution duri functionnally constrained evolution of protein sequences.* Protein Engineering, 1994. 7(11): p. 1323-1332.

36. Sjolander, K., Karplus, K., Brown, M., Huguey, R., Krogh, A., Saira, M Haussler, D., *Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology.* Comp. App. in Biosci., 1996. 12(4): p. 327-345.

37. Overington, J., Donnelly, D., Johnson, M. S., Sali, A., and Blundell *Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds.* Protein Science, 1992. 1: p. 216-226.

38. Benner, S. A., Cohen, M. A., and Gonnet, G. H., *Empirical and structural mode for insertions and deletions in the divergent evolution of proteins.* J. Mol. Biol., 1993. 229: p. 1065-1082.

39. Pascarella, S. and Argos, P., *Analysis of insertions/deletions in protein structures.* J. Mol. Biol., 1992. 224: p. 461-471.

40. Chou, P. Y. and Fasman, G. D., *Empirical predictions of protein conformation.* Ann. Rev. Biochem., 1978. 47: p. 251-276.

41. Gu, X. and Wen-Hsiung, L., *The size distribution of insertions and deletions in human and roent pseudogenes suggests the logarithmic gap penalty for sequence alignmnent.* J. Mol. Evol., 1995. 40: p. 464-473.

42. Altschul, S. F., *Gap costs for multiple sequence alignment.* J. Theor. Biol., 1989. 138: p. 297-309.

43. Waterman, M. S., *Efficient sequence sequence alignment.* Journal of Theoretic Biology, 1984. 108: p. 333-337.

44. Miller, W. and Myers, E. W., *Sequence comparison with concave weighting functions.* Bull. Math. Biol., 1988. 50: p. 97-120.

45. Gotoh, O., *An improved algorithm for matching biological sequences.* J. Mol. Biol., 1982. 162: p. 705-708.

46. Vingron, M. and Waterman, M. S., *Sequence alignment and penalty choice.* Journal of Molecular Biology, 1994. 235: p. 1-12.

47. Taylor, W. R., *Motif-Biased Protein Sequence Alignment.* Journal of Computational Biology, 1994. .

48. Wilbur, W. J. and Lipman, D. J., *Rapid similarity searches of nucleic acid and protein data banks.* Proceedings of the National Academy of Sciences, 1983. 80: 730.

49. Pearson, W. R. and Lipman, D. J., *Improved tools for biological sequence comparison.* Proceedings of the National Academy of Sciences, 1988. 85: p. 244

50. Smith, T. F. and Waterman, M. S., *Identification of common molecular subsequences.* J. Mol. Biol., 1981. 147: p. 195-197.

51. Pearson, W. R., *Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms.* Genomics, 1991. 11: p. 635-650.

52. Waterman, M. S. and Vingron, M., *Rapid and accurate estimates of statistical significance for sequence database searches.* Proc. Natl. Acad. Sci. USA, 1994. 91: p. 4625-4628.

53. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., M W., and Lipman, D., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. .

54. Bucher, P. and Hofmann, K., *A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system.* Ismb, 1996. 4(44): p. 44-51.

55. Murata, M., Richardson, J. S., and Sussman, J. L., *Simultaneous comparison three protein sequences.* Proceedings of the National Academy of Sciences, 1985. 3073-3077.

56. Gotoh, O., *Alignment of three biological sequences with an efficient traceback procedure.* J. Theor. Biol., 1986. 121: p. 327-337.

57. Carrillo, H. and Lipman, D. J., *The multiple sequence alignment problem in biology.* SIAM J. Appl. Math., 1988. 48: p. 1073-1082.

58. Sankoff, D., *Minimal mutation trees of sequences.* SIAM J. Appl. Math., 1975. 28: p. 35-42.

59. Fredman, M. L., *Algorithms for computing evolutionary similarity measures with length-independent gap penalties.* Bull. Math. Biol., 1984. 46: p. 553-566.

60. Altschul, S. F. and Erickson, B. W., *Optimal sequence alignment using affine gap costs.* Bull. Math. Biol., 1986. 48: p. 603-616.

61. Bairoch, A. and Apweiler, R., *The SWISS-PROT protein sequence data bank and its new supplement TrEMBL.* Nucleic Acids Research, 1997. 25: p. 31-36.

62. Vingron, M. and Sibbald, P., *Weighting in sequence space: a comparison of methods in terms of generalized sequences.* Proceedings of the National Academy Sciences, 1993. 90: p. 8777-8781.

63. Sibbald, P. R. and Argos, P., *Weighting aligned protein or nucleic acid sequences to correct for unequal representation.* Journal of Molecular Biology, 1990. 216: p. 8 818.

64. Felsenstein, J., *Inferring evolutionary trees from DNA sequences,* in *Statistical analysis of DNA sequences,* B.S. Weir, Editor. 1983, Marcel Dekker Inc.: New York 133-150.

65. Thompson, J. D., Higgins, D. G., and Gibson, T. J., *Improved sensitivity profile searches through the use of sequence weights and gab excision.* Computer Applications in Biological Science, 1994. 10: p. 19-29.

66. Altschul, S. F., Carroll, R. J., and Lipman, D. J., *Weights for data related by tree.* Journal of Molecular Biology, 1989. 207: p. 647-653.

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

68. Henikoff, S. and Henikoff, J. G., *Position-based sequence weights.* Journal Molecular Biology, 1994. 243: p. 574-578.

69. Gotoh, O., *Optimal alignment between groups of sequences and its application to multiple sequence alignment.* Comput Appl Biosci, 1993. 9(3): p. 361-70.

70. Neuwald, A. F., Liu, J. S., Lipman, D. J., and Lawrence, C. E., *Extra protein alignment models from the sequence database.* Nucleic Acids Res, 1997. 25(9): p. 1665-77.

71. Gibson, T. J., Thompson, J. D., and Heringa, J., *The KH domain occurs i diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding nucleic acid.* FEBS Letters, 1993. 324: p. 361-366.

72. Taylor, W. R., *A non-local gap-penalty for profile alignment.* Bull Math Biol, 1996. 58(1): p. 1-18.

73. Gotoh, O., *Further improvement in methods of group-to-group sequence alignment with generalized profile operations.* Comput Appl Biosci, 1994. 10(4): p. 379-87.

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

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

76. Higgins, D. G. and Sharp, P. M., *CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.* Gene, 1988. 73: p. 237-244.

77. Higgins, D. G., Bleasby, A. J., and Fuchs, R., *CLUSTAL V: improved sofwar for multiple sequence alignment.* Computer Applications in Biological Science, 1992. 189-191.

78. Bucher, P., Karplus, K., Moeri, N., and Hofmann, K., *A flexible motif sea technique based on generalized profiles.* Comput Chem, 1996. 20(1): p. 3-23.

79. Luthy, R., Xenarios, I., and Bucher, P., *Improving the sensitivity of the sequence profile method.* Protein Sci, 1994. 3(1): p. 139-46.

80. Rabiner, L. R., *A tutorial on hidden Markov Models and selected applications in speech recognition.* Proc. IEEE, 1989. 77: p. 257-286.

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

82. Hughey, R. and Krogh, A., *Hidden Markov models for sequence analysis: extension and analysis of the basic method.* Computer Applications in Biological Scien 1996. 12: p. 95-107.

83. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A and Wootton, J. C., *Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment.* Science, 1993. 262: p. 208-214.

84. Bryant, S. H. and Altschul, S. F., *Statistics of sequence-structure threading.* Current Opinion in Structural Biology, 1995. 5: p. 236-244.

85. Crick, F. H. C., *Central dogma of molecular biology.* Nature, 1970. 227: p. 56 563.

86. Richards, R. G., *5 S RNA An analysis of Possible Base Pairing Schemes.* Eur J Biochem, 1969. 10: p. 36-42.

87. Gultayaev, A. P., van Batenburg, F. D. H., and Pleij, C. W. A., *The com Simulation of RNA Folding Pathways Using a Genetic Algorithm.* J. Mol. Biol., 1995. 250: p. 37-51.

88. Shapiro, B. A. and Wu, J. C., *Predicting RNA HType pseudoknots with the massively parallel genetic algorithm.* Comp. Applic. in Biosci., 1997. 13(4): p. 459-47

89. Zuker, M., *Computer prediction of RNA structure.* Meth. Enzymol., 1989. 180: p 262-288.

90. Kim, J., Cole, J. R., and Pramanik, S., *Alignment of possible secondary structures in multiple RNA sequences using simulated annealing.* Comp. Applic. Biosci., 1996. 12(4): p. 259-267.

91. Eddy, S. R. and Durbin, R., *RNA aligment using covariance models.* Nucleic Acic Res., 1994. 22(11): p. 2079-2088.

92. Corpet, F. and Michot, B., *RNAlign program: alignment of RNA sequences using both primary and secondary structures.* Comp. Applic. Biosci., 1994. 10(4): p. 389-99.

93. Sakakibara, Y., Brown, M., Underwood, R. C., Mian, I. S., and Hauss *Stochastic Context-Free Grammars for Modeling RNA*. in *27th Hawaii International Conference on System Sciences*. 1994. Wailea, HI, U.S.A.: Los Alamitos, CA: IEEE Computer Society Press.

94. Lefebvre, F. *An optimized parsing algorithm well suited to RNA folding*. in *ISMB-95*. 1995. Cambridge, England: AAAI Press.

95. Pleij, C., *Pseudoknots: a new motif in the RNA GAme.* Trends Biochem. Sci., 1990. 15: p. 143-147.

96. Westhof, E. and Jaeger, L., *RNA pseudoknots.* Current Opinion In Struc Biology, 1992. 2: p. 327- 333.

97. Notredame, C., O'Brien, E. A., and Higgins, D. G., *RAGA: RNA Sequer Alignment by Genetic Algorithm.* Nucleic Acids Res.(in press), 1997. .

98. Notredame, C. and Higgins, D. G., *SAGA: sequence alignment by genetic algorithm.* Nucleic Acids Res., 1996. 24: p. 1515-1524.

99. Hirschberg, D. S., *A linear space algorithm for computing maximal common subsequences.* Comm. ACM, 1975. 18: p. 341-343.

100. Myers, E. W. and Miller, W., *Optimal alignments in linear space.* Comput. Appli Biosci., 1988. 4: p. 11-17.

101. Naor, D. and Brutlag, D. L., *On Near-Optimal Alignmnets of Biological Sequences.* Journal of Computational Biology, 1994. 1(4): p. 349-366.

102. Saqi, M. A. S. and Sternberg, M. J. E., *A simple method to generate non-trivial alternate alignments of protein sequences.* Journal of Molecular Biology, 1991. 219: 727-732.

103. Chao, K. M., *Computing all the suboptimal alignments in linear space.* Lecture Notes In Computer Science, 1994. 807: p. 31-42.

104. Vingron, M. and Argos, P., *Determination of reliable regions in protein sequence alignment.* Protein Eng., 1990. 3: p. 565-569.

105. Taylor, W. R. and Orengo, C. A., *Protein structure alignment.* Journal o Molecular Biology, 1989. 208: p. 1-22.

106. Sneath, P. H. A. and Sokal, R. R., *Numerical Taxonomy*. 1973, San Franci Freeman, W.H.

107. Barton, G. J. and Sternberg, M. J. E., *A strategy for the rapid multiple alignmen of protein sequences: confidence levels from tertiary structure comparisons.* Journal of Molecular Biology, 1987. 198: p. 327-337.

108. Gusfield, D., *Efficient methods for multiple sequence alignment with guaranteed error bounds.* Bull. Math. Biol., 1993. 55: p. 141-154.

109. Gupta, S. K., Kececioglu, J. D., and Schaffer, A. A., *Improving the practi apace and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment.* Journal of Computational Biology, 1995. 2(3): p. 459-472.

110. kececioglu, J. D., *The maximum weight trace problem in multiple sequence alignmnet.* Lecture Notes in Computer Science, 1993. 684: p. 106-119.

111. Reinert, K., Lenhof, H. P., Mutzel, P., Melhorn, K., and Kececioglu, J *branch-and-cut Algorithm for multiple sequence alignmnet.* Recomb97, 1997. : p. 241-249.

112. Kauffman, S. and Levin, S., *Toward a General Theory of Adaptative Walks on Rugged Landscapes.* J. Theor Biol, 1987. 128(1): p. 11-45.

113. Charleston, M. A., *Toward a characterisation of landscapes of combinatorial optimisation problems, with special attetion to the phylogeny problem.* Journal of Computational Biology, 1995. 2(3): p. 439-450.

114. Berger, M. P. and Munson, P. J., *A novel randomized iterative strategy for aligning multiple protein sequences.* Comput. Appl. Biosci., 1991. 7: p. 479-484.

115. Ishikawa, M., Toya, T., and Tokoti, Y. *Parallel Iterative Aligner with Genetic Algorithm*. in *Artifificial Intelligence and Genome Workshop, 13th International Conference on Artificial Intelligence*. 1993. Chambery, France:

116. Hirosawa, M., Totoki, Y., Hoshida, M., and Ishikawa, M., *Comprehensive s on iterative algorithms of multiple sequence alignments.* Comp. App. Biosci., 1995. 11(1): p. 13-18.

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

118. Geman, D. and Geman, D., *Gibbs Sampling.* Trans. Pattern Anal. Mach. 1984. 6: p. 721.

119. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., an E., *Equation of state calculations by fast computing machines.* J. Chem. Phys., 1953. 21: p. 1087-1092.

120. Kirkpatrick, S., Gelatt, C. D. J., and Vecchi, M. P., *Optimization by Simula Annealing.* Science, 1983. 220: p. 671-680.

121. Ingber, L. and Rosen, B., *Genetic Algorithm and Very Fast simulated Reannealing: a comparison.* Mathematical Computer Modeling, 1993. 16: p. 87-100.

122. Ishikawa, M., Toya, T., Hoshida, M., Nitta, K., Ogiwara, A., and Kaneh *Multiple sequence alignment by parallel simulated annealing.* Comp. Applic. Biosci., 1993. 9: p. 267-273.

123. Kim, J., Pramanik, S., and Chung, M. J., *Multiple Sequence Alignment usin Simulated Annealing.* Comp. Applic. Biosci., 1994. 10(4): p. 419-426.

124. Godzik, A. and Sander, C., *Conservation of residue interactions in a family of Ca-binding proteins.* Prot. Eng., 1989. 2: p. 589-96.

125. Schmitz, M. and Steger, G., *Description of RNA Folding by Simulate Annealing.* J. Mol. Biol, 1996. 255: p. 254-266.

126. Eddy, S. R. *Multiple alignment using hidden Markov models*. in *Third International converence on Intelligent Systems for Molecular Biology (ISMB)*. 1995. Cambridge, England: Menlo Park, CA: AAAI Press.

127. Holland, J. H., *Adaptation in natural and artificial systems*. 1975, Ann Arbour, MI University of Michigan Press.

128. Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*. ed. D.E. Goldberg. 1989, New York: Addison-Wesley.

129. Davis, L., *The handbook of Genetic Algorithms*. ed. L. Davis. 1991, New York: Van Nostrand Reinhold.

130. Rabow, A. A. and Scheraga, H. A., *Improved genetic algorithm for the protein folding problem by use of a cartesian combination operator.* Protein Science, 1996. 5: p. 1800-1815.

131. Pedersen, J. T. and Moult, J., *Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithms.* Proteins: Structure, Function, an Genetics, 1995. 23: p. 454-460.

132. Legrand, S. M. and Merz, K. M., *The genetic Algorithm and the conformational search of polypeptides and proteins.* Molecular Simulation, 1994. 13: p. 299-320.

133. Unger, R. and Moult, J., *Genetic Algorithms for Protein Folding Simulations.* J. Mol. Biol., 1993. 231: p. 75-81.

134. Schulze-Kremer, S., *Genetic algorithms for protein tertiary structure prediction,* in *Proceedings of the 2nd Conference on Parallel Problem Solving from Nature,* R. Männer and B. Manderick, Editor. 1992, Elsevier Science Publishers: Amsterdam. p. 3

135. Sun, S., *Reduced representation model of protein structure prediction: statistical potential and genetic algorithms.* Protein Science, 1993. 2(5): p. 762-785.

136. Dandekar, T. and Argos, P., *Potential of genetic algorithms in protein folding and engineering simulations.* Protein Eng., 1992. 5: p. 637-645.

137. Verkhivler, G. M., Rejto, P. A., Gelhaar, D. K., and Freer, S. T., *Exploring energy landscape of molecular recognition by a genetic algorithm: analysis of the requirement for robust docking of HIV-1 Protease and FKBP-2 Complexes.* Proteins:Structure, Function and Genetics, 1996. 250: p. 342-353.

138. Shapiro, B. A. and Wu, J. C., *An annealing mutation operator in the genetic algorithm for RNA folding.* Comp. Applic. Biosci., 1996. 12(3): p. 171-180.

139. Ogata, H., Yutaka, A., and Minoru, K., *A genetic algorithm based molecular modeling technique for RNA stem-loop structures.* Nucleic Acids res., 1995. 23(3): p 419-426.

140. McClure, M. A., Vasi, T. K., and Fitch, W. M., *Comparative analysis of multiple protein-sequence alignmnent methods.* Molecular Biology Evolution, 1994. 11(4): p. 5 592.

141. Subbiah, S. and Harrison, S. C., *A method for multiple sequence alignments with gaps.* J. Mol. Biol., 1989. 209: p. 539-548.

142. Sali, A. and Overington, J. P., *Derivation of Rules for comparative protein modeling from a database of protein structure alignments.* Protein Sci., 1994. 3: p. 1582-1596.

143. Holm, L. and Sander, C., *The FSSP database: fold classification based on structure-structure alignment of proteins.* Nucleic Acids Res., 1996. 24: p. 206-210.

144. Barton, G. J., *Scop: structural classification of proteins.* Trends Biochem Sci, 1994. 19(12): p. 554-5.

145. Gotoh, O., *A weighting System and Algorithm for aligning many phylogenetically related Sequences.* Comput. App. in Biosci., 1995. 11(543-551).

146. Tabaska, E. J. and D., S. G. *Automated alignment of RNA sequences to pseudoknotted structures*. in *ISMB-97*. 1997. Halkidiki, Greece: AAAI Press.